ESUKA – JEFUL 2017, 8–2: 93–125

AN AUTOENCODER-BASED NEURAL NETWORK MODEL FOR SELECTIONAL PREFERENCE: EVIDENCE FROM PSEUDO-DISAMBIGUATION AND CLOZE TASKS

Aki-Juhani Kyröläinen, Juhani Luotolahti, and Filip Ginter
University of Turku

Abstract. Intuitively, some predicates have a better fit with certain arguments than others. Usage-based models of language emphasize the importance of semantic similarity in shaping the structuring of constructions (form and meaning). In this study, we focus on modeling the semantics of transitive constructions in Finnish and present an autoencoder-based neural network model trained on semantic vectors based on Word2vec. This model builds on the distributional hypothesis according to which semantic information is primarily shaped by contextual information. Specifically, we focus on the realization of the object. The performance of the model is evaluated in two tasks: a pseudo-disambiguation and a cloze task. Additionally, we contrast the performance of the autoencoder with a previously implemented neural model. In general, the results show that our model achieves an excellent performance on these tasks in comparison to the other models. The results are discussed in terms of usage-based construction grammar.

Keywords: neural network, autoencoder, semantic vector, usage-based model, Finnish

DOI: https://doi.org/10.12697/jeful.2017.8.2.04

1. Introduction

Intuitively it is clear that predicates have a better fit with certain arguments than others. For example, *I ate* is more likely to combine with *apple* than with *car*. In usage-based models of language, semantic similarity plays a crucial role in the formation of constructions, mappings between form and meaning/function. Semantic similarity is assumed to be one of the primary factors that influences the formation of new usage patterns (Bybee and Eddington 2006, Kalyan 2012). Importantly, Goldberg (1995) has formulated the principle of semantic compatibility that constrains the usage of argument structure constructions in

language. Thus, these types of models strongly rely on the notion of semantic similarity in determining the goodness-of-fit of a particular lexical item in a given construction. (Goldberg 2006, Bybee 2010) and there is ample evidence demonstrating how the goodness-of-fit also influences processing as measured by eye-movements (Ehrlich and Rayner 1981) and event-related brain potentials (Kutas and Hillyard 1984). It is, however, an open question how to exactly model the semantics of constructions. In this study, we focus on modeling the semantics of transitive constructions – who did what to whom – in Finnish. To model the semantic structure, we implemented a neural network to model the goodness-of-fit of lexical items in a given transitive construction. Specifically, we focus on the realization of the object in this argument structure construction as objects have shown to have high-cue validity in disambiguating the semantics of transitive constructions compared to predicates and subjects, at least in English (see Yarowsky 1993). In this respect, this study is closely connected to models of selectional preference, i.e., the semantic fit of a given word relative to its context (Erk, Padó and Padó 2010, Baroni and Lenci 2010, Lenci 2011, Van de Cruys 2014).

Usage-based models emphasize the role of the low-level generalizations rather than abstract structures in the formation of semantic information. Additionally, these models assume that semantic information is shaped by experience. Thus, semantic information is assumed to be formed by forming associations over usage patterns (see, for example, Bybee 2010, Ramscar *et al.* 2014). This notion follows the distributional hypothesis according to which the degree of semantic similarity between words is primarily driven by their context of use (Harris 1951, Firth 1957). Given that directly modeling prior experience is not feasible, there is a long tradition in computer and cognitive science to utilize corpus-based co-occurrence information to model the structuring of semantic relations, such as the Hyperspace Analog to Language (HAL; Lund and Burgess 1996) and Latent Semantic Analysis (LSA; Landauer and Dumais 1997) commonly referred to as semantic vector models. Related to this, Suttle and Goldberg (2011) have shown using LSA that people are more confident in accepting a newly formed verb when it is semantically similar to existing ones in English. In general, the estimated semantic similarities based on these models have been extensively investigated in experimental and corpus settings as a general purpose model of semantic memory (see Durda and Buchanan 2008, Baroni and Lenci 2010). Thus, these types of models assume that words that share similar usage patterns are also likely to be

semantically similar. However, these types of models rely on counting the co-occurrence of words in a given corpus and they can become computationally demanding when the co-occurrences are estimated based on a large-scale corpus.

Recently, a paradigm shift has emerged and, rather than counting the co-occurrence patterns of words, neural networks are used to model this type of structuring. Specifically, these types of models are used to predict the co-occurrence patterns associated with words in a language. Artificial neural networks are models originally inspired by the functioning of biological systems. These models consist of connected nodes called "neurons" and learning takes place by adjusting the activation weights of these nodes. Modeling linguistic structures with neural networks has a long tradition in usage-based models, i.e., connectionist models of language. Neural networks have been used to model the structuring of irregular verbs in English (Rumelhart and McClelland 1986), syntactic production (Chang, Dell, and Bock 2006) and morphological processing (Baayen *et al.* 2011), among others. Importantly, these types of models share the assumption of the distributional hypothesis with usage-based models. For the purposes of the present study, we implemented a neural network called word2vec to model semantic similarity relations among words (Mikolov *et al.* 2013). This algorithm has been shown to have excellent performance compared to the traditional count-based models (see Baroni, Dinu, and Kruszewski 2014, for example) and this model is discussed in detail in Section 2. Similar to count-based models, word2vec can be used to model the semantic similarity between pairs of words based on the contextual information of the words, for example, the similarity between *eat* and *apple*. However, the semantic structure of argument structure constructions such as the transitive construction investigated in this study do not necessarily depend solely on the relationship between word pairs, but also the semantics of the construction must also be considered in terms of the goodness-of-fit (see Suttle and Goldberg 2011: 1157, for discussion). This is an interesting empirical question given that a transitive construction minimally consists of three obligatory slots in Finnish: subject, verb and object. This allows us to test whether there is a substantial difference between models that rely on word pairs and those that include the whole structure of the argument structure construction. In this study, we specifically contrast the performance of these two types of models.

To model the whole semantic structure of the transitive construction, we implemented an autoencoder-based neural network architecture, as these are widely used in different scientific domains (Hinton and Zemel

1994, Bengio 2009). Autoencoders are a type of neural networks that encode the input by smoothing it and then reconstructing it. Often, these are used to create representations of the data in lower dimensionality. In cognitive science, autoencoders have been previously used to model, for example the structuring of categories in adults (Kurtz 2007) and the formation of categories in children (Mareschal, French, and Quinn 2000). Conceptually, this makes the architecture of an autoencoder highly suitable for modeling the semantics of constructions as constructions are argued to be generalizations over usage patterns (see Goldberg 2006, Bybee 2010, Croft 2001). The details of the implemented model are discussed in Section 3. At the same time, an autoencoder is only one possible neural network model that can be used to model semantic structuring. Recently, Van de Cruys (2014) implemented a binary neural classifier to model the semantics of transitive constructions in English. In order to compare the performance of the implemented autoencoder as a neural model of semantic information for constructions, we reimplemented the binary neural classifier for Finnish, discussed in Section 4. Thus, this allows us to directly compare the performance of these two types of neural models.

To evaluate and compare the performance of the neural models, two tasks were implemented that have been previously used to model the structuring of semantic information. The first is a corpus-based pseudo-disambiguation task (Yarowsky 1993). This task makes it possible to evaluate the performance of a model in terms of discriminating between semantically plausible and implausible realizations of the object in a given transitive construction. The details of this task are discussed in Section 5. The second task used in this study is a cloze task (Taylor 1953) and it is commonly used in experimental studies to evaluate how predictable a specific completion is in a given context (see Rayner *et al.* 2011, for example). The task is described in Section 6. Finally, we discuss the performance of the models and their conceptual basis in relation to usage-based models and outline possible directions of future research in Section 7.

2. Modeling the semantic information of words with Word2vec

To approximate semantic structuring in language, semantic models are typically trained on some corpus data. For the purposes of the present study, the data were extracted from the Finnish Internet Parsebank. This corpus contains approximately 3.7 billion tokens (Kanerva *et al.* 2014).

The corpus is automatically tagged for syntactic and morphological information. The performance of the parser is estimated to have a labeled attachment score of 81.4%. The resources are publicly available and can be found at <http://bionlp.utu.fi>. To construct the semantic vector presentation for the Finnish lexicon, we used the skip-gram version of the Word2vec algorithm (Mikolov *et al.* 2013). This type of model learns to predict the context words of a given target word by changing the activation weights of the nodes in the hidden layer. This type of neural model is illustrated in Figure 1.

