Capturing Anomalies in the Choice of Content Words in Compositional Distributional Semantic Space

Ekaterina Kochmar
Computer Laboratory
University of Cambridge
ek358@cl.cam.ac.uk

Ted Briscoe
Computer Laboratory
University of Cambridge
ejb@cl.cam.ac.uk

Abstract

In this work, we present a new task for testing compositional distributional semantic models. Recently, there has been a spate of research into how distributional representations of individual words can be combined to represent the meaning of phrases. Vecchi *et al.* (2011) have shown that some compositional models, including the additive and multiplicative models of Mitchell and Lapata (2008; 2010) and the linear map-based model of Baroni and Zamparelli (2010), can be applied to detect semantically anomalous adjective–noun combinations. We extend their experiments and apply these models to the combinations extracted from texts written by learners of English.

Our work contributes to the field of compositional distributional semantics by introducing a new test paradigm for semantic models and shows how these models can be used for error detection in language learners' content word combinations.

1 Introduction

Vector-based (*distributional*) models are widely used for representing the meaning of single words. They rely on the assumption that word meaning can be learned from the linguistic environment and can be approximated by a word's *distribution* across contexts. Words are represented as vectors in a high-dimensional space, with vector dimensions encoding word co-occurrence with contextual elements – other words within a local window, words linked by specific dependencies to the target word, and so forth. Distributional models provide a clear basis for interpreting word meaning, as well as a simple means for measuring semantic similarity. These properties have been exploited in many NLP tasks, including automatic thesaurus extraction (Grefenstette, 1994), word sense induction (Schütze, 1998) and disambiguation (McCarthy et al., 2004), collocation extraction (Schone and Jurafsky, 2001) and others.

In contrast to single words, the distribution of phrases cannot be used as a reliable approximation of their meaning, as phrase vectors are much sparser. Irrespective of the size of the corpus considered, some content word combinations will remain unattested as a consequence of their Zipf-like distributions. For example, Vecchi *et al.* (2011) have shown that both semantically acceptable and semantically deviant word combinations will be absent from large English corpora. A promising alternative is to use *compositional* models which combine distributional vectors for the component words in some way, for example, using a direct vector combination function (Kintsch, 2001; Mitchell and Lapata, 2008; Erk and Padó, 2008; Thater et al., 2010) or linear transformations on vectors (Baroni and Zamparelli, 2010).

In spite of the spate of recent work in this area, the question of how to combine word representations is far from answered. Compositional models can be assessed by their ability both to provide a solid theoretical basis for meaning composition and to represent composite meaning for relevant practical tasks. Promising results have been shown with such models on similarity detection and paraphrase ranking (Mitchell and Lapata, 2008; Erk and Padó, 2008; Thater et al., 2010), adjective–noun vector prediction (Baroni and Zamparelli, 2010) and semantic anomaly detection (Vecchi et al., 2011). Of these tasks, the latter appears to be particularly challenging since it addresses the ability of compositional models to account for linguistic productivity.

No corpus can effectively sample all possible content word combinations. On the other hand, some corpus-attested word combinations may appear semantically deviant when considered out of context (for example, when they are used metaphorically). Vecchi *et al.* (2011) have focused on unattested adjective–noun (AN) combinations and noted that if a combination does not occur in a corpus, it may be due to various reasons including data sparsity as well as nonsensicality. The task of distinguishing between the two cases is challenging. Vecchi *et al.* use the following examples:

(1) a. *blue rose*
b. *residential steak*

Whereas both may well be unattested in a corpus, the concept of *blue rose* is perfectly conceivable while that of *residential steak* is nonsensical and only interpretable in specifically-constructed discourse contexts. Vecchi *et al.* argue that there should be a detectable difference between the model-generated representations for the semantically deviant combinations and those for the acceptable ones, and assess compositional models by their ability to capture this difference. Vecchi *et al.* have created a set of corpus-unattested AN combinations, annotated them as semantically acceptable or deviant, and applied the *additive (add)* and *multiplicative (mult)* models of Mitchell and Lapata (2008) and *adjective-specific linear maps (alm)* of Baroni and Zamparelli (2010).

