Word Category Maps based on Emergent Features Created by ICA

Jaakko J. Väyrynen and Timo Honkela

Helsinki University of Technology
Laboratory of Computer and Information Science
Neural Networks Research Centre
P.O. Box 5400, FIN-02015 HUT, Finland
http://www.cis.hut.fi/{jjvayryn, tho}

Abstract. In this paper, we assume that word co-occurrence statistics can be used to extract meaningful features, exhibiting syntactic and semantic behavior, from text data. Independent component analysis (ICA), an unsupervised statistical method, is applied to word usage statistics, calculated from a natural language corpora, to extract a number of features. With a self-organizing map (SOM), we will demonstrate that the extracted vector representation for words can further be applied to other tasks. It is also demonstrated, that the ICA-based encoding scheme is a good alternative to random projection (RP), a method commonly used in text analysis.

1 Introduction

Our goal is to analyze words and to learn interesting features from natural language data in an unsupervised manner. In this paper, we use independent component analysis [1, 2], a statistical method for blind source separation. The learning is done at word level from running text. Here a *word*, e.g. "blue" or "12", is a unique string of letters separated by white-spaces or possibly punctuation marks. The needed statistics are estimated from a text corpus, which represents the usage of written English.

It is assumed that the word usage statistics tell something about the structure and the rules of the language. First-order statistics (word frequencies) tell how common a word is. In this paper, higher-order statistics (co-occurrences of words in contexts) are used.

In [3] an analysis of a text corpus was conducted in which the relationships between the 150 most frequent words of the Grimm tales were studied. For the study, the statistical contextual relations of these words were represented two-dimensionally by the Self-Organizing Map (SOM) algorithm [4]. Encoding of the words was made using a 90-dimensional random real vector for each word. The code vectors of the words in a contextual window were then concatenated into a single input vector $x(t)$. In order to equalize the mapping for the selected 150 words statistically and to speed up computation, a solution used in [5] was to average the contexts relating to a particular focus word. As the main result

of the SOM-based analysis, the general organization of the map reflected both syntactical and semantical categories.

In [3], however, it remained an open question whether one could create a system that would provide automatically a set of meaningful features for each word. Rather than having a random encoding [6] for each word, one could ideally represent each word as a feature vector that would take into account its syntactic and semantic characteristics. This kind of sparse feature representation can be created automatically using independent component analysis (ICA) [2] as we have shown in [7]. In this paper, we first introduce shortly the basic idea how to analyze words in contexts using ICA. Then we compare the SOMs of words where random encoding and ICA-based encodings are used.

2 Analysis of Words in Contexts using ICA

In linguistics, a syntactic word category is usually defined as a set of words, where the syntax of the language is not broken, when a word is replaced with another word from the same word category. This is called is a replacement test. For instance, in the sentence "Mary gave John two flowers", the word "flowers" can be replaced with any plural noun without violating the syntactic rules of the language. In this article, it is assumed that there is no mechanism for checking whether a sentence is syntactically correct or not, and a statistical replacement test of context similarity is used instead. The idea is, that if two words occur in similar contexts, they should be assigned to the same category. If there are enough examples of the use of the language, the co-occurrence statistics should tell something about the structure of the language. In our analysis, only the closest words are used in the statistical replacement test. The co-occurrence statistics are collected into *context histograms*, in which the co-occurrences of focus words in a given context are collected.

It might be reasonable to assume that the context histograms are mixtures of word sets that resemble word categories. As an example, consider what kind of words could immediately precede nouns (e.g. "boys" and "girl"). Here each noun (the context) creates a new context histogram, and the immediately preceding word (the focus word) counts are elements in the context histograms. The mentioned nouns are countable, so they could be preceded by numerals ("two boys", "four girls"). Nouns can also have other attributes ("good boys", "young girls"). In case of noun contexts, the histograms might have a high frequency for adjectives. For plural noun contexts, in addition to adjectives, there might also be high frequency counts for numerals. With enough contexts, a statistical feature extraction method might be able to find features that resemble adjectives and numerals. If independent component analysis is used to extract the features, it is assumed that the context histograms are linear mixtures of features.

