A Geometric Method for Detecting Semantic Coercion

Stephen McGregor
Queen Mary University of London
s.e.mcgregor@qmul.ac.uk

Elisabetta Jezek
University of Pavia
jezek@unipv.it

Matthew Purver
Queen Mary University of London
m.purver@qmul.ac.uk

Geraint Wiggins
Queen Mary University of London
geraint.wiggins@qmul.ac.uk

Abstract

In this paper we present state-of-the-art results on the computational classification of semantic type coercion, accomplished using a novel geometric method which is both context-sensitive and generalisable. We show that this method improves accuracy on a SemEval dataset over previous work, and gives promising results on a new more challenging experimental setup involving the same data. In addition to a description of our distributional semantic methodology and the results obtained on an established dataset, we offer an overview of the linguistic phenomenon of coercion and an analysis of the geometric features by which our results are achieved.

1 Introduction: Computers and Language in Context

Computers are notoriously literal devices. Provided that communication remains grounded in straightforward propositional expressions about named entities with categorical properties involved in unambiguously labelled processes, a computer has some hope of tracking the development of a linguistic exchange. In the pragmatic domain of natural language, however, we are never far from a slide into the webs of implication and inference that characterise communication between environmentally situated agents, capable of resorting to assumptions of isomorphic conceptual schemes in order to optimise the *quality*, *efficiency*, and *relevance* of linguistic constructs (Grice, 1975; Wilson and Sperber, 2012).

The computational representation of contextual shifts in lexical semantics presents a particularly significant challenge, in that it, at first glance, requires the establishment of a rule-based system for indicating the open-ended ways in which rules may be broken. Approaches have typically relied on the construction, in some way or another, of categorical conceptual representations – ontologies – designed for the transfer of properties between classes. So, for instance, motivated by the cognitive linguistic *conceptual metaphor* model of Lakoff and Johnson (1980), Shutova (2013) uses clustering techniques to define classes based on a statistical analysis of dependency relationship in a parsed corpus, and then uses class transgressions in verb-noun relationships to detect metaphor. Alternatively, Veale and Hao (2008) draw inspiration from the *conceptual blending* work of Fauconnier and Turner (2003) in their description of a system that combines information extracted from the WordNet knowledge base with statistical corpus analysis in order to treat metaphor as the porting of categorical information between conceptual domains. Of particular relevance to the research presented here is the work of Shutova et al. (2013), who likewise use a combination of corpus analysis and knowledge base extraction to predict classes of words in order to identify instances of logical metonymy.

Notwithstanding the impressive results generated by these and other similar models, they tend to require a certain degree of preprocessing, annotation, or often direct access to an existing knowledge base in order to achieve effective semantic extrapolation and are prone to falling short of the truly exponential compositionality that characterises natural language. As an alternative, we propose a method for building geometric semantic representations which, in their infinitely adaptable spatial situation, mirror the versatility of language in use. Our method offers three crucial features. First, it is context sensitive, in

that it dynamically generates a subspace and a corresponding array of semantic relationships in response to online linguistic input, including the words being modelled as well as their sentential context where available. Second, it is generalisable, in that a straightforward classification model built on a relatively small training set can subsequently be applied to any given linguistic input, regardless of whether any of the words involved were observed in the training data. Third, it is built on unannotated raw textual data, extrapolating semantic relationships using distributional semantic techniques (see Clark, 2015, for an overview). In particular we have refrained from adorning our representations with information derived from, for instance, dependency parsing, allowing us to present a model that avoids downstream commitments regarding the cognitive role of grammatical class (Langacker, 1991).

We apply our method to a task involving the classification of semantic type coercion, a linguistic phenomenon which will be described in detail in the next section. The section following that will present our methodology for building a base space of co-occurrence dimensions from unlabelled data and then selectively projecting subspaces of these dimensions in order to contextually analyse the semantic relationships between words. Section 4 will describe the results of a logistic regression model applied to the geometric features generated by our subspace projection technique, trained on the labelled coercion data described in Section 2. Section 5 will analyse these results, examining the way that the typical geometry of semantic relationships shift as they move from selectional to coerced uses.