Given that promising results have been obtained in their experiments, we propose that a useful extension of this task is to test the compositional models on errors in content word combinations extracted from texts written by learners of English. This task provides a natural setting for testing semantic models on genuine examples and is a potential practical application for such models.

Language learners' errors are diverse, but many of them can naturally be explained in terms of nonproductive, semantically anomalous combination of content words (Leacock et al., 2010). Learners may lack robust intuitions about words' selectional preferences and subtle differences in meaning, so they may confuse near-synonyms, overuse words with broad meaning, and otherwise choose words inappropriately. Consider the following examples extracted from our data:

(2) a. **big importance* vs *great importance*
b. **economical crisis* vs *economic crisis*
c. **deep regards* vs *kind regards*
d. *best moment* vs *best time*

These examples illustrate that learner errors can often be explained by confusions stemming from similar meaning (*2a*) or form (*2b*). When a word combination appears to be nonsensical as in *2c*, the words chosen might still be related to the appropriate ones in the learner's mental lexicon. We recognise that although error detection in learners' content word combinations is a natural extension to semantic anomaly detection, it also poses additional difficulties that semantic models might not be able to deal with. For example, some erroneous word combinations may not be completely devoid of compositional meaning, while violating language conventions. However, semantic models might still be able to capture some of these conventions. Another challenge is that some expressions cannot be unambiguously classified as either correct or incorrect, as their interpretation depends on the context of use: *best moment* (*2d*) is appropriate when used to denote a short period of time, but it is often incorrectly used by learners instead of *best time*.

To make our work comparable with previous work on semantic anomaly, we investigate AN combinations extracted from texts written by non-native speakers of English, and apply the *add*, *mult* and *alm* models of semantic composition. The main contributions of this work are to show that error detection in content word combinations provides a natural testbed and useful application for the compositional distributional models, and that the results obtained on this task provide a more natural estimate of the models' performance than ones based on artificially constructed examples. If the compositional distributional models can distinguish between correct and incorrect content word combinations, these models can then be used for writing or pedagogical assistance. To the best of our knowledge, this is the first attempt to handle learner errors in the choice of content words using compositional distributional semantics.

Plan of the paper. We overview related work on error detection and discuss the three models of semantic composition in Section 2. Section 3 presents the data and experimental setup. We discuss the results of our experiments in Section 4 and conclude in Section 5.

2 Related Work

2.1 Error Detection in Content Words

Research on error detection has mostly been concerned with function words, such as determiners and prepositions (Leacock et al., 2010; Dale et al., 2012). Such errors are more frequent, but they are also more systematic which makes them easier to detect. Function words constitute a closed class, so the set of possible corrections is also limited. By comparison, errors in content word combinations pose a bigger challenge. Since content words primarily express meaning rather than encode syntax, detection and correction of such errors depend on a system's ability, in the limit, to recognise the communicative intent of the writer. Moreover, the set of possible corrections is much larger than for function words.

Previous work has either focused on correction alone assuming that errors are already detected (Liu et al., 2009; Dahlmeier and Ng, 2011), or has reformulated the task as *writing improvement* (Shei and Pain, 2000; Wible et al., 2003; Chang et al., 2008; Futagi et al., 2008; Park et al., 2008; Yi et al., 2008). In the former case error detection, which is a difficult task in itself, is not addressed, while in the latter case it is integrated into that of suggesting alternatives according to some metric (for example, frequency or mutual information). In some cases, a database of typical errors in word combinations is collected from learner texts and suggestions are only made for these error-prone combinations. Otherwise suggestions will be made for many acceptable phrases.

In this work, we treat error detection in the choice of content words as an independent task and assess the ability of compositional distributional models to discriminate incorrect from correct AN combinations – a frequent source of error in learner texts.