A connection to hidden Markov models (HMM) can be seen, where each state (context) has it's own probabilities for each word, and word sequences are generated by emitting words using the word probabilities for each state and the transition probabilities between states.

A numerical representation for the words is needed in order to use them for calculations. This can be accomplished by using a *vector space model* [8] and attaching a real vector $\mathbf{v}_i$ to each word w_i.

2.1 Context Histograms

The data for the ICA algorithm are the word co-occurrence frequencies in different contexts (context histograms). A context, c, is defined with the surrounding words of the focus word w. The co-occurrence frequencies of words in contexts are the un-normalized ML estimates of the conditional word probabilities $P(w|c)$.

Only a fraction of the the possible focus words and contexts were selected. This operation was conducted to lower the dimension of the representation, and to select a representative set of the context histograms. In the following, the selection process is examined in detail.

Focus Words The focus words are the words being modeled. The most common terms (different word forms) are usually chosen simply because they are what we want to model, and they contain much of the frequency information. In this paper, the focus words were chosen more specifically, as we wanted to examine the separation of verbs and adjectives, and the focus words were chosen to consist of only words that are verbs or adjectives.

Context Words The context words are not necessarily the same as the focus words. The context words define what co-occurrences of the focus words and the context words are calculated. The context words should be chosen to capture as much of the interesting information as possible, and can be chosen in many ways, e.g. by examining word frequency [9], function words, variance [10] or by statistical analysis of context histogram consistency [11]. The most frequent words were chosen as the context words. It is a simple method, but it should give good results [9].

The most frequent words overlap greatly with the so called function words (determiners, pronouns, auxiliary verbs etc.), that convey much of the syntactic information by binding together other words (verbs, nouns, adjectives etc.) that carry more specific meaning in a sentence.

An intuitive argument in favor of choosing the most common words is that since they are common, their co-occurrence histograms with other words is not very sparse and they represent most of the frequency information. Furthermore, less frequent words as context words might have more of a semantic role and their co-occurrence histograms are sparser. As an example, consider the portion of nouns that might occur in the context of the adjective "humid". The linear ICA might not find traditional syntactic categories, but more semantic ones. On the other hand, function words are very likely appear with any word, depending on the syntactic properties of the words.

If ICA is used to estimate the underlying components from context histograms, and the goal is to find components for word categories, the most frequent words are a good choice for the context words, since their context histograms might be combinations of different categories. Since the less frequent words have less co-occurrences, their context histograms might not contain a linear combination of syntactic categories, but subsets of categories.

If the goal is to find components with more semantic information, the context words could be chosen differently.

Context Types In addition to selecting context words, the calculation of the co-occurrences requires that the context word relation of the focus words is defined. Naturally, a context can be defined in many different ways. A simple method is to place the context word or words related to the focus word w. This is closely related to n-grams and Markov models, where the next word in a sequence, w_k is modeled using only the $n-1$ preceding words

$$P(w_k|w_{k-n+1}, w_{k-n+2}, \ldots, w_{k-1}) \ , \tag{1}$$

and simple bigrams ($n = 2$) and trigrams ($n = 3$) are commonly used in natural language processing. In this paper, bigrams were used.

The context word can also occur in a context window around the context word. Examples of different context types would be the the word immediately preceding the focus word ($P(w_k|w_{k-1} = c)$), the two words following the focus words ($P(w_k|w_{k+1} = c_1, w_{k+2} = c_2)$), or the context word might appear in a window around the focus words ($P(w_k|w_{k-1} = c \vee w_{k+1} = c)$). Naturally, different context types can be mixed to for the data.