2 Background: Coercive Verbs, Susceptible Nouns

Coercion as a theoretical tool has been used in linguistic studies to account for several kinds of semantic shifts occurring in different linguistic structures. For example, *aspectual type coercion* (Moens and Steedman, 1988) identifies the shift occurring when a predicate denoting an event type is coerced to a different type by contextual triggers, as in (1a), where the punctual adverb *suddenly* coerces the predicate *know* from State to Transition. *Grinding* in the nominal domain (Copestake and Briscoe, 1995) consists of a mass construal of a count noun, as for *pillow* in (1b), which is coerced to mass by the quantifier *some*. Finally, *coercion by construction* (Michaelis, 2004) identifies a shift in the meaning of a verb as a result of its insertion in a specific construction, as in the causative construction in (1c).

(1) a. She suddenly knew it.
    b. Give me some pillow.
    c. He barked them back to work.

In this paper, we focus on semantic coercion in predicate-argument combination, intended as the compositional mechanisms that resolves an apparent mismatch between the semantic type expected by a predicate for a specific argument position (in one or more of its specific senses, should the predicate be polysemous) and the semantic type of the argument filler, by adjusting the type of the argument to satisfy the type requirement of the function (*argument type coercion*; Pustejovsky, 1991). An example is (2), where *wine* is coerced to an Activity (drinking) as a result of the semantic requirements the predicate imposes on its object, i.e. *finish* applies to an activity.[1]

(2) When they finished the wine, he stood up. (drinking)

In predicate-argument composition, the semantics of the argument plays a crucial role in two ways. First, it provides the semantic purport on which selection or coercion may apply;[2] second, in the presence of a coercion environment, it constrains the resulting interpretation. While the default interpretation of (2) is "drinking", the one in (3) is "eating"; in other words, different nouns grant privileged access to different activities, particularly those which are most frequently performed with the entities they denote.

[1] Such cases of coercion to event are referred to as *logical metonymies* (see Verspoor, 1997, and Lapata and Lascarides, 2003).

[2] Coercions are not always successful; that is, some predicate-argument combinations are not interpretable. Constraints on interpretability are clearly related to cognition and the way we conceptualize entities and relations among them, an aspect we will return to later in the paper.

(3) They finished their cake. (*drinking, eating)

It has been noted, however, that linguistic and situational contexts play a crucial role in the interpretation of coercions: for example, in the corpus fragment in (4), taken from the EnTenTen corpus, the context triggers a different interpretation for wine (preparing, making) as object of *finish*. In other words, the "reconstructed hidden event" may be assigned contextually.

(4) So unless the winemakers add tannin by finishing the wine in oak ...

Extensive corpus work on both English and Italian data (Pustejovsky and Jezek, 2008; Jezek and Quochi, 2010, *inter alia)* has shown that coercion in predicate-argument composition is particularly frequent with certain verb classes, including event-selecting verbs (attend, cancel, organize) of which aspectual verbs constitutes a subclass (finish, interrupt, start, continue), perception verbs (hear, listen), communication verbs (announce, inform), directed motion verbs (arrive, reach), and verbs indicating motion performed using a vehicle (land).

Data on mismatches between expected type and argument type offer several options of linguistic modelling. Pustejovsky (2011) for example proposes a two-layered coercion mechanism: *coercion by exploitation* takes an available part of the argument's type (modelled as *quale* to the type) to satisfy the function, whereas *coercion by introduction* wraps the argument with the type required by the function (for example, in "the passengers read the walls of the subway", read wraps the walls with an informational content, which is present in the selecting type but absent in the argument type). Asher (2011), on the other hand, acknowledging the role played by discourse context in the interpretation of mismatches, uses dependent types to model coercion. Both authors assume, in addition to the Montague types, e and t, a richer subtyping over the entity domain than is typically assumed in type theory, including complex types such as the one associated with *book*, which comprises a physical as well as an informational component.