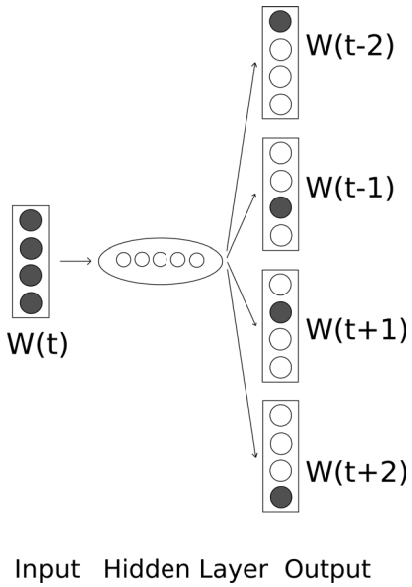

Figure 1. A visualization of the Word2vec neural model.

Given that we are interested in modeling the semantic structure of the transitive construction in Finnish, we used the lemmatized version of the corpus as we are not interested in morphological relations of the transitive construction. The skip-gram model was trained on the whole corpus using a window size of five, i.e., up to five words before and after a given target word as a larger window size has been shown to more closely reflect global semantic information (Levy and Goldberg 2014). Additionally, the semantic information associated with the words were represented using a semantic space of 200 dimensions. Conceptually, these dimensions can be understood as variables that together form the semantic space. All the other parameters were kept at their default value.

Once the model is trained, it is possible to compute a measure of semantic similarity in this space between words using a cosine similarity. Cosine similarity ranges between -1 and 1 where -1 means that two words are diametrically opposite and 1 when they are the same. It is important to emphasize that, in this context, semantic similarity refers to degree of similarity between words based on their shared context (see Turney 2006 for discussion). Importantly, words can be semantically similar even if they do not co-occur in a given corpus. We illustrate these types of semantic relations for five Finnish words with pairwise semantic similarities in Table 1.

Table 1. Pairwise cosine similarities for five Finnish words.

	äiti	*valtio*	*hoitaa*	*lapsi*	*tehtävä*
äiti 'mother'	1	0,058	0,152	0,759	0,104
valtio 'government'	0,058	1	0,19	0,201	0,402
hoitaa 'take care of'	0,152	0,19	1	0,202	0,246
lapsi 'child'	0,759	0,201	0,202	1	0,261
tehtävä 'task'	0,104	0,402	0,246	0,261	1

In Table 1, the diagonal is always one because the usage pattern of a given word is always identical to itself. The results indicate that *äiti* 'mother' is estimated to be highly semantically similar to *lapsi* 'child', as expected, and dissimilar to *tehtävä* 'task'. In terms of transitive constructions, it is now possible to estimate pairwise similarities between lexical realizations of the arguments in a given transitive construction. This model serves two purposes. First, we can use this semantic vector representation of words to estimate the pairwise semantic similarities among words in transitive constructions. For example, to represent the lexical realization of *äiti* 'mother' as a subject in a transitive construction relative to the realization of the object such as *lapsi* 'child'. Second, we can use this type of semantic vector representation of words as input for other neural models. For the purposes of the present study, the latter property is the most important because we can use these vectors to model the whole semantic structuring of the transitive construction. To achieve this, two neural models were implemented and these are discussed in the following two sections.

3. Modeling the semantics of transitive constructions with an autoencoder

For the purpose of the present study, an autoencoder-based neural network architecture (AE) was implemented (Hinton and Zemel 1994, Bengio 2009). A simple autoencoder is a three-layer neural network where the input and the output are directly connected. The implemented model is visualized in Figure 2. This type of model first encodes the input in a lower dimensional space (hidden layer) and then tries to reconstruct the input (output layer). In our case, the model received as its input the word vectors of the subject and the verb, encoded them and, finally, reconstructed them. The semantic vectors of these words were estimated with word2vec as described in the previous section. The semantic vectors of the subject and the verb were concatenated to form a single vector, which was fed to a dense neural network layer with hyperbolic tangent activation and an output size of 200, the same size as a single word vector. The second part of our model architecture consists of the mapping function for the object. Given the encoded semantic vectors of the subject and verb, the model predicts the semantic vector of the object. Importantly, the realizations of the object were never given as an input for the model. Similarly, hyperbolic tangent was used as an activation function for this layer. In this respect, our system could be seen as a mapping in the vector space from the subject and verb vectors into their most probable object vector. In this model, the estimated semantic similarity of the mapped object is always relative to the subject and the verb slots. Following a standard practice, the network, implemented in Keras (Chollet 2015), was trained to minimize the mean squared error of these three vectors. The training data contained 1,428,439 unique transitive triplets consisting of a subject, verb and object.

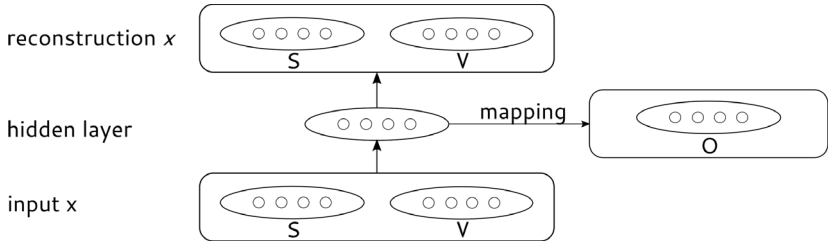

Figure 2. A visualization of the implemented autoencoder model.

The model both reconstructs the original subject and verb word vectors, and also produces a predicted word vector for the object. Since the model predicts an object which is based on examples on the training corpus, we can use the distance between the predicted, mapped object vector, and an arbitrary object vector to model the semantic fit of a given lexical realization of the object in a transitive construction. This semantic fit can be expressed using a cosine similarity between the predicted object vector and the semantic vector of the object. We illustrate the semantic structure learnt by the AE with the verb *hoitaa* 'take care of' in Table 2. The left side of the table contains the estimated six closest semantic neighbors for the object when the subject slot was filled with *äiti* 'mother'. On the right side, the semantic neighbors are given for the object when the subject slot was filled with *hallinto* 'government'. Additionally, the cosine similarity for the six closest semantic neighbors is provided in the table.

Table 2. Estimated best objects with the autoencoder based on the realization of the subject argument with the verb hoitaa 'take care of'.

äiti hoitaa		*hallinto hoitaa*	
Object	Cosine similarity	Object	Cosine similarity
lapsi 'child'	0,662	*tehtävä* 'task'	0,772
vauva 'baby'	0,642	*käytäntö* 'practice'	0,665
vanhempi 'parent'	0,597	*asia* 'thing'	0,659
koira 'dog'	0,568	*perustehtävä* 'basic task'	0,655
perhe 'family'	0,568	*toimenpide* 'procedure'	0,653
koti 'home'	0,55	*toimi* 'deed, post'	0,641

At least from a qualitative perspective, the implemented model appears to be capable of modeling selectional preference in simplex transitive constructions as the semantic fit of the object is modulated by the realization of the subject and the verb as expected. However, the goal of this study is to test how well the implemented model generalizes across different tasks. Before evaluating the performance of this model, we will introduce a previously implemented neural network model in the following section. In this way, it is possible to compare the impact of different architectures on modeling semantic similarity relations in Finnish transitive constructions.

4. Modeling the semantics of transitive constructions with a binary neural classifier

To contrast the performance of the autoencoder, we also implemented a binary neural network classifier (BiNN) based on the work of Van de Cruys (2014). Van de Cruys (2014) used this architecture to model selectional preference in English, i.e., the realization of the object in a transitive construction. The architecture is visualized in Figure 3. Similar to the AE, this model was trained on the semantic vectors estimated with the word2vec and the model was also implemented in Keras. The structure of the BiNN is a feed-forward neural network, which receives as its input word vectors consisting of a triplet, i.e., the subject, verb and object. During the training of the model, these vectors were fed into a dense neural network layer consisting of 200 neurons and the output of this hidden layer was fed to another neural network layer with a single output, ranging from zero to one. Because the output value is a measure of probability, we can use this model to evaluate and rank subject-verb-object triplets on their meaningfulness.