2.2 Composition by Component-wise Operations

In the *additive* and *multiplicative* compositional models of Mitchell and Lapata (2008; 2010), the components of the composite vector are obtained by component-wise operations applied to the word vectors. If **c** is a word combination vector and **a** and **b** are word vectors, then **c**'s i-th component is the sum of the i-th components of **a** and **b** for the *add* model:

$$c_i = a_i + b_i \qquad (1)$$

and the product of the corresponding components for the *mult* model:

$$c_i = a_i b_i \qquad (2)$$

An advantage of using these models is that they provide a clear and simple interpretation of vector composition, requiring no training or tuning. They have also been shown to be promising models of composition in a number of NLP tasks, including semantic anomaly detection (Vecchi et al., 2011). However, the principal weakness of these models is that they use commutative operations, and therefore fail to represent the difference in the grammatical function of the component words, their order, and "headedness". For example, these models would produce the same composite vectors for *component vector* and *vector component*.

In addition, the *add* model does not take "incompatibility" of constituent vectors along individual dimensions into account. If one vector has a high value in its i-th dimension while another vector has 0, the composed vector will receive the high value from the first input vector, even though, intuitively, this dimension should get 0 or near-0 value. This problem does not arise with the *mult* model. On the other hand, the *mult* model is heavily biased towards dimensions with high values in both input vectors (Baroni et al., 2012).

2.3 Distributional Functions and Linear Maps

The *adjective-specific linear maps* of Baroni and Zamparelli (2010) take the grammatical functions of the words within a combination into account. Focusing on AN combinations, they try to model the fact that adjectives modify nouns and the resulting combination is nominal. They note that the meaning of nouns can be represented with their distributional vectors, but the meaning of attributive adjectives cannot be fully captured by their distribution alone: for example, *new* in *new friend* is not the same as *new* in *new shoes*. The meaning of the adjective *new* is defined through its application to the denotations of the nouns. Therefore, Baroni and Zamparelli (2010) suggest treating adjectives as *distributional functions* that map between semantic vectors representing nouns to ones representing AN combinations.

Within this approach, adjectives are represented with weight matrices. The composition is defined by matrix-by-vector multiplication as follows:

$$f(noun) =_{def} \mathbf{F} \times \mathbf{a} = \mathbf{b} \qquad (3)$$

where **F** is the matrix representing an adjective and encoding function *f*, which maps the input noun vector **a** to the output AN vector **b**. The ij-th cell of the matrix contains the weight determining how much the component corresponding to the j-th context element in the noun vector contributes to the value assigned to the i-th context element in the AN vector (Baroni et al., 2012). These weights are estimated separately for each adjective from all corpus-observed noun–AN vector pairs using (multivariate) partial least squares regression.

3 Experimental Setup

3.1 Test Data

We have extracted a set of AN combinations from the publicly available CLC-FCE dataset (Yannakoudakis et al., 2011), a subset of the Cambridge Learner Corpus (CLC),[1] which is a large corpus of texts produced by English language learners sitting Cambridge Assessment's examinations.[2]

These texts have been manually error-coded (Nicholls, 2003). Using the error annotation, we have divided extracted ANs into two subsets – correctly used ANs and those that are annotated with error codes due to inappropriate choice of an adjective or/and noun.[3] For the ANs that are used correctly in some contexts and incorrectly in others we use the most frequent annotation from the data.

Our test set contains 4681 correct and 530 incorrect combinations. In contrast to Vecchi *et al.* (2011), who have used a limited set of constituent adjectives and nouns and an approximately equal number of semantically acceptable and deviant combinations, our test set is more skewed towards correct combinations and consists of a wider range of constituent words. It also includes ANs occurring in the BNC[4] – 3294 of the correct test ANs and 256 of the incorrect ones are corpus-attested. The set of corpus-attested ANs annotated as incorrect in our data includes low-frequency combinations from the BNC, as well as combinations whose error-annotation depends on context. We believe that this test set reflects practical applications of semantic anomaly detection more closely.[5]

3.2 Semantic Space Construction

In constructing the *semantic space* we follow the procedure outlined in Vecchi *et al.* (2011). We populate the semantic space with a large number of distributional vectors for the *target elements* – constituent nouns and adjectives from the test ANs, and the most frequent nouns and adjectives from a corpus of English as well as AN combinations of these words. To estimate the frequency rankings, we use a concatenation of two well-formed English corpora – the 100M word BNC and the Web-derived 2B word ukWaC corpus.[6]

The semantic space is represented by a matrix encoding word co-occurrences, with the rows representing the target elements and the columns representing a set of 10K *context words* consisting of 6,590 nouns, 1,550 adjectives and 1,860 verbs most frequent in the combined corpus. The ij-th cell of the original matrix contains a sentence-internal co-occurrence count of the i-th target element with the j-th context word. The raw sentence-internal co-occurrence counts from the original matrix have been transformed into Local Mutual Information scores (Baroni and Zamparelli, 2010; Evert, 2005).