The number of context words in the context c is usually limited for practical reasons. Long contexts would mean higher n-grams and more unreliable probability estimates. It would also mean either having a huge number of context histograms as the combinations of all $n-1$ context words, or creating a method of choosing the context histograms.

Limiting the context length and the position of the focus word, in relation to the context words, restricts the context histograms and the estimated features. It is hoped that enough, if not most, of the syntactic and semantic information is captured by the statistics of near-by words. For better results, it might be a good idea to vary the length of the context and the location of the focus word as much as it is computationally possible. This means using context histograms with different lengths and focus word positions.

2.2 Creating the Data Matrix

Contextual information was calculated as context histograms for focus words. Given a particular context (m chosen words) and their positions given the focus word, the corresponding co-occurrence frequencies were extracted from the text corpus.

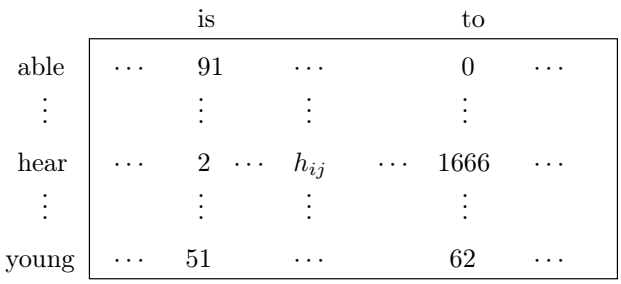

Fig. 1. An example context histogram matrix H. The rows are the modeled focus words w_i and the columns are the context histograms for different context words. Here the context is the word immediately preceding the focus word

When the context histograms have been calculated, a real matrix $H = (h_1, \ldots, h_C)$ of size $F \times C$ is ready. Here F is the number of focus words, and C the number of different context histograms. An example histograms is shown in Fig. 1, where the context word is the first word immediately preceding the focus word. The rows of H are the signals for focus words w_i, $i = 1, \ldots, F$ and the columns h_j, $j = 1, \ldots, C$ are the context histograms. The order of the columns is inconsequential to the ICA algorithm.

2.3 Preprocessing

The data matrix H gives the frequencies of the focus words in each context as the columns. The raw frequency counts are not the best input to the ICA algorithm, because of the large variations in frequencies, so some preprocessing is needed.

The frequency data is concentrated on the most frequent words. To lessen the effect of the word frequency, the logarithm of the elements h_{ij} increased by one could was taken. This kind of preprocessing should make the ICA algorithm model the words more equally, instead of modeling only the most frequent ones.

For computational reasons it might be necessary to perform dimension reduction before independent component analysis. The FastICA package [12] can be used to reduce the dimension of the problem with principal component analysis (PCA) before the actual ICA algorithm.

2.4 Feature Extraction with ICA

Similar to the experiments in [7] and [13], independent component analysis is applied to the the preprocessed context histogram matrix H to extract a number of features. The estimated mixing matrix A and the source matrix $S = (s_1, \ldots, s_K)^T$ are the result of applying ICA to the context histogram matrix H. The ICA model is illustrated in Fig. 2. The columns of the mixing matrix A are the features, and the rows give a vector representation for the focus words.

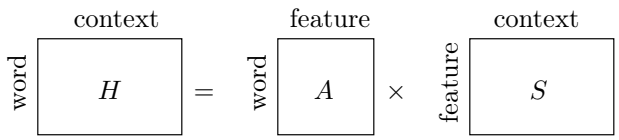

Fig. 2. The ICA model explains the data H using the source signals S and the mixing matrix A

2.5 Interpretation of the ICA-based Features

The columns a_k of the mixing matrix A are the features that linearly model the context histograms. Each feature can be seen as a sum of word vectors scaled by the intensity of the word in the component. According to our hypothesis, the extracted features a_k represent syntactic and semantic information in the corpus. An analysis of the ICA-based features has been conducted in [13], where the extracted features were compared to traditional syntactic word categories. However, applying the self-organizing map algorithm requires only latent structure to be present in the encoding scheme.