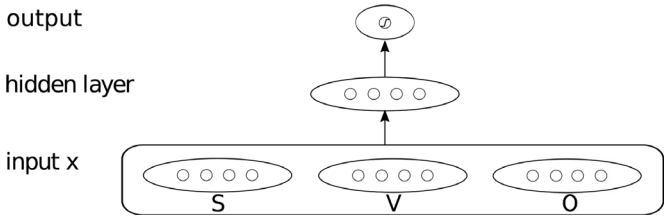

Figure 3. A visualization of the implemented binary neural classifier (BiNN) after Van de Cruys (2014).

There are, however, critical differences between these two architectures implemented in this study. The first difference concerns the number of inputs available in the models. The AE predicts the object vector based on the combination of a subject and verb vector. In contrast, the predictions of the BiNN are based on the whole triplet, i.e., subject, verb and object. The second difference concerns the estimates. The BiNN produces a single estimate of goodness-of-fit whereas the AE produces an estimate for the object given the semantic structure of the subject and the verb. The third difference concerns the training of these models. The AE was only trained on positive instances. In contrast, the BiNN is a binary classifier and requires that the input for the model explicitly

contain both positive and negative instances in order for this type of model to learn representations. We followed the same training procedure as was used in Van de Cruys (2014). To reconstruct false instance of the transitive construction, we implemented the same procedure as in the case of the pseudo-disambiguation task described in Section 5. For example, the training data contained pairs of attested triplets such as subject *mies* 'man', verb *dokata* 'booze' and object *sossuraha* 'social security money' and unattested instances in which the object was replaced with a random object such as *rikollispomo* 'kingpin'. Thus, the AE was only trained on the attested instances of the transitive construction whereas the BiNN also received false instances.

5. Experiment 1: pseudo-disambiguation task

To evaluate the performance of the neural network models, we implemented a pseudo-disambiguation task (see Yarowsky 1993) as it has been previously used to model selectional preference in English transitive constructions (Van de Cruys 2014, Erk, Padó, and Padó 2010). The task itself is effectively a binary classification task where the performance of the a given model is evaluated in terms of its ability to discriminate between true and false objects in a given transitive construction. This is a purely corpus-based task but it can be understood as mimicking a plausibility rating task were participants are asked to rate the goodness-of-fit of a given object in a sentence (see Rayner *et al.* 2004 for example).

In order to implement the task, we first created a corpus of lemmatized triplets consisting of a subject, verb and object extracted from the Finnish Internet Parsebank. In total, this corpus contained 2000 lemmatized triplets, for example subject *mies* 'man', verb *dokata* 'booze' and object *sossuraha* 'social security money' and all verbs were unique in these transitive constructions. Importantly, these instances were not part of the data set used to train the neural networks in order to avoid overfitting, i.e., a model simply learned the distributional properties of the training data but cannot properly generalize to unseen data. In the pseudo-disambiguation task, the objects of these triplets are considered as the true instances. In order to create the false instances, we extracted all the objects from the triplets to form a corpus of possible objects. In the pseudo-disambiguation task, the true object of a given transitive construction is replaced with an object selected at random from the corpus of possible objects, for example subject *mies* 'man', verb *dokata*

'booze' and object *rikollispomo* 'kingpin'. In this vein, the task tests whether the models can discriminate between these two types objects by selecting the true/original object of a given transitive construction.

For the purposes of the present study, we implemented two versions of this task. In the first one, the false objects are assigned at random. We will refer to this as the random condition. In the second one, not only are the object assigned at random but they were also matched in frequency (see Dagan, Lee, and Pereira 1999). We will refer to this as the matched condition. Count-based models such as HAL have been shown to be highly sensitive to differences in frequency distributions (Shaoul and Westbury 2010). The inclusion of the latter condition allows us to see the possible impact of frequency on the performance of the models. However, it is currently unclear whether frequency also influences the performance of word2vec. At the same time, it is worth pointing out that naturally occurring linguistic elements are not balanced in frequency but the matching condition, nonetheless, enables us to evaluate the potential impact of frequency (see also Erk, Padó, and Padó 2010: 737–738, for discussion).

In order to make the possible contribution of frequency even more tangible, we implemented a simple fallback n-gram model (Ngram) for the pseudo-disambiguation task. This model first attempts to discriminate between the true and false object based on a trigram (SVO) frequency. In the case that the trigram frequency is not observed in the corpus, this model falls back to a bigram frequency (verb and object) and, finally, to a unigram frequency (object) if the bigram frequency was not covered in the corpus. These frequency counts are based on the whole Finnish Internet Parsebank. For this corpus-based task, this n-gram model also serves as a baseline. In sum, the following models were evaluated in this task: 1) n-gram (Ngram), 2) word2vec-based pairwise similarity between the subject and the object (Word2vec_SO), 3) word2vec-based pairwise similarity between the verb and the object (Word2vec_VO), 4) an autoencoder (AE) and 5) a binary neural classifier (BiNN).

5.1. Evaluation of the models in a pseudo-disambiguation task

Given that the false objects were sampled at random, the pseudo-disambiguation task was repeated 1000 in each condition as this also allows us to construct confidence intervals for accuracy (Efron and Tibshirani 1993). For the vector-based models, the correct instances

corresponded to those cases where a given model assigned a higher cosine similarity to the true object relative to the false one. In the case of the fallback n-gram model, the difference in frequency was used as a measure of accuracy where the correct instances correspond to the true object that received a higher frequency. Finally, in the case of BiNN, a classification was considered correct if it received a higher probability than the false instance. To evaluate the performance of the models in this task, we report the average classification accuracy, i.e., the average accuracy of a particular model in a given run over the 2000 triplets. The distribution of the classification accuracy of the models is visualized in Figure 4 using a violin plot that combines a boxplot and a density plot.

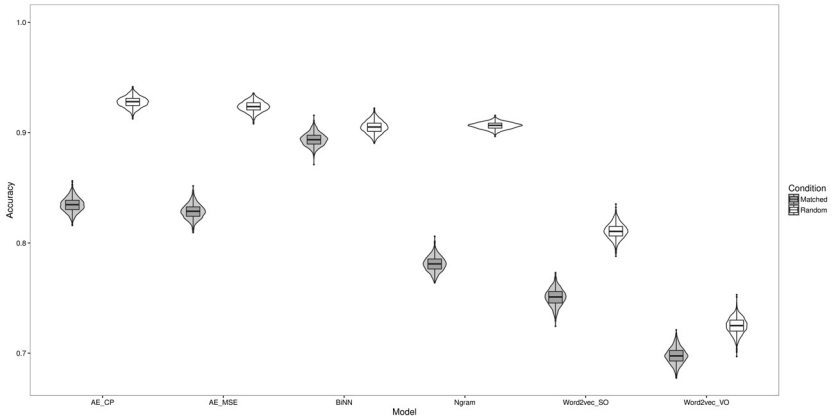

Figure 4. A zoomed in violin plot for the distribution of the classification accuracy of the models in the pseudo-disambiguation task across the two conditions. Each condition was repeated 1000 times.

In general, the results show that all the models performed well above chance which would correspond to an average classification accuracy of 0.5. Additionally, all the models obtained the best performance in the random condition. The density plots indicate that all the models appeared to be fairly consistent as the peak of the distributions is located around the median classification accuracy (bar inside the boxplot). In terms of the random condition, the AE achieved the highest average classification accuracy (M = 0.923, 95% CI [0.914, 0.933]) compared to all the other models. The BiNN obtained the second best performance (M = 0.905, 95% CI [0.894, 0.916]) and the difference, albeit small, between it and the AE was statistically significant, t(1944.8) = 80.2,

$p < .001$. Importantly, both neural network models outperformed the simple Ngram model and the two purely vector-based models in this task, although the average classification accuracy of the Ngram model was not overly poor, $M = 0.906$, 95% CI [0.9, 0.912]. Interestingly, both of the simple vector-based models were outperformed by the other models: cosine similarity between the subject and the object (Word2vec_SO) or between the verb and the object (Word2vec_VO). At the same time, the results suggest that a relatively decent average classification accuracy can be obtained even by simply computing the similarity between the subject and object, $M = 0.811$, 95% CI [0.797, 0.824], in the random condition.

In terms of the matched condition, all the models performed worse than in the random condition. However, the performance of the AE dropped drastically, $M = 0.828$, 95% CI [0.816, 0.842], in contrast to the BiNN, $M = 0.893$, 95% CI [0.883, 0.905]. In this respect, the BiNN appears fairly immune to differences in frequency distributions, although the difference in the average classification accuracy between the matched and the random conditions was statistically significant even with the BiNN, $t(1996.2) = –44.469$, $p < .001$. As expected, a similar decrease in performance was observed with the Ngram model, which only reached an average classification accuracy of 0.781 (95% CI [0.768, 0.794]), although it outperformed both purely vector-based models.