Coercion detection has been addressed as a specific NLP task in the context of SemEval 2010[3], with the goal of testing the ability of computational models to identify whether the type that a verb selects is satisfied directly by the argument (selection), or whether the argument must change type to satisfy the verb typing (coercion), and classify it accordingly.[4] A dataset was produced for both English and Italian, using the methodology described by Pustejovsky et al. (2010).[5] First, five coercive verbs that impose semantic typing on one of their arguments in at least one of their senses (*arrive*, *cancel*, *deny*, *finish*, and *hear*) were selected by examining the data from the BNC, using the Sketch Engine corpus query tool. Sense inventories were compiled for each verb using OntoNotes as a reference. For each sense, a set of type templates was identified following the Corpus Pattern Analysis (CPA) technique (Pustejovsky et al., 2004; Hanks, 2013): every argument in the syntactic pattern associated with a given sense was assigned a type specification. The coercive senses of the chosen verbs were associated with type templates. Type templates and senses for the five verbs are summarized below:

(5) a. HUMAN arrive at LOCATION (reach a destination or goal)
b. HUMAN cancel EVENT (call off)
c. HUMAN deny PROPOSITION (maintain that something is untrue)
d. HUMAN finish EVENT (complete an activity)
e. HUMAN hear SOUND (perceive physical sound)

A set of sentences was randomly extracted for each target verb from the BNC. The extracted sentences were parsed automatically, and organized according to the grammatical relation the target verb was involved in. Word sense disambiguation of the predicate was performed manually on each extracted sentence, matching it against the sense inventory and the corresponding type template. The appropriate senses were then saved into the database along with the associated type template. The sentences containing coercive senses of the verbs were annotated for selection or coercion in the specified grammatical

[3] A metonymy resolution task not focused on verb-argument composition is described in Markert and Nissim (2009).

[4] Complex types and the distinction between exploitation and introduction as described above are not included in the task.

[5] For the purposes of this paper, we focus on the English data set in the following.

Source Type	Target Type	Verb	Train	Test
event	location	arrive	38	37
artifact	event	cancel	35	35
		finish	91	92
event	proposition	deny	56	54
artifact	sound	hear	28	30
event	sound	hear	24	26
document	event	finish	39	40

Table 1: Coercion Shifts in the English SemEval data set

relation (object). Only the six most recurrent coercion types were selected; these are reported in Table 1. Examples of annotated data tagged as coercions are given in (6).

(6) a. Mr Templeton said that when he arrived at the *fire* after 10 pm ... (Event → Location)
b. Her *milk* and *newspapers* will have to be cancelled. (Artifact → Event)
c. I can hear that *car* like it is just going past here. (Artifact → Sound)

The distribution of selectional and coercive instances were skewed to increase the number of coercions. The final English data set contains about 30% coercions. The data set was randomly split in half into a training set and a test set. The training data has 1032 instances, 311 of which are coercions, whereas the test data has 1039 instances, 314 of which are coercions. Of the 1992 sentences used in our tests (see Section 4), there were 20 unique surface forms for the 5 verbs analysed and 697 objects.

3 Methodology: Projecting Semantic Context

In this section, we describe a method for projecting distributional semantic subspaces based on contextual input in the form of a word or groups of words. Our hypothesis is that there should be a way to classify the coerciveness in a verb-object pairing in terms of the absolute and relative geometric features of the corresponding word-vectors in a subspace delineated in terms of a set of co-occurrence dimensions salient to the context in which the pairing arises. The intuition underlying this hypothesis is that there should be a distinction between the co-occurrence profiles of verbs and objects selected by the verb's argument class (expressed in the form of type specification) versus objects coerced by the same verb's expected argument class, and that this distinction should be particularly evident in the context of co-occurrences relevant to the conceptual domain indicated by the word pairing.