The rows of the mixing matrix A give a K-dimensional vector representation for each word. We will compare the ICA-based encoding of words to a random vector encoding with the help of the self-organizing map.

3 Word Category Maps

Here we will explain how the word category maps were created.

Word vectors were encoded with independent component analysis or random projection. A self-organizing map was taught with the word vectors using the SOM Toolbox for Matlab. Each word was placed on the map in its best matching unit (BMU), analogous to the results in [3]. As an emergent consequence, the self-organizing map organized the words according to their syntactic and semantic characteristics.

In order to be able to examine the syntactic categorization of the emerging map, we selected the modeled focus words to include only word forms that could be used as verbs, adjectives or both. More specifically, we selected the most common 187 word forms which had been tagged with the `VB` (verb, base: uninflected present, imperative or infinitive) or the `JJ` (adjective) tag in a subset of the tagged Brown corpus. Of the 187 words, 100 were at least once tagged with the `JJ` tag, and 103 with the `VB` tag. The two syntactic categories were selected, because they contain a lot of words and overlap only slightly. The information known about the verb and adjective categories was not used in the learning.

To qualitatively examine the separation of the `VB` and the `JJ` categories, the self-organizing map was plotted with the categories for words, not the actual words, in their BMUs, i.e., each word in was replaced with the category information collected from the Brown corpus.

4 Comparing Word Category Maps with Different Encoding Schemes

Here we will show comparisons of word category maps taught on vector representations for words acquired with independent component analysis and random projection. The corpus was the same as in [13]. The context histogram matrix H was calculated using the 187 selected focus words. Contextual information was calculated using the most common 5000 words in the corpus as the context words. Context histograms were calculated for the immediately following word, i.e, the elements h_{ij} in the context matrix H represented the number of occurrences the focus word w_i followed the context word j in the text corpus.

4.1 Experiments with Random Projection

Random projection was applied to the context histogram matrix H analogous to the experiments in [3]. The random projection method requires the destination dimension to have enough dimensions to make the projected vectors to be pseudo-orthogonal. This was clearly seen in the rapidly decreasing quality of the self-organizing maps taught with random projected data, and random projection methods fails completely with the low dimensions that can be achieved with ICA. The word category map thought with random projected contexts shows qualitatively similar properties when the dimension is high enough. Fig. 3 shows a word category map for 100-dimensional random projection of the preprocessed context histogram matrix H. Fig. 4 shows the same map with the possible syntactic categories `VB` and `JJ` marked for each word, and most of the map units model one of the categories. The self-organizing maps reflects the syntactic and semantic properties of the modeled words similar to the results in [3, 5].

4.2 Experiments with ICA

The FastICA package [12] was applied to extracting a chosen number of features from the postprocessed context histogram matrix H. Parameter selections were similar to those in [13].

Fig. 5 shows the ten-dimensional ICA-based encodings projected to two of the features. This illustrates the explicit categories found by ICA, first reported in [7]. It should be noted, that ICA can be defined as principal component analysis and whitening of the data followed by an ICA-rotation. [2] The rotation doesn't affect the Euclidean distances, and thus it doesn't affect the self-organizing map algorithm. However, as we are also interested in the explicit features provided by ICA, we are motivated to use independent component analysis over principal component analysis or singular value decomposition (SVD).