5.2. Discussion

The results of the pseudo-disambiguation task showed that an excellent average classification accuracy can be obtained with distributional models of semantics and, importantly, the models also appear to be consistent in their predictions. In this experiment, we included two pure semantic vector models based on the word2vec algorithm (Word2vec_SO and Word2vec_VO) as these can be viewed as serving as a baseline for distributional models of semantics. The former model is based on the semantic similarity between the subject and the object and the latter on the similarity between verb and the object. The results of these semantic vector models demonstrate that these types of models are capable of learning basic semantic structures. Specifically, subject and object arguments of a transitive construction appear to be in closer proximity in the vector space than verbs and object arguments as expected since these arguments tend to be realized as nominals. This indicates that

the word2vec model has implicitly learnt basic part-of-speech information based on similar usage patterns of words in a text, as the underlying model has never seen information associated with part-of-speech. Given a large enough corpus, the word2vec model appears to be able to learn similarity relations among words and, importantly, abstract over them. Interestingly, the results presented here show that by combining these semantic vector models with neural networks, even better average classification accuracy can be obtained, at least in the pseudo-disambiguation task. At the same time, this is to be expected as both the AE and the BiNN have access to a greater amount of information, specifically to the semantic vectors associated with the verbs (see Erk, Padó, and Padó 2010 for discussion).

Our AE model showed the best performance in the random condition, indicating that this architecture is fully capable of generalizing to unseen data and outperformed the BiNN model. Importantly, both of the neural network models outperformed the simple fallback n-gram model in this task. This suggests that by combining the semantic vectors associated with the subject and the verb, these models learn a vector representation that affords a meaningful mapping to the object argument. In this way, the model appears to capture a low-level semantic representation of a transitive construction. For example, in the case of the verb *hoitaa* 'take care of', the realization of the subject argument influences the semantic fit of the object argument. Surprisingly, the AE model showed a drop in performance when the frequency of the objects was matched but this was not the case with the BiNN. Given that both of the neural network models were trained on the same semantic vectors, this difference is unlikely to be simply related to frequency distributions. The simplest explanation for this difference is most likely related to the amount of information available to a given model. The AE model was only trained on positive instances whereas the BiNN was explicitly trained also on negative instances. This appears to offer a greater degree of discriminatory power. Another possibility could be related to the semantic structuring learnt by the AE. Specifically, the model predicts the most probable object vectors. In case of low frequency objects, both true and false instances could be located further away from the predicted most probable object vector, making it difficult to discriminate between them. We will leave this type of investigation for future studies. However, it is also possible that the predictions of these models are qualitatively different. We will investigate this property of the models in the following cloze task.

6. Experiment 2: cloze task

Another aspect related to goodness-of-fit is the lexical predictability of a given item in a sentence. Lexical predictability has been shown to influence language processing in experimental studies as measured by eye-movements and event-related brain potentials (Ehrlich and Rayner 1981, Kutas and Hillyard 1984). A commonly used method to measure lexical predictability is a cloze task in which people are asked to complete a given sentence and the probability that a particular lexical item was used as a completion is referred to as cloze probability (Taylor 1953). In order to implement the present cloze task, several measures were taken. First, 5000 transitive verbs were extracted from the Finnish Internet Parsebank. Second, these verbs were divided into three quantile groups based on frequency. Third, we sampled 50 verbs from each of the three quantile groups, i.e., 150 verbs in total. These two measures were taken to ensure that a wide range of verbs based on frequency was included in the cloze task. Fourth, for each verb we constructed subject arguments that referred to human and each subject argument was unique in the task. Fifth, all of the verbs were presented in imperfect tense, for example subject *tutkija* 'researcher' and verb *kloonasi* 'cloned'.

For a typical cloze task, participants are instructed to produce a single completion. However, it is plausible that typical transitive constructions tend not to be highly constrained lexically, reducing cloze probability. Therefore, there might be multiple possible completions for a given combination of subject and verb, thus creating noise (see Shaoul, Baayen and Westbury 2014: 440–441, for discussion). In order to reduce this potential source of noise, the participants were instructed to produce three completions for a given combination of subject and verb (see Federmeier *et al.* 2007). We will refer to these preference groups simply the first, the second and the third. Given that this procedure increases the time required to complete the experiment, the 150 verbs were first randomized and then divided into three list, each containing 50 combinations of subjects and verbs. Thus, each participants produced 150 completions. We used an on-line questionnaire to collect the completions, with each participant providing completions for a single list. In total, 69 participants (Age: $M = 28.1$, $SD = 14.3$, 12 men) from across Finland voluntarily took part in the experiment. The participants appeared to represent a diverse population as they reported 44 different birth places and 17 different current places of living and an average year of education of 17.7 ($SD = 3.42$). Each list had 23 participants and, in

total, 10,350 completions were produced. The completions were automatically lemmatized and morphologically tagged using the software OMorFi (Pirinen 2011). After which, the results were manually verified and cleaned by removing words containing typographical errors (less than 1% of the data). For the purposes of the present study, we only included in the final data set those instances that can be considered to function as objects in the transitive construction excluding, for example, adverbs. The final data set contains 9681 completions.

For the purposes of this study, we present two analyzes of the data in which the performance of the vector-based models are compared to the productions in the cloze task. The first one concentrates on the correlation between the estimated semantic fit by the models and cloze probability discussed in Section 6.1. Cloze probability reflects the probability of producing a particular lexical item for a given combination of the subject and verb. For example, in case of *isoveli asensi* 'the big brother installed' the most probable object was *lamppu* 'lamp' with a cloze probability of 0.09. The second one focuses on the frequency of producing a given object for a particular transitive construction presented in Section 6.2. In this respect, this variable can be understood as measuring subjective frequency that has been shown to influence, for example, processing times similar to objective frequency (Balota, Pilotti, and Cortese 2001). It is plausible that cloze probability does not necessarily capture production preferences in its totality for transitive constructions that are associated with low lexical predictability.

6.1. Model estimates and cloze probability

For these data, the cloze probability was calculated separately for each of the preference groups as the participants were instructed to produce three completions in order of preference. For example, the following combination of the subject and verb *lääkäri amputoi* 'the doctor amputated' was most often completed with the word *jalka* 'foot', $n = 20$, in the first preference group. Thus, the cloze probability for this realization is showing a high degree of lexical predictability for this particular combination. In general, cloze probability values between 0.7 and 0.9 are considered to indicate high lexical predictability whereas values 0.1 and less are taken to indicate low predictability. We illustrate completions associated with high and low cloze probability in Table 3. In the case of the amputate event, the produced completions indicate a

high degree of lexical specificity in the first preference group as only two lexical realizations were produced.

Table 3. High and low cloze probability completions for two transitive constructions.

lääkäri amputoi 'the doctor amputated'

Object	Cloze probability
jalka 'foot'	0.87
raaja 'limb'	0.13

puuseppä aitasi 'the carpenter enclosed'

object	Cloze probability
piha 'yard'	0.35
alue 'area'	0.13
pelto 'field'	0.9

The distribution of the cloze probability across the preference groups is visualized with a boxplot in Figure 5.

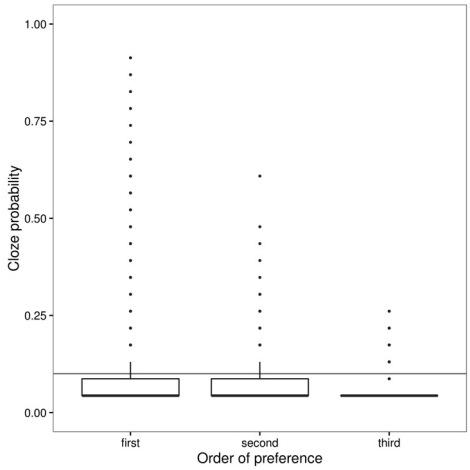

Figure 5. The distribution of the cloze probability across the three preference groups in the cloze task. The horizontal line indicates the cloze probability value of 0.1.