An interesting research question is how much data are needed to obtain reliable word co-occurrence counts. We estimate the word co-occurrence statistics using the BNC only, and leave it for future research to explore the impact of estimating them from larger corpora, for example, the ukWaC or the concatenated corpus mentioned above. We lemmatise, tag and parse the data with the RASP system (Briscoe et al., 2006; Andersen et al., 2008), and extract all statistics at the lemma level.

The target elements are selected as follows: we first select the 4K adjectives and 8K nouns which are most frequent in the concatenated corpus. In each case, we exclude the top 50 most frequent words since those may have too general meanings.

Next, we extract the constituent adjectives and nouns from our test data and populate the semantic space with the words not yet contained in it. As a result, our semantic space contains 8,364 nouns.

Since we aim at investigating AN behaviour in a highly-populated semantic space, we add more AN combinations to that. We select 218 very frequent adjectives (occurring more than 100K but

[1] http://www.cup.cam.ac.uk/gb/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus/

[2] http://www.cambridgeenglish.org

[3] The corresponding error codes are `RJ` and `RN`.

[4] http://www.natcorp.ox.ac.uk/

[5] The examples extracted for our experiments are publicly available at http://www.cl.cam.ac.uk/~ek358/.

[6] http://wacky.sslmit.unibo.it/

less than 740K times) and merge them with the adjectives from the test ANs. We generate all possible AN combinations by crossing this combined set of adjectives and the set of 8,364 nouns. This results in a set of ANs of which 1,6M combinations are corpus-attested. From these we randomly choose 62,205 ANs that occur more than 100 times in the corpus. As a result, we populate our semantic space with ANs with the number of unique corpus-attested combinations per adjective ranging from 1 to 1,226 and being 84.52 on average. Since we apply our approach to real data, we cannot avoid having a different number of training examples for different adjectives. It is worth exploring how many training examples are needed for a single adjective, since some highly frequent adjectives may have more training examples in the data, while some adjectives may require more training examples than others due to polysemy or lack of strong selectional preferences.

Finally, we check our test set against the combined corpus and add 1,131 test ANs which are corpus-attested but not yet contained in the semantic space. Our final semantic space consists of 8,364 nouns, 4,353 adjectives and 63,336 corpus-attested ANs.

We perform all operations on vectors in the full semantic space, using a 76,053 × 10K matrix. We leave it for future research to perform dimensionality reduction (for example, using Singular Value Decomposition) and to compare the results with the ones reported here.

3.3 Composition Methods

For the *add* and *mult* models, the AN vectors are obtained by component-wise addition and multiplication without normalisation. For the *alm* model, the weight coefficients are estimated with multivariate partial least squares regression using the R `pls` package (Mevik and Wehrens, 2007), using the leave-one-out training regime. This model is computationally expensive since a separate weight matrix must be learned for each adjective and since we use the non-reduced semantic space. Therefore, for the experiments presented here we limit the number of test adjectives to 38. The selected adjectives are, on the one hand, frequently misused by language learners, and, on the other, have a manageable number of training examples. The reduced set of test ANs consists of 347 combinations.

The number of latent variables used by the training algorithm depends on the number of available noun–AN training pairs. We have gradually changed this number from 3 to 20 depending on the adjective and the number of available training pairs with the aim of keeping the independent-variable-to-training-item ratio stable. However, we have not optimised this number and leave it for future research.

3.4 Measures of Semantic Anomaly

Once the composite vectors are obtained, the next question is how to distinguish between the vectors for correct and anomalous combinations. Vecchi *et al.* (2011) propose three simple measures for distinguishing between the two sets of vectors:

1. **Vector Length (VLen)**: they hypothesise that vectors for anomalous ANs are shorter than those for acceptable ones. Since the distributional vectors encode word occurrence, words that do not "match" semantically should have their co-occurrence counts distributed differently along the dimensions, and their composition is expected to have many near-0 values.