The data was the same as in the experiments in Sect. 4.1. Ten features were extracted and a self-organizing map was taught on the emergent ICA-based encoding for the words w_i. The word category maps are shown in Fig. 6 and Fig. 7. The maps reflects the syntactic and semantic properties of the modeled

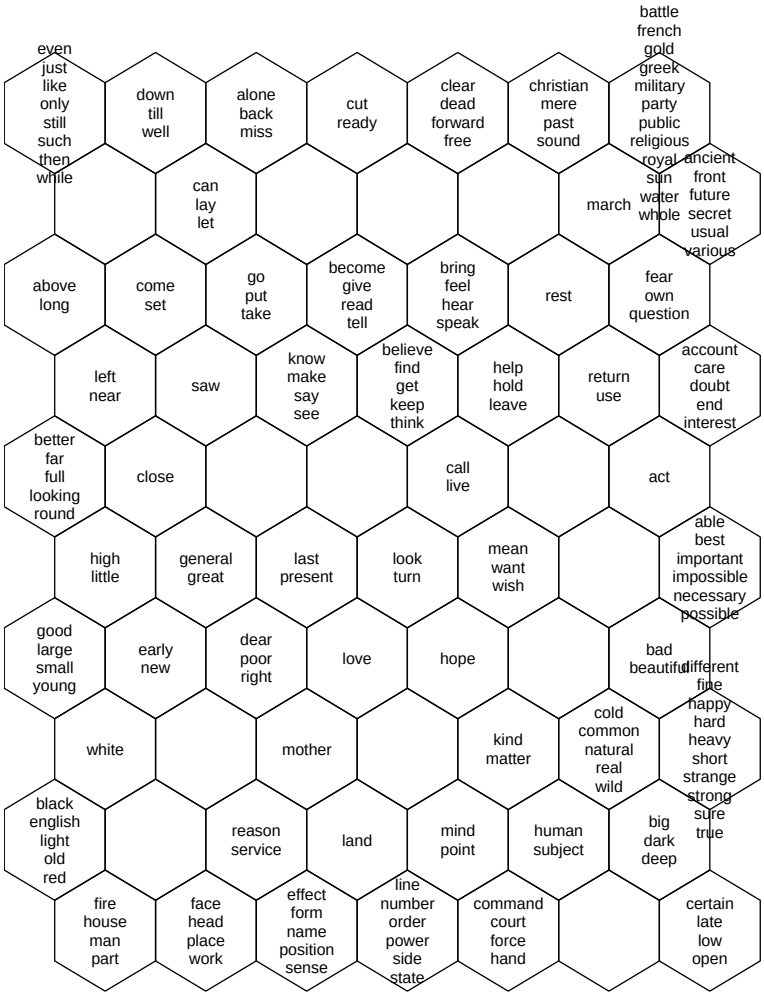

Fig. 3. A word category map created using a 100-dimensional random projection for words