The distribution of the cloze probability across the preference groups indicates that most of the transitive constructions used in this experiment, do not appear to pertain to semantic domains with a high degree of lexical specificity as the mass of the probability distribution is located below the threshold value of 0.1. Only certain lexical combinations evoked a high cloze probability. These extreme values of the cloze probability are depicted with a dot in the figure. Out of the 150 verbs only 13 were associated with a cloze probability value equal to or greater than 0.7, such as *tyrehdyttää* 'suppress', *raottaa* 'open slightly' and *jynssätä* 'scrub'. Additionally, the distribution clearly brings forth the nature of the task, i.e., cloze probability steadily declines when moving from the first preference group to the third. For most transitive events there are multiple possible lexical completions for objects and only those events which appear to be associated with a higher degree of lexical specificity such as the amputation event, do we find high values of cloze probability. Consequently, this demonstrate that cloze task is an expensive task; hundreds of participants would be required to obtain stable estimates for cloze probability for transitive events in general (see Shaoul, Baayen, and Westbury 2014 for discussion).

For the purposes of the present study, we focus on the lexical completions for the objects that had the highest cloze probability in the first preference group as this set appears to be the most stable, as expected. Thus, we extracted the highest cloze probability associated with the object in a given combination of the subject and verb allowing us to evaluate the degree of correspondence between the distributional models and average subjective preference indexed by the cloze probability. Given that the BiNN is a binary classifier, we used the predicted probability for the object in a given transitive construction as a proxy for semantic fit. For the other distributional models, cosine similarity was used as a measure of semantic fit. For all four distributional models, a Pearson correlation was calculated between these measures of semantic fit and the cloze probability. The results are given in Table 4.

Table 4. Pearson correlation coefficients between the model estimates of selectional preference and cloze probability.

Model	*r*	Lower bound	Upper bound	*P*-Value
AE	0.35	0.29	0.49	< 0.001
BiNN	0.31	0.16	0.45	< 0.001
Word2vec_VO	0.29	0.14	0.44	< 0.001
Word2vec_SO	0.26	0.09	0.39	< 0.001

The results showed that all the models captured some facets of cloze probability in this task as all the estimated correlations were statistically significant. Additionally, all the correlations displayed the expected sign where an increase in cloze probability was positively correlated with an increase in semantic similarity in the distributional models and, in the case of the BiNN, with increased probability. The AE achieved the highest correlation with cloze probability compared to all other models investigated in this study. Finally, we evaluated the difference in the correlations between the AE against all the other models based on Fisher's *r*-to-*z* transformation (Cohen and Cohen 1983). Although numerically the estimated correlation with the AE was the highest, the differences were not statistically significant as all *p*-values were greater than 0.05 at the nominal α-level of 0.05. In sum, the results show that the different methods implemented in this study to model semantic similarity are correlated with cloze probability. This demonstrates that these models are able to capture, at least, certain aspects of lexical predictability.

6.2. Model estimates and cloze frequency

To further evaluate the fit of the models and the productions in a cloze task, we calculated the frequency of the lexical completions for the objects across the three preference group for a particular combination of subject and verb, for example, the frequency of the completions for the combination *tutkija kloonasi* 'the researcher cloned'. From this set, the realization with the highest frequency was extracted. We will refer to this measure as cloze frequency. Thus, the difference between these two constructs is how well they can approximate lexical predictability. It is worth pointing out that the lexical items are the same when calculated either based on cloze frequency or cloze probability.

The results indicated that the cloze frequency displayed a greater variation than the cloze probability for these transitive constructions, $M = 13.18$, $SD = 5.74$. This suggests that cloze frequency might be a better construct for constructions with lower lexical predictability.

Similar to the evaluation of cloze probability, we calculated Pearson correlations between the semantic similarity measures estimated with the implemented models and the cloze frequency. The results are given in Table 5.

Table 5. Pearson correlation coefficients between the model estimates of selectional preference and cloze frequency.

Model	*r*	Lower bound	Upper bound	*P*-Value
AE	0.45	0.31	0.57	< 0.001
BiNN	0.35	0.2	0.48	< 0.001
Word2vec_VO	0.34	0.19	0.47	< 0.001
Word2vec_SO	0.2	0.04	0.35	0,015

Similar to the results presented in Section 6.1 for cloze probability, all the estimations of semantic fit were correlated with cloze frequency and were statistically significant. Additionally, these correlations displayed the same pattern were higher cloze frequency was positively correlated with increase in semantic fit, as expected. Interestingly, the estimated correlations indicated, however, a considerably better correspondence between the models and cloze frequency in contrast to cloze probability, although the correlation with Word2vect_SO was numerically lower. It seems that cloze frequency appeared to approximate lexical preference better in comparison to cloze probability, at least for these data. Finally, we evaluated the statistical significance of the difference between the models, similar to the evaluation procedure for cloze probability. The results indicated that only the difference between the AE and the Word2vec_SO was statistically significant, $p = 0.015$, at the nominal α-level of 0.05.

6.3. Discussion

We investigated the correspondence between the implemented models and two subjective measures of lexical predictability estimated based on a cloze task. In the task, the participants were instructed to produce three completions in order of preference for a given transitive construction, for example *tutkija kloonasi* ‘the researcher cloned’. This design was implemented to obtain a large number of completions for a particular transitive construction and possibly stabilize the estimates in the task. For the purposes of the present study, two subjective measures of lexical predictability were constructed, specifically cloze probability and cloze frequency. The former was constructed based on the probability of producing a given lexical item in the first preference group and the most probable completion was used to index cloze probability.

This is the commonly used construct in experimental studies. The latter measure was calculated over all the three preference groups and simply represents the frequency of occurrence of a particular completion with a given combination of the subject and verb. The results showed that although both of the constructs selected the same lexical items, cloze frequency appeared to offer a better fit, at least for these constructions.

The analysis indicated that all the implemented models were correlated with the subjective measures of lexical predictability, although the correlation between the cloze frequency and the cosine similarity between the subject and the object was not statistically significant. Additionally, the analysis based on the correlations implied that the AE offered the best fit to these data. The difference between the AE and the BiNN, however, was not statistically significant. This is most likely an issue of statistical power and a larger number of transitive constructions would be required.

Interestingly, the results presented in the previous sections suggest that there are distributional differences between the cloze probability and the cloze frequency, although both subjective measures selected the same lexical items. This appears to be the case at least for transitive constructions associated with low lexical predictability. To gain a better understanding of the correspondence between the model estimates and the subjective measures, we visualized the distributions in Figure 6. The density plots are given on the inherent scale of a given measure; it should be noted that scaling the distributions did not influence the shape of the distributions. The cloze probability ranges between 0 and 1 and the cloze frequency is simply a count variable. The estimates of the BiNN also range between 0 and 1 as it is a binary classifier. The other neural network models are all based on cosine similarity ranging between -1 and 1. For the purposes of the present study, the important aspect is the shape of the distribution (see Griffiths *et al.* 2007 for discussion about distributions and categorization).

The visualization of the distributions brings forth the functional form learnt by the neural models. As the BiNN is a binary classifier, the mass of the distribution is located around 0.9 and 1 as all these completions are plausible and, importantly, suitable completions for these transitive constructions. In contrast, the functional form estimated with the AE appears to follow a normal distribution and the mass of the distribution is located around the value 0.5.

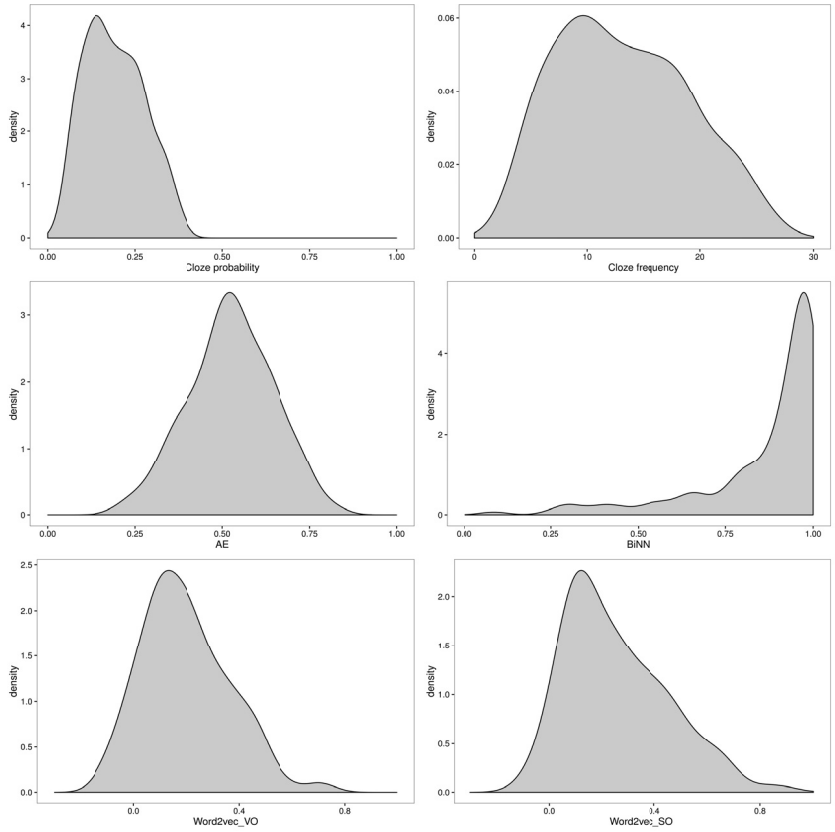

Figure 6. The distribution of the estimated lexical preferences across the models in the cloze task in comparison to cloze probability and cloze frequency. The distributions are given on the inherent scale of a given model.