2. **Cosine with the Noun Vector (CosN)**: they hypothesise that in nonsensical ANs the meaning of the input nouns is degraded and their model-generated vectors are situated further away from the original noun vectors. For example, since a *big dog* is still a *dog* and an **extensive dog* is less clearly so, in the semantic space the vector for *big dog* would be closer to that of *dog* than the vector for **extensive dog* to *dog*. Semantically deviant ANs are expected to have lower cosine between their vectors and the original noun vectors.

3. **Density of the AN Neighbourhood (Dens)**: it is hypothesised that deviant ANs will have fewer close neighbours and be more "isolated" in the semantic space. This is measured by the average cosine with the top 10 nearest neighbours, which is assumed to be lower for anomalous ANs.

We hypothesise that some cues alternative to the ones already proposed may also be effective:

1. **Cosine with the Adjective Vector (CosA)**: since both *add* and *mult* models are symmetric and both input vectors contribute to the

Measure	*all*	*attest*	*unattest*
VLen	0.1992	0.6226	0.1840
CosN	0.0797	0.1538	0.00001$^{(*)}$
Dens	0.9792	0.3921	0.5589
CosA	0.6867	0.3790	0.0026$^{(*)}$
RDens	0.6915	0.7493	0.1414
Num	0.8756	0.5753	0.1050
COver	0.6028	0.2126	0.1200

Table 1: p values for the *add* model

Measure	*all*	*attest*	*unattest*
VLen	0.0033$^{(*)}$	0.1549	0.0004$^{(*)}$
CosN	0.0017$^{(*)}$	0.0182$^{(*)}$	0.0083$^{(*)}$
Dens	0.3531	0.6656	0.2703
CosA	0.00002$^{(*)}$	0.0144$^{(*)}$	0.3352
RDens	0.0002$^{(*)}$	0.0300$^{(*)}$	0.0001$^{(*)}$
Num	0.0001$^{(*)}$	0.0091$^{(*)}$	0.0001$^{(*)}$
COver	0.0041$^{(*)}$	0.0096$^{(*)}$	0.7317

Table 2: p values for the *mult* model

output combination equally, we also measure the distance to the original adjective vector.

2. **Ranked Density (RDens)**: we define *close proximity* to the model-generated AN vector as the neighbourhood populated with vectors for which the cosine to the AN vector is higher than 0.8. Since the number of close neighbours is different for different ANs, we measure *ranked density* as $\sum_{i=1}^{N} rank_i \, distance_i$, where N is the number of neighbours.

3. **Number of Neighbours within Close Proximity (Num)**: the number of close neighbours itself can be used as a measure.

4. **Component Overlap (COver)**: we assume that AN combinations, unless they are idiomatic, are similar to the constituent words or combinations with the same constituents. The models can be assessed by their ability to place the AN vector in the neighbourhood populated by similar words and combinations. We measure this as the proportion of nearest neighbours containing same constituent words as in the tested ANs.

4 Results

We use the measures described above and compute the difference between the mean values for the correct and incorrect model-generated ANs. We apply the unpaired t-test, assuming a two-tailed distribution, to assess the statistical significance of the difference between these values. In Tables 1 to 3 we report p values estimating statistical significance at the 0.05 level, and statistical significance is marked with an asterisk $(*)$.

We assume that there might be a difference between the corpus-attested and corpus-unattested test ANs, with each of the subgroups being more homogeneous than the entire test set. Our corpus-unattested examples are more similar to the ANs considered by Vecchi *et al.* (2011). We report the results on the full set of test ANs, as well as on each of the two subgroups separately.

Our goals are to:

- comparatively evaluate performance of the three composition models;
- assess the appropriateness of the proposed metrics;
- investigate models' performance on the corpus-attested and corpus-unattested combinations.

4.1 Comparative Performance of the Models

Of the three composition models, the *mult* model (Table 2) shows the best results overall.