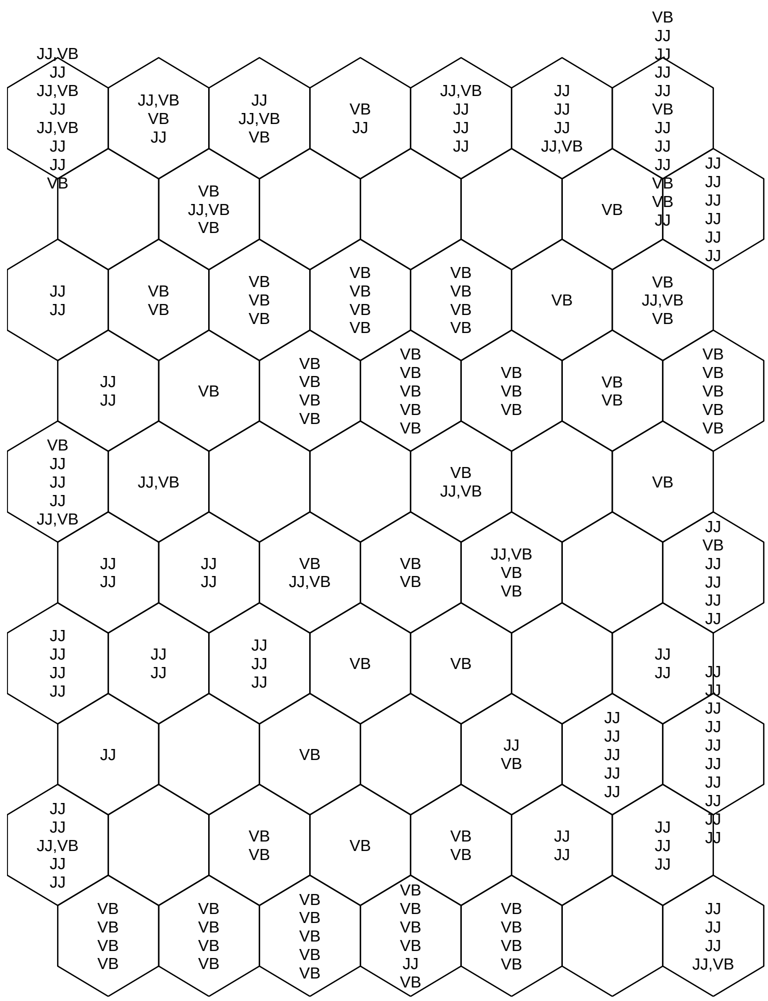

Fig. 4. The same map as in Fig. 3, but here the possible adjective (`JJ`) and verb (`VB`) categories for the focus words are shown

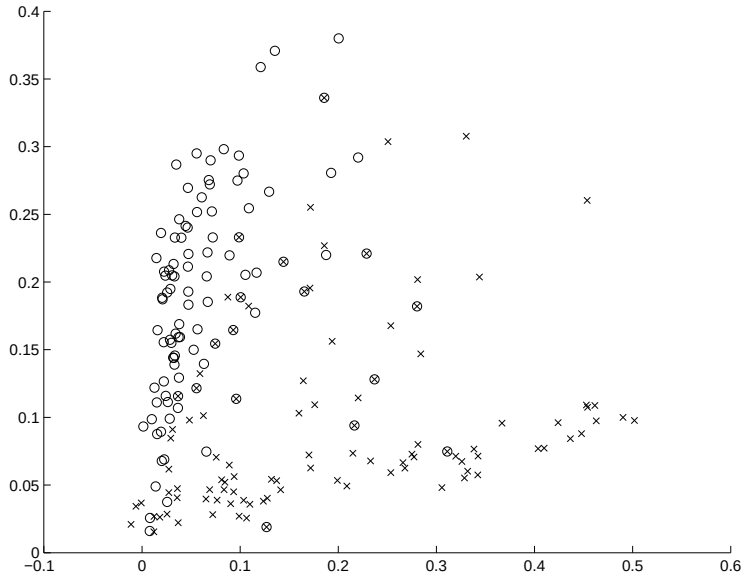

Fig. 5. A projection to two of the extracted features from a ten-dimensional ICA-representation. The selected dimensions were selected as the best features modeling the `JJ` (*circle*) and the `VB` (*cross*) categories of the 187 focus words

words similar to the results in [3, 5]. The studied `JJ`/`VB` categorization is visible in the map. Additionally, the verbs are clearly separated into to two regions, one modeling verbs witch can also act as nouns, and the other modeling verbs that are not used as nouns.

5 Conclusions

A self-organizing map was taught on vector representation for words. The vector representation was calculated from statistical information of words in contexts. The word category maps taught with ICA-based encoding were qualitatively better when compared to random projection encoding.

Random projection requires the vector encoding to have a sufficiently high dimension to work, but the method requires very little computation. Independent component analysis is a more computationally expensive method, but it is able to find a good representations with extremely low dimensions, where random encoding fails. Additionally, independent component analysis finds explicit features that show syntactic and semantic characteristics. The ICA-based features also enable nonlinear processing, such as removing "noise" from the data by thresholding, which should be explored further. Although principal component analysis would have been enough for creating the word category maps shown in this paper, we were motivated to use ICA for the properties mentioned above.