Given that the AE has access to more information compared to the two word2vec models, the distribution of the cosine similarity appears to be shifted more towards 1, indicating a better completion. In terms of the subjective measures, we can see that the shape of the distribution estimated with the cloze frequency is closer to the cosine similarities. It seems that people tend to produce similar completions in a cloze task, as expected, but for less predictable completions this systematicity is not reflected in the cloze probability unless the participant pool were considerably larger. In contrast, the cloze frequency appears to offer a smooth distribution across the productions. This appears to make the fit better between the cloze frequency and the model estimates.

7. General discussion

In this study, we explored the use of semantic vectors in modeling the semantic structure of transitive constructions in Finnish. This type of argument structure constructions follows the semantics of who did what to whom. Specifically, we focused on modeling the lexical realization of the object, i.e., selectional preference, for example *tutkija kloonasi X* 'the researcher cloned X' where the X denotes the lexical realization of the object. Intuitively, it is clear that the object can be filled with a number of different lexical realizations and the semantic fit of a given realization forms a continuum. Related to this, usage-based models emphasize the role of semantic similarity in the formation of the structure of argument constructions (Bybee 2010, Goldberg 2006). Importantly, these types of models assume that semantic information and, ultimately the structuring of the mental lexicon, is shaped by experience. The role of experience is connected to the concept of distributional properties where the structuring of a given construction is connected to the context in which it occurs. These types of co-occurrence patterns form the basis of distributional models and they have a long traditional in cognitive science to model semantic information (Lund and Burgess 1996, Landauer and Dumais 1997). In this respect, distributional models of semantic structure follow the same fundamental assumptions of usage-based models. Recent developments in distributional models, however, have shifted away from counting co-occurrences to predicting them using neural networks such as the word2vec model. Here, we extended this line of investigation by implementing an autoencoder-based neural network to model selectional preference in the Finnish transitive construction. Specifically, in this model, the realization of the object in the transitive construction is achieved through mapping in semantic space. This mapping function can be viewed as forming an abstract representation for the object given the realization of the subject and the verb in the transitive construction. In this study, we took the first steps in evaluating the performance of this model in a pseudo-disambiguation and cloze task. Additionally, we contrasted the performance of the AE model relative to a binary neural classifier and word2vec.

In the pseudo-disambiguation task, the AE offered the best performance when the objects were not matched in frequency. Interestingly, both the AE and the BiNN outperformed a purely frequency-based model in this task. Importantly, people have been shown to be sensitive to differences in frequency distributions, even in the case of multi-word

units. For example, Arnon and Snider (2010) showed that people read multi-word phrases faster when the unit as a whole is more frequent (see also Tremblay and Baayen 2010 for EEG results). However, this raises the question about the size of the mental lexicon as is discussed, for example, in Baayen, Hendrix, and Ramscar (2013). If people store distributional information associated with multi-word units as such, in addition to word frequency information, the size of the mental lexicon is drastically increased. In this respect, both of the neural networks offer a more economical model of semantic memory compared to a purely frequency-based model because these types of models smooth the semantic space. In the case of the AE, the subject and the verb are represented by a layer of 200 neurons. This is an additional smoothing because the underlying word2vec model already represents the semantic space as a smooth distribution.

Interestingly, we saw a reduction in performance with the AE when the objects were matched in frequency, although naturally occurring realizations of the transitive construction are unlikely to be matched in frequency. In contrast, the BiNN displayed only a minor reduction in performance. This results indicated that the AE was substantially more sensitive to difference in the distributional properties of the input. Furthermore, the BiNN also has access to a greater amount of information compared to the AE as it was explicitly trained on negative instances. This leads to an important difference in the architecture between these two models, specifically, in terms of the conceptual basis of how semantic memory is structured. The use of negative evidence is problematic if a model is assumed mirror, at least up to a degree, the formation of semantic memory. It is highly unlikely that during language acquisition people are exposed to negative evidence to the extent that the positive and the negative evidence are balanced (see Ambridge *et al.* 2009, Bowerman 1988). From this perspective, the AE offers a cognitively more plausible model of semantic memory.

A cloze task was implemented in this study to further evaluate the performance of the models. A cloze task is often used to measure the degree of lexical predictability in a particular construction in experimental studies. Additionally, the role of cloze probability has been shown to influence reading times, for example (see Matsuki *et al.* 2011, Rayner *et al.* 2004). From this perspective, it is desirable to estimate the degree of correspondence between models of semantic fit and subjective estimates. Two subjective measures were constructed for the purposes of the present study, namely cloze probability and cloze frequency. The

results showed for low predictability constructs the cloze frequency might offer a better estimate for the degree of lexical predictability compared to the traditional measure of cloze probability. Importantly, the analysis indicated that all the implemented models captured certain aspects of these two subjective estimates. Additionally, the AE offered the best fit to the data, although the correlations were only statistically significant between the AE and the pairwise semantic similarity between the subject and object.

Intuitively, it seems clear that the frequency of use in itself cannot be the only factor driving semantic fit. In the case of the transitive construction, it is conceivable that certain realizations of the object may not be frequent but can belong to the same semantic domain as a frequent realization, for example (see Goldberg 2006, Barðdal 2008). We illustrate this possibility with the transitive construction *tutkija kloonasi* X 'the researcher cloned X' and *lääkäri amputoi* X 'the doctor amputated X'. The former appears to represent a lexically more open-ended event type compared to the latter. This is also reflected in the number of unique completions produced in the task for these combinations of the subject and the verb; see Appendix for the full list. For example, the participants produced such completions for the *tutkija kloonasi* X 'the researcher cloned X' as *ihminen* 'human', *itse* 'oneself', *koira* 'dog' and *hevonen* 'horse'. This suggests that there is more structure in the data than is actually reflected either in cloze probability or cloze frequency for these completions. To bring forth this type of structuring, we extracted the semantic vectors for these completions and visualized them using t-SNE via Barnes-Hut algorithm (Van Der Maaten 2014) and the visualization is given in Figure 7. Additionally, we included for both constructions the mapped semantic vector of the object estimated with the AE denoted with the label mapped in the Figure 7.

The similarity relations among the objects appear to form small clusters indicating the presence of semantic subdomains as have been previously proposed for argument structure constructions (Goldberg 2006, Barðdal 2008). The amputate event appears to contain two clusters and the clone event three or four. Additionally, the visualization highlights the properties of the AE. The semantic structure learnt by the AE represents low-level generalizations over event structures as advocated in constructionist approaches to argument structures (see Suttle and Goldberg 2011 for recent discussion). The mapped object represents an abstraction over the distributional properties associated with the lexical realization of the object with a particular relation of the transitive

construction. As supported by the results of this study, this type of representation of semantic information is suitable for modeling plausibility and predictability of lexical items. These results strongly support the view that the AE offers an attractive model for estimating the semantic fit of particular constructions.