The particular methodology we propose has been developed from work originally described by Agres et al. (2015) and McGregor et al. (2015), and early versions of the subspace selection techniques outlined here have been applied by Agres et al. (2016) to a metaphoricity rating task. We begin by building a base space of word co-occurrence statistics, using a typical pointwise mutual information metric for representing the expectedness of observing a co-occurrence term c within n words of a target word w. This results in a co-occurrence matrix where the dimension corresponding to c for word-vector $\overrightarrow{w}$ is determined as follows:

$$PMI_{w,c} = \log_2 \left(\frac{f_{w,c} \times W}{f_w \times (f_c + a)} + 1 \right) \tag{1}$$

Here, $f_{w,c}$ is the number of times w and c are observed to co-occur, f_w is the independent frequency of w, f_c is the frequency of c, W is the total count of word tokens in the corpus, and a is a smoothing constant to avoid the proliferation of obscure dimensions in our dimension selection process, set here at 10,000. The ratio is incremented by 1 to ensure that all values are positive: a PMI score of 0 then corresponds to no observed co-occurrences between w and c.

This base matrix is very large – for the model applied here, in which 200,000 vocabulary word-vectors were extrapolated from an analysis of the English language component of Wikipedia, there are approximately 7.5 million unique co-occurrence types and corresponding dimensions – and very sparse

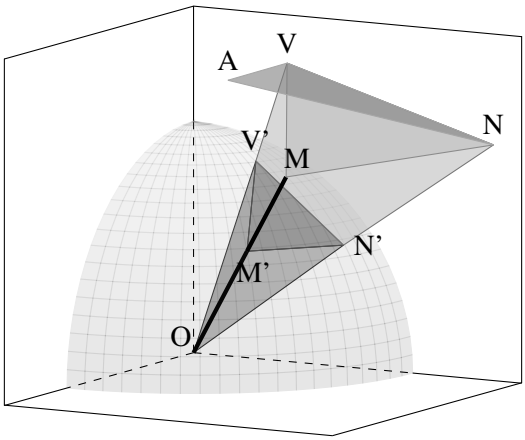


Figure 1: Semantics in Space: Verb-object pairs are projected into a subspace in which the geometric features of the relationship between the word-vectors, the origin, and salient points in the subspace are expected to collectively indicate semantic relationships such as coerciveness.

due to the long tail of relatively rare word types. From this base matrix, we project subspaces based on an analysis of the word-vectors corresponding to a group of input terms. Our objective is to discover a mechanism for identifying a set of co-occurrence features which is in some sense salient to these input terms, the idea being that semantic properties of relevant words should be apparent in their geometric situation in such a subspace. For the purposes of the experiments reported here, we explore three different subspace selection techniques:

Joint For input terms T, select the k co-occurrence dimensions that have non-zero values for all terms and the highest mean PMI values across all terms;

Indy For each term in T, select the $k/|T|$ dimensions that have the highest value for each term independent of other terms and combine them to form a k dimensional subspace;

Zipped From the subset of dimensions with non-zero values for all terms in T, select the $k/|T|$ terms with the highest value for each term, again combining for a k dimensional subspace.

For the purposes of the experiments described in this paper, we analyse the geometric relationship between word-pairs in a projection in order to determine the properties of each word-vector's situation in a space which correspond to instances of coercion. The geometric features we explore are illustrated in Figure 1, where V represents the position of the verb in a subspace and N the noun, M is the point representing the mean value for all non-zero word-vectors on each dimension, and A is the point representing the maximum value found on each dimension in the subspace. V', N', and M' are normalised vectors of V, N, and M respectively, and thus sit on the surface of a hypersphere emanating from the origin O. The features we examine are the lengths of $\overline{VO}$ and $\overline{NO}$ (the norms of each word-vector in the pair), the distances $\overline{VN}$ and $\overline{V'N'}$, the mean values of the pairs $(\overline{VO}, \overline{NO})$, $(\overline{V'M'}, \overline{N'M'})$, $(\overline{VM}, \overline{NM})$, and $(\overline{VA}, \overline{NA})$, as well as the ratios of the elements of each of those pairs, dividing the smaller constituent by the larger. We also examine the angles $\angle VON$, $\angle V'M'N'$, $\angle VMN$, and $\angle VAN$.