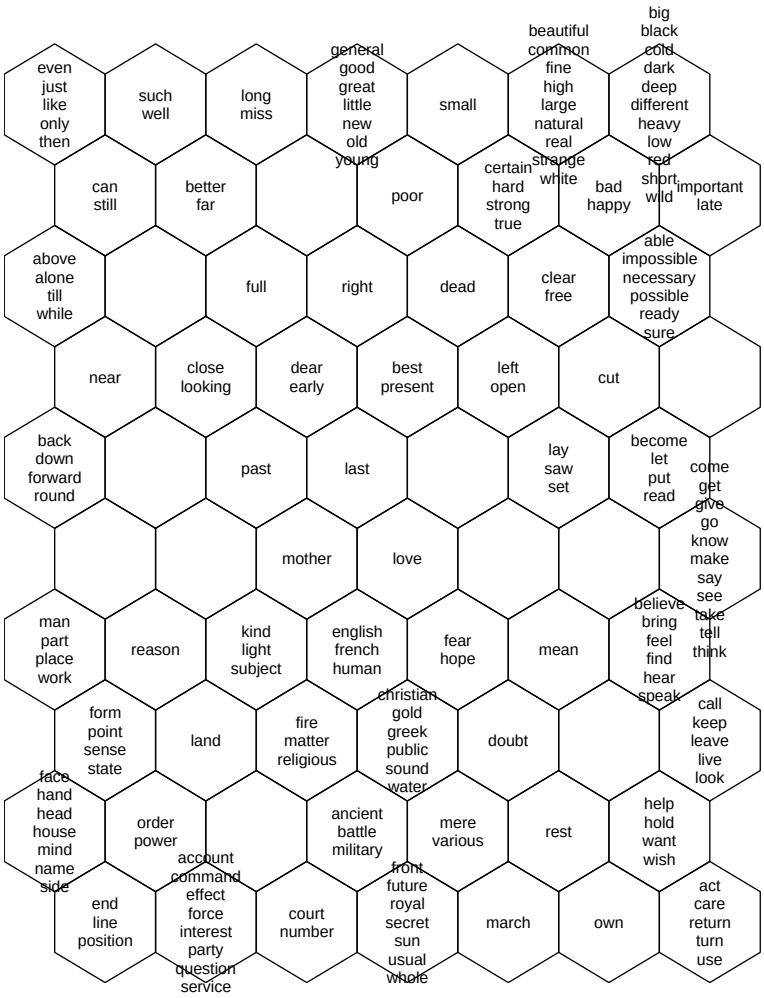

Fig. 6. A word category map created using a 10-dimensional ICA-based encoding. The selected and shown focus words appear as adjectives or verbs in a tagged Brown corpus. The map shows both syntactic and semantic characteristics