The *alm* model (Table 3) shows statistically significant difference between the model-generated vectors for the correct and incorrect combinations with the cosines and component overlap, but it does not detect the difference on the corpus-unattested subset with any of the metrics.

The *add* model (Table 1) shows statistically significant differences only with the cosine measures on the corpus-unattested subset. The poor performance of this model may be due to its weaknesses outlined in Section 2.2. Also, Baroni and Zamparelli (2010) note that normalisation may help improving its performance.

4.2 Appropriateness of the Metrics

Cosines to the original input vectors show promising results with all three models. In contrast to the results reported by Vecchi *et al.* (2011), the density of the semantic neighbourhood does not differ significantly with any of the models, but since

Measure	*all*	*attest*	*unattest*
VLen	0.6537	0.2840	0.5557
CosN	0.00003$^{(*)}$	0.0003$^{(*)}$	0.1555
Dens	0.8160	0.4902	0.1799
CosA	0.0188$^{(*)}$	0.0070$^{(*)}$	0.8440
RDens	0.9106	0.6804	0.8588
Num	0.5959	0.9619	0.1402
COver	0.00001$^{(*)}$	0.0004$^{(*)}$	0.1484

Table 3: p values for the *alm* model

AN	bad intention	*bad information
add	***bad***, ***bad*** *company*, ***bad*** *image*	***information***, *other* ***information***, *real* ***information***
mult	*uncomplicated*, *improbable*, *suggestive*	*uncomplicated*, *improbable*, *humane*
alm	***intention***, *main* ***intention***, *real* ***intention***	*people*, *blind people*, *like-minded*

Table 4: Top 3 neighbours for each model

many of the combinations tested in our experiments are not genuinely anomalous, the fact that they are situated in densely populated semantic neighbourhoods is not surprising. Measures based on close proximity neighbourhood – RDens and Num – show statistical difference when applied to the *mult*-generated vectors only.

With COver, the *alm* model, followed by the *mult* model, produce sensible results. Table 4 shows the top 3 nearest neighbours found by the models for the correct AN *bad intention* and the incorrect **bad information*. The latter is annotated as incorrect since its meaning is quite vague and a possible correction is *inaccurate information*. Note that only the *alm* model is able to discriminate between the correct and the incorrect word combinations suggesting sensible nearest neighbours for *bad intention* and less sensible ones for **bad information*.

4.3 Attested vs Unattested Combinations

Our results show that the models perform differently on the two subsets and somewhat better on corpus-attested ANs. However, the results also confirm that appropriate models and metrics can be found to distinguish between correct and incorrect ANs in both subsets.

5 Conclusion

In this paper we have introduced a new task on which compositional distributional semantic models can be tested. Our results support the hypothesis that semantic models can be applied to detect errors in the choice of content words by English language learners. The original contribution of our paper is to show how compositional and distributional semantics can be linked to error detection to provide a solution to a practical task.

Our results suggest that with the metrics considered it is easier to detect the difference between the model-generated vectors for the correct and incorrect word combinations with the *multiplicative* model. On the other hand, qualitative analysis suggests that the *adjective-specific linear maps* of Baroni and Zamparelli (2010) are superior, since they place the model-generated vectors in semantically sensible neighbourhoods.

We plan to investigate further whether the use of a bigger corpus for collecting word co-occurrence statistics provides more reliable counts, and whether dimensionality reduction and/or normalisation of the models improves the results. We also plan to apply the *alm* model to a larger number of examples. Some other models such as the ones by Erk and Padó (2008) and Thater *et al.* (2010) which take selectional preferences and context into account may yield better results on this task, and we plan to test this experimentally in the future. Finally, since these models can discriminate between correct and anomalous combinations, the next step is to incorporate them into an error detection classifier.

Acknowledgments

We are grateful to Cambridge ESOL, a division of Cambridge Assessment, and Cambridge University Press for supporting this research and for granting us access to the CLC for research purposes. We would like to thank Helen Yannakoudakis, Øistein Andersen and the anonymous reviewers for their valuable comments.

References

Andersen Ø., Nioche J., Briscoe T. and Carroll J. 2008. *The BNC parsed with RASP4UIMA*. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC).