Our objective is to establish mechanisms for systematically gauging the geometric relationships between the word-vectors corresponding to word-pairs, as well as the relative relationships between these word-vectors and some anchor points within a given subspace. With regard to these anchor points, it is important to note that, unlike typical distributional semantic methods which build normalised spaces through either the factorisation of a matrix of co-occurrence statistics (Baroni and Lenci, 2010; Pennington et al., 2014) or the application of neural networks for the learning of abstract word-vectors across iterations of a corpus (Mikolov et al., 2013), our spaces are not normalised, and so there may be considerable variance in terms of the distribution of values across different dimensions. Our case is that, in non-normalised context-specific subspaces, we should be able to find a richer range of geometric features

with which to analyse various semantic properties of words relevant to the specific context determining a given projection.

In fact, the contextually indifferent nature of co-occurrence based models subjected to principal component analysis (Lebret and Collobert, 2014), the aforementioned neural network models, and hybrid models applying both word counting and neural network techniques (Pennington et al., 2014) are a motivation for the model we describe in this paper. While these established methodologies have achieved impressive results on a variety of language processing tasks, the representations composing them are static and abstract, and are therefore not susceptible to the online influence of contextual factors at play in our dimension selection techniques. Our case is that, for a phenomenon such as coercion, we require, as Pustejovsky (1995) has put it, "a model of meaning in language that captures the means by which words can assume a potentially infinite number of senses in context, while limiting the number of senses actually stored in the lexicon," (ibid, p. 104). As a point of comparison, we will also present results from the `word2vec` model of Mikolov et al. (2013) trained on the same underlying corpus as our models. We also test models derived from a principal component analysis of one of our base co-occurrence spaces, applying a version of the standard singular value decomposition technique in order to build a matrix of abstract dimensions optimally capturing the statistical variance between features of word-vectors.

4 Results: Detecting Coercion in a SemEval Dataset

We train a model for the identification of coercion based on a logistic regression of features of the subspaces described in the previous section. We generate JOINT, INDY, and ZIPPED type subspaces for each verb-object pair in the training portion of the dataset described in Section 2 (Pustejovsky et al., 2010), extracting the 16 geometric features identified in Section 3, illustrated in Figure 1, and enumerated again in Table 5 in Section 5. We also experiment with three other feature extraction techniques:

Verb Select only the k co-occurrence dimensions with the highest values for the verb's word-vector;

Object Select only the k co-occurrence dimensions with the highest value for the object's word-vector;

Merged Take the average feature values for the VERB and OBJECT methods.

In the case of each subspace selection technique, we generate a 993 x 16 matrix, expressing 16 geometric features for each sentence in the training data (38 examples were withheld because the targeted argument was a multi-word token, and at this point our model has only been trained for single words). We perform mean-zero, std-one normalisation on this matrix, and then perform a logistic regression trained to classify selectionality versus coerciveness. We apply L2 regularisation to the regression coefficients, with a relatively strong regularisation strength of 1.67, determined experimentally.[6] We then similarly extract data from the testing data (here 40 examples are withheld), in this case, crucially, normalising the data reusing the mean and standard deviation from the training data in order to test the generality of this method and our ability to apply it arbitrarily to any given input. We apply the model learned from the training data to the normalised test matrix, evaluating each verb-object pair as either coercive or non-coercive. We experiment with models based on co-occurrence windows of both 2 and 5 words on either side of a vocabulary word as observed in the underlying corpus (Wikipedia), and with projected subspaces consisting of 20 and 200 dimensions.

Results for these experiments are reported in Table 2, with the 200 dimensional subspaces outperforming the 20 dimensional subspaces across the board, and the 5x5 word co-occurrence window models generally doing better, but only slightly better, than the 2x2 window models. The INDY subspace selection technique outperforms all other techniques, and its strong performance is particularly pronounced in terms of f-scores, indicating that this method, in addition to learning that most instance of word-pairs are not coercive, is also learning something about when to positively indicate coercion. The stronger performance of higher dimensional spaces suggests that significant information is available across a wider

[6] We implement the regression using the `scikit-learn` LogisticRegression module for python.