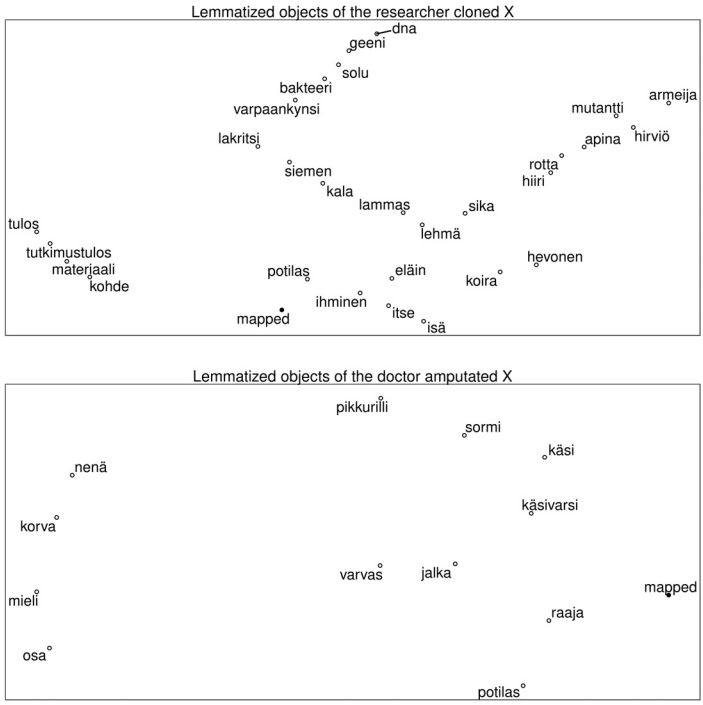

Figure 7. Visualization of the semantic similarity relations among the objects produced in the cloze task for two transitive constructions in vector space using word2vec. The mapped object of these two constructions predicted by the AE is indicated with the label mapped and a filled dot. The translations for the objects are provided in Appendix.

In this study, we have, however, only focused on the realization of the object in a transitive construction. Furthermore, we utilized only corpus and off-line data in the evaluation of the models. Off-line tasks such as the cloze task are known to be sensitive to decision processes. For example, cloze probability is known to be modulated by the frequency of the word itself, among other variables (see Smith and Levy 2011 for

discussion). To further substantiate the results presented here, the evaluation of the models should be carried out based on tasks that are not or are, at least, less sensitive to decision processes. We are currently evaluating the performance of these models relative to online processing of transitive constructions using eye-tracking. This method offers an online measure associated with processing cost that are not influenced by decision making processes. This allows us to compare the performance of the AE and the BiNN in relation to online processing. We will, however, leave this type of inquiry for future studies.

8. Conclusion

We presented a neural network, i.e., an autoencoder, for modeling the semantic structure of transitive constructions in Finnish. The model received as its input semantic vectors based on the word2vec algorithm and the architecture of this model consists of two parts. The first encodes the subject and the verb slots in a transitive construction following a standard autoencoder model architecture. The second part of the model contains a mapping function from the encoded representation of the subject and verb to the object. The mapping function can be understood as an abstract representation of the most probable object in a given realization of the transitive construction. Thus, the conceptual basis of the proposed model follows the basic premise of usage-based construction grammar by representing low-level semantic generalizations of a transitive construction. In order to evaluate the performance of this model, we concentrated on the realization of the object in a given transitive construction, i.e., on selectional preference. Two tasks were implemented to evaluate the performance of the proposed model: a pseudo-disambiguation and a cloze task in Finnish. Additionally, we compared the performance of the proposed model to a neural binary classifier that has been previously used to model selectional preferences in English, to a semantic-based similarity measure obtained from word2vec algorithm and a to a purely frequency-based n-gram model. The results showed that the AE offered the best fit to the data in the pseudo-disambiguation task and it obtained the highest correlation with human productions in the cloze task. In sum, the results presented here take the first steps towards representing the semantic structure of an argument construction in a conceptually and computationally harmonious manner.

Acknowledgements

This research was funded by the Kone Foundation and the Academy of Finland (Project 285739). We would like to thank the two anonymous reviewers and the editors for their valuable comments and suggestions. Any errors that may remain are, of course, our own.

Address
Aki-Juhani Kyröläinen
Department of Finnish and Finno-Ugric Languages
20014 Turun yliopisto
Turku
Finland
E-mail: akkyro@utu.fi

Juhani Luotolahti
E-mail: mjluot@utu.fi

Filip Ginter
E-mail: figint@utu.fi

Appendix: List of all the unique objects produced in the cloze for two subject and verb combinations

tutkija kloonasi X 'the researcher cloned X'			
Object	Translation	Object	Translation
apina	monkey	*lakritsi*	licorice
armeija	armey	*lammas*	sheep
bakteeri	bacteria	*lehmä*	cow
dna	dna	*materiaali*	material
eläin	animal	*mutantti*	mutant
geeni	gene	*potilas*	patient
hevonen	horse	*rotta*	rat
hiiri	mouse	*siemen*	seed
hirviö	monster	*sika*	pig
ihminen	human	*solu*	cell
isä	father	*tulos*	result
itse	oneself	*tutkimustulos*	finding
kala	fish	*varpaankynsi*	toenail
kohde	target		
koira	dog		

lääkäri amputoi X 'the doctor amputated X'			
Object	Translation	Object	Translation
jalka	foot	*osa*	part
käsi	hand	*pikkurilli*	little finger
käsivarsi	arm	*potilas*	patient
korva	ear	*raaja*	limb
mieli	mind	*sormi*	finger
nenä	nose	*varvas*	toe

References

Ambridge, Ben, Julian M. Pine, Caroline F. Rowland, Rebecca L. Jones, and Victoria Clark (2009) "A semantics-based approach to the 'no negative evidence' problem". *Cognitive Science* 33, 7, 1301–1316.

Arnon, Inbal and Neal Snider (2010) "More than words: frequency effects for multi-word phrases". *Journal of Memory and Language* 62, 1, 67–82.

Baayen, R. Harald, Peter Hendrix, and Michael Ramscar (2013) "Sidestepping the combinatorial explosion: an explanation of n-gram frequency effects based on naive discriminative learning". *Language and Speech* 56, 3, 329–347.

Baayen, R. Harald, Petar Milin, Filipović Đurđević Dušica, Peter Hendrix, and Marco Marelli (2011) "An amorphous model for morphological processing in visual comprehension based on naive discriminative learning". *Psychological Review* 118, 3, 438–481.

Balota, David A., Maura Pilotti, and Michael J. Cortese (2001) "Subjective frequency estimates for 2,938 monosyllabic words". *Memory & Cognition* 29, 4, 639–647.

Barðdal, Jóhanna (2008) *Productivity: evidence from case and argument structure in Icelandic*. Amsterdam: John Benjamins.

Baroni, Marco and Alessandro Lenci (2010) "Distributional memory: a general framework for corpus-based semantics". *Computational Linguistics* 36, 4, 673–721.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski (2014) "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors". In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 238–247. Available online at <http://anthology.aclweb.org/P/P14/P14-1023.pdf>. Accessed on 25.05.2017.

Bengio, Yoshua (2009) "Learning deep architectures for AI". *Foundations and Trends in Machine Learning* 2, 1, 1–127.

Bowerman, Melissa (1988) "The "no negative evidence" problem: How do children avoid constructing an overly general grammar?" In John A. Hawkins, ed., *Explaining language universals*, 73–101. Oxford: Blackwell.

Bybee, Joan L. (2010) *Language, usage and cognition*. Cambridge: Cambridge University Press.

Bybee, Joan L. and David Eddington (2006) "A usage-based approach to Spanish verbs of 'becoming'". *Language* 10, 5, 425–455.

Chang, Franklin, Gary S. Dell and Kathryn Bock (2006) "Becoming syntactic". *Psychological Review* 133, 2, 234–272.

Chollet, François (2015) *Keras*. *GitHub repository*. Available online at <https://github.com/fchollet/keras>. Accessed on 25.05.2017.

Cohen, Jacob and Patricia Cohen (1983) *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale: NJ: Erlbaum.

Croft, William (2001) *Radical construction grammar: syntactic theory in typological perspective*. Oxford: Oxford University Press.

Dagan, Ido, Lillian Lee, and Fernando C. N. Pereira (1999) "Similarity-based models of word cooccurrence probabilities". *Machine Learning* 34, 1, 34–69.

Durda, Kevin and Lori Buchanan (2008) "WINDSOR: Windsor improved norms of distance and similarity of representations of semantics". *Behavior Research Methods* 40, 3, 705–712.

Efron, Bradley and Robert J. Tibshirani (1993) *An introduction to the bootstrap*. New York: Chapman & Hall.