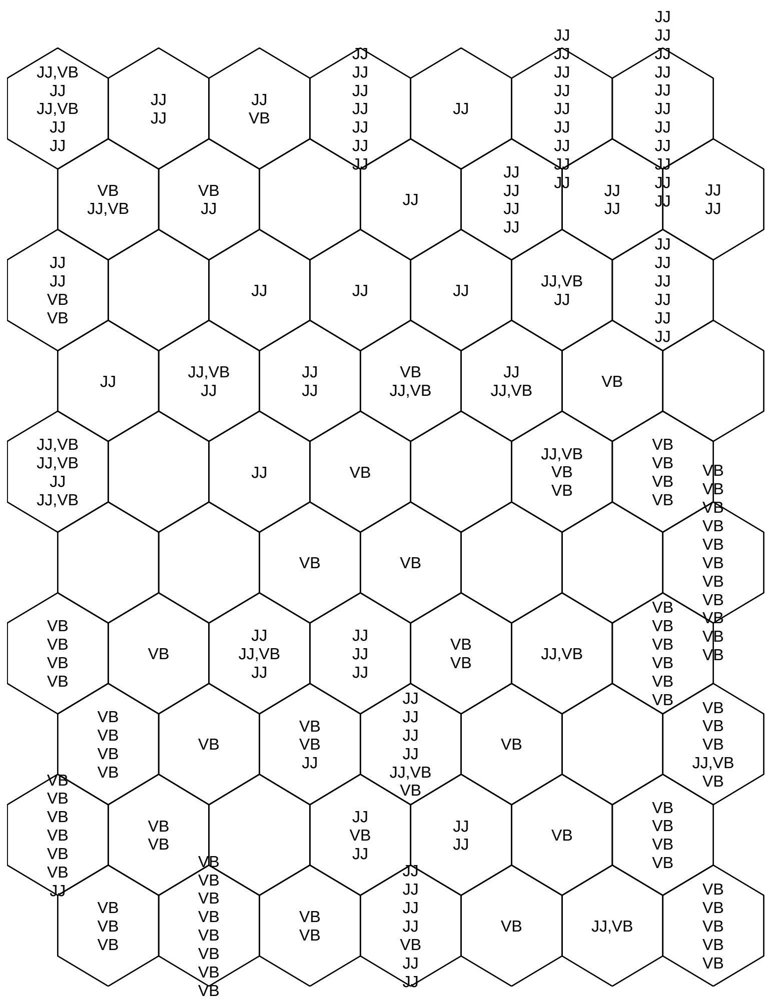

Fig. 7. The same map as in Fig. 6, but here the possible adjective (`JJ`) and verb (`VB`) categories for the focus words are shown. Notice the regions of the syntactic categories

We have demonstrated that the ICA-based vector representation for words could be a useful tool in the practical applications of language technology. Other language technology applications that could benefit of the ICA-based encoding approach include information retrieval and machine translation.

References

1. Comon, P.: Independent Component Analysis, a new concept ? Signal Processing, Elsevier **36** (1994) 287–314
2. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons (2001)
3. Honkela, T., Pulkki, V., Kohonen, T.: Contextual relations of words in Grimm tales analyzed by self-organizing map. In Fogelman-Soulié, F., Gallinari, P., eds.: Proceedings of ICANN-95, International Conference on Artificial Neural Networks. Volume 2., Nanterre, France, EC2 (1995) 3–7
4. Kohonen, T.: Self-Organizing Maps. Springer (1995)
5. Ritter, H., Kohonen, T.: Self-organizing semantic maps. Biological Cybernetics **61** (1989) 241–254
6. Kaski, S.: Dimensionality reduction by random mapping: fast similarity computation for clustering. In: Proc. Int. Joint Conf. on Neural Networks (IJCNN'98), Anchorage, Alaska (1998) 413–418
7. Honkela, T., Hyvärinen, A., Väyrynen, J.: Emergence of linguistic representation by independent component analysis. Technical report, Helsinki University of Technology (2003)
8. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18** (1975) 613–620
9. Levy, J.P., Bullinaria, J.A.: Learning lexical properties from word usage patterns: Which context words should be used? In French, R., Sougne, J., eds.: Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop, London, Springer (2001) 273–282
10. Lund, K., Burgess, C.: Producing highdimensional semantic spaces from lexical cooccurrence. Behavior Research Methods, Instruments and Computers **28** (1996) 203–208
11. Lowe, W., McDonald, S.: The direct route: Mediated priming in semantic space. In Gernsbacher, M., Derry, S., eds.: Proceedings of the 22nd Annual Meeting of the Cognitive Science Society, New Jersey, Lawrence Erlbaum Associates (2000) 675–680
12. The FastICA Team at the Helsinki University of Technology: The FastICA MATLAB package. Available at `http://www.cis.hut.fi/projects/ica/fastica/` (1998)
13. Väyrynen, J.J.: Learning linguistic features from natural text data by independent component analysis. Master's thesis, Helsinki University of Technology (2004) Unpublished.
14. Schütze, H.: Ambiguity resolution in language learning. CSLI Publications (1997)