Baroni M., Bernardi R. and Zamparelli R. 2012. *Frege in Space: A Program for Compositional Distributional Semantics*. http://clic.cimec.unitn.it/composes/materials/frege-in-space.pdf

Baroni M. and Zamparelli R. 2010. *Nouns are vectors, adjectives are matrices: Representing adjective–noun constructions in semantic space*. In Proceedings of the EMNLP-2010, pp. 1183–1193.

Briscoe E., Carroll J., and Watson R. 2006. *The Second Release of the RASP System*. In Proceedings of the COLING/ACL-2006 Interactive Presentation Sessions, pp. 59–68.

Chang Y.C., Chang J.S., Chen H.J., and Liou H.C. 2012. *An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology*. Computer Assisted Language Learning, 21(3):283–299.

Dahlmeier D. and Ng H.T. 2011. *Correcting Semantic Collocation Errors with L1-induced Paraphrases*. In Proceedings of the EMNLP-2011, pp. 107–117.

Dale R., Anisimoff I., and Narroway G. 2012. *HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task*. In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 54–62.

Erk K. and Padó S. 2008. *A Structured Vector Space Model for Word Meaning in Context*. In Proceedings of the EMNLP-2008, pp. 897–906.

Evert S. 2005. *The Statistics of Word Cooccurrences*. Dissertation, Stuttgart University.

Futagi Y., Deane P., Chodorow M., and Tetreault J. 2009. *A computational approach to detecting collocation errors in the writing of non-native speakers of English*. Computer Assisted Language Learning, 21(4):353–367.

Grefenstette G. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

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

Leacock C., Chodorow M., Gamon M. and Tetreault J. 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan and Claypool Publishers.

Liu A. L.-E., Wible D., and Tsao N.-L. 2009. *Automated suggestions for miscollocations*. In Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 47–50.

McCarthy D., Koeling R., Weeds J. and Carroll J. 2004. *Finding predominant word senses in untagged text*. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pp. 280–287.

Mevik B. and Wehrens R. 2007. *The pls package: Principal component and partial least squares regression in R*. Journal of Statistical Software, 18(2).

Mitchell J. and Lapata M. 2008. *Vector-based models of semantic composition*. In Proceedings of ACL, pp. 236–244.

Mitchell J. and Lapata M. 2010. *Composition in distributional models of semantics*. Cognitive Science, 34:1388–1429.

Nicholls D. 2003. *The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT*. In Proceedings of the Corpus Linguistics conference, pp. 572–581.

Park T., Lank E., Poupart P., and Terry M. 2008. *Is the sky pure today? AwkChecker: an assistive tool for detecting and correcting collocation errors*. In Proceedings of the 21st annual ACM symposium on User interface software and technology, pp. 121–130.

Schone P. and Jurafsky D. 2001. *Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem?*. Pittsburg, PA, pp. 100–108.

Schütze H. 1998. *Automatic word sense discrimination*. Computational Linguistics, 24(1):97–123.

Shei C.C. and Pain H. 2000. *An ESL Writer's Collocation Aid*. Computer Assisted Language Learning, 13(2):167–182.

Thater S., Fürstenau, H., and Pinkal M. 2010. *Contextualizing Semantic Representations Using Syntactically Enriched Vector Models*. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 948–957.

Vecchi E., Baroni M. and Zamparelli R. 2011. *(Linear) maps of the impossible: Capturing semantic anomalies in distributional space*. In Proceedings of the DISCO Workshop at ACL-2011, pp. 1–9.

Wible H., Kwo C.-H., Tsao N.-L., Liu A., and Lin H.-L. 2003. *Bootstrapping in a language-learning environment*. Journal of Computer Assisted Learning, 19(4):90–102.

Yannakoudakis H., Briscoe T. and Medlock B. 2011. *A New Dataset and Method for Automatically Grading ESOL Texts*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1:180–189.

Yi X., Gao J., and Dolan W.B. 2008. *A Web-based English Proofing System for English as a Second Language Users*. In Proceedings of the third International Joint Conference on Natural Language Processing (IJCNLP), pp. 619–624.