Ehrlich, Susan F. and Keith Rayner (1981) "Contextual effects on word perception and eye movements during reading". *Journal of Verbal Learning and Verbal Behavior* 20, 6, 641–655.

Erk, Katrin, Sebastian Padó, and Ulrike Padó (2010) "A flexible, corpus-driven model of regular and inverse selectional preferences". *Computational Linguistics* 36, 4, 723–763.

Federmeier, Kara D., Edward W. Wlotko, Esmeralda Ochoa-Dewald, and Marta Kutas (2007) "Multiple effects of sentential constraint on word processing". *Brain Research* 1146, 75–84.

Firth, John Rupert (1957) "A synopsis of linguistic theory 1930–1955". In *Studies in linguistic analysis*. Oxford: Blackwell. Reprinted in Frank R. Palmer, ed. *Selected papers of J.R. Firth (1952–59)*, 168–205. London and Bloomington: Longman and Indiana University Press.

Goldberg, Adele E. (1995) *Constructions: a construction grammar approach to argument structure*. Chicago: Chicago University Press.

Goldberg, Adele E. (2006) *Constructions at work. The nature of generalization in language*. Oxford: Oxford University Press.

Griffiths, Thomas L., Kevin R. Canini, Adam A. Sanborn, and Daniel J. Navarro (2007) "Unifying rational models of categorization via the hierarchical Dirichlet process". In *Proceedings of Cognitive Science Society*, 323–328.

Harris, Zellig (1951) *Methods in structural linguistics*. Chicago: University of Chicago Press.

Hinton, Geoffrey E. and Richard S. Zemel (1994) "Autoencoders, minimum description length, and Helmholtz free energy". In D. Cowan, G. Tesauro and J. Alspector, eds. *Advances in neural information processing systems 6*, 3–10. Available online at <https://papers.nips.cc/paper/798-autoencoders-minimum-description-length-and-helmholtz-free-energy.pdf>. Accessed on 21.04.2017.

Kalyan, Siva (2012) "Similarity in linguistic categorization: The importance of necessary properties". *Cognitive Linguistics* 23, 3, 539–554.

Kanerva, Jenna, Matti Luotolahti, Veronika Laippala, and Filip Ginter (2014) "Syntactic N-gram collection from a large-scale corpus of Internet Finnish". In *Proceedings of the sixth international conference Baltic HLT*, 184–191. <doi:10.3233/978-1-61499-442-8-184>

Kurtz, Kenneth J. (2007) "The divergent autoencoder (DIVA) model of category learning". *Psychonomic Bulletin & Review* 14, 4, 560–576.

Kutas, Marta and Steven A. Hillyard (1984) "Brain potentials during reading reflect word expectancy and semantic association". *Nature* 307, 161–163.

Landauer, Thomas K. and Susan T. Dumais (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.

Lenci, Alessandro (2011) "Composing and updating verb argument expectations: a distributional semantic model". In *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics*, 58–66. Available online at <http://dl.acm.org/citation.cfm?id=2021103>. Accessed on 21.04.2017.

Levy, Omer and Yoav Goldberg (2014) "Dependency-based word embeddings". In *Proceedings of the ACL* (2), 302–308. Available online at <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.722.3988&rep=rep1&type=pdf>. Accessed on 21.04.2017.

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

Mareschal, Denis, Robert M. French, and Paul C. Quinn (2000) "A connectionist account of asymmetric category learning in early infancy". *Developmental Psychology* 36, 5, 635–645.

Matsuki, Kazunaga, Tracy Chow, Mary Hare, Jeffrey L. Elman, Christoph Scheepers, and Ken McRae (2011) "Event-based plausibility immediately influences on-line language comprehension". *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37, 4, 913–934.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013) "Efficient estimation of word representations in vector space". In *Proceedings of workshop at ICLR*. Available online at <https://arxiv.org/abs/1301.3781>. Accessed on 25.05.2017.

Pirinen, Tommi (2011) "Modularisation of Finnish finite-state language description – towards wide collaboration in open source development of a morphological analyser". *In Proceedings of the 18th Nordic Conference of Computational Linguistics*, 299–302. Available online at <https://www.aclweb.org/anthology/W/W11/W11-46.pdf#page=313>. Accessed on 25.05.2017.

Ramscar, Michael, Peter Hendrix, Cyrus Shaoul, Petar Milin, and Harald Baayen (2014) "The myth of cognitive decline: non-linear dynamics of lifelong learning". *Topics in Cognitive Science* 6, 1, 5–42.

Rayner, Keith, Timothy J. Slattery, Denis Drieghe, and Simon P. Liversedge (2011) "Eye movements and word skipping during reading: effects of word length and predictability". *Journal of Experimental Psychology: Human Perception and Performance* 37, 2, 514–528.

Rayner, Keith, Tessa Warren, Barbara J. Juhasz, and Simon P. Liversedge (2004) "The effect of plausibility on eye movements in reading". *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30, 6, 1290–1301.

Rumelhart, David E. and James L. McClelland (1986) "On learning the past tenses of English verbs". In David E. Rumelhart and James L. McClelland, eds. *Parallel distributed processing: explorations in the microstructure of cognition*, 216–271. Cambridge: MIT Press.

Shaoul, Cyrus and Chris Westbury (2010) "Exploring lexical co-occurrence space using HiDEx". *Behavior Research Methods* 42, 2, 393–413.

Shaoul, Cyrus, R. Harald Baayen, and Chris F. Westbury (2014) "N-gram probability effects in a cloze task". *The Mental Lexicon* 9, 3, 437–472.

Smith, Nathaniel J. and Roger Levy (2011) "Cloze but no cigar: the complex relationship between cloze, corpus, and subjective probabilities in language processing". In *Proceedings of the 33rd annual conference of the Cognitive Science Society*, 1637–1642. Available online at <https://vorpus.org/papers/smith-levy-2011-cogsci.pdf>. Accessed on 25.05.2017.

Suttle, Laura and Adele E. Goldberg (2011) "The partial productivity of constructions as induction". *Linguistics* 49, 6, 1237–1269.

Taylor, Wilson L. (1953) "Cloze procedure: a new tool for measuring readability". *Journalism and Mass Communication Quarterly* 30, 4, 415–433.

Tremblay, Antoine and R. Harald Baayen (2010) "Holistic processing of regular four-word sequences: a behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall". In D. Wood, ed. *Perspectives on formulaic language: Acquisition and communication*, 151–173. London; New York: Continuum.

Turney, Peter D. (2006) "Similarity of semantic relations". *Computational Linguistics* 32, 3, 379–416.

Van de Cruys, Tim (2014) "A neural network approach to selectional preference acquisition". In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 26–35. Available online at <http://www.aclweb.org/old_anthology/D/D14/D14-1004.pdf>. Accessed on 25.05.2017.

Van Der Maaten, Laurens (2014) "Accelerating t-SNE using tree-based algorithms". *Journal of Machine Learning Research* 15, 1, 3221–3245.

Yarowsky, David (1993) "One sense per collocation". In *Proceedings of the ARPA human language technology workshop*, 266–271. Available online at <http://dl.acm.org/citation.cfm?id=1075731>. Accessed on 25.05.2017.

Kokkuvõte. Aki-Juhani Kyröläinen, M. Juhani Luotolahti ja Filip Ginter: Autokoodril põhinev närvivõrkude mudel valikulisel eelistamisel. Intuitiivselt tundub, et mõned argumendid sobivad teatud predikaatidega paremini kokku kui teised. Kasutuspõhised keelemudelid rõhutavad konstruktsioonide struktuuri (nii vormi kui tähenduse) kujunemisel tähendusliku sarnasuse olulisust. Selles uurimuses modelleerime soome keele transitiivsete konstruktsioonide semantikat ja esitame närvivõrkude mudeli ehk autokoodri. Mudel põhineb distributiivse semantika hüpoteesil, mille järgi kujuneb semantiline info peamiselt konteksti põhjal. Täpsemalt keskendume uurimuses objektile. Mudelit hindame nii valeühestamise kui ka lünkülesande abil. Kõrvutame autokoodri tulemusi varem välja töötatud neurovõrgumudelitega ja tõestame, et meie mudel töötab võrreldes teiste mudelitega väga hästi. Tulemused esitame kasutuspõhise konstruktsioonigrammatika kontekstis.

Võtmesõnad: neurovõrk, autokooder, tähendusvektor, kasutuspõhine mudel, soome keel