	JOINT	INDY	ZIPPED	VERB	OBJECT	MERGED
2x2, 20	0.484/0.761	0.564/0.776	0.464/0.753	0.546/0.764	0.494/0.752	0.539/0.766
2x2, 200	0.537/0.793	0.631/0.795	0.524/0.778	0.630/0.800	0.598/0.789	0.632/0.801
5x5, 20	0.463/0.763	0.536/0.765	0.519/0.776	0.571/0.775	0.482/0.755	0.521/0.765
5x5, 200	0.577/0.801	0.652/0.804	0.556/0.786	0.623/0.799	0.543/0.764	0.626/0.802

Table 2: F-score/accuracy results for coercion classification using various subspace selection techniques, adjusting parameters for co-occurrence window size (2x2 and 5x5) and subspace dimensionality (20 and 200). Baseline scores and scores from other studies are reported in Table 3.

range of co-occurrence profiles for a given target word, and inclusion of this information is desirable, but it's also interesting to note this dimensional gain deteriorates for smaller co-occurrence windows as information in our base matrix becomes sparser. We use the top performing 5x5 co-occurrence window, 200 dimensional subspaces in the rest of our experiments below.

In order to test the hypothesis that coercion is always ultimately contextually determined, we add information about the sentential context of the examples provided in the data. We do this by parsing each sentence in the data and then creating two additional sets of features: we generate new subspaces based on the other words in the sentence, and then extract features of the verb/object vector geometry as above; we do this first using only content words (other verbs, nouns, adjectives, and adverbs in the sentence), and then using only function words. We extract the geometric features from these spaces as described above, normalise them, and then concatenate them with the original features extracted using the corresponding technique. In the rare instances where no appropriate sentential analysis is available, we concatenate a feature vector of zeros, reasoning that, given the application of zero-mean normalisation, this should have relatively little impact on our model while maintaining the shape of the data. Results for the logistic regression experiment run on this enhanced data are reported in Table 3. The results from the INDY type space in particular are notable in that they outperform a number of other methods which we will now describe, and moreover return an improvement in accuracy on the non-contextual results of 0.014 and in f-score of 0.021. More generally, while accuracy scores don't admit significant improvement, f-scores are generally up in the range of about 0.040 points, indicating a particular increase in the models' abilities to detect coercion with increased contextual data.

We report a minority class baseline where all verb-object pairs are classified as coercive and a majority class where all are considered non-coercive. We also test an *example based learning* method in which we learn a single rule for each surface form of the five verb stems found in the data, and discover that fairly good results can be achieved by simply assuming a given verb is either coercive or not. (In practice, all verbs other than *finish* are observed to be typically non-coercive in the training data.) Because many of the objects also occur multiple times in both the training and testing data, we can learn an object-based rule for guessing coercion, resorting to the verb-based rule in cases where we encounter an object which hasn't been observed in the training data. The very strong results achieved using this method, designated EBL* in Table 3, which take tagged observations of word combinations into account, can be thought of as something of a ceiling for models such as ours: where the EBL and EBL* methods learn to predict semantic relationships between priorly observed words based on the actual identity of the words, our method simply learns something about the geometry that indicates a particular semantic relationship.

We also report results from two models defined by static lexical representations: a principle component model built using singular value decomposition,[7] and a model constructed using the skip-gram methodology described by Mikolov et al. (2013).[8] In the case of the former, we factorised our 5x5 word co-occurrence window base space and extrapolated a 200 dimensional matrix in which each dimension is orthogonal, capturing an optimal degree of variance between word-vectors (see Deerwester et al., 1990, for a classic overview of this approach). For the latter, we built a likewise 200 dimensional space of word-vectors derived over 10 traversals of our corpus, applying negative sampling at a rate of 10. In

[7]Implemented through the python scikit-learn `TruncatedSVD` module, `http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html`.

[8]Implemented using the `gensim` package for python, `https://radimrehurek.com/gensim/`.

	prec	rec	f-score	acc		prec	rec	f-score	acc
JOINT	0.687	0.562	0.619	0.794	MINORITY	0.297	1.000	0.458	0.297
INDY	0.727	0.626	0.673	0.819	MAJORITY	0.000	0.000	0.000	0.703
ZIPPED	0.672	0.532	0.594	0.784	EBL	*0.630*	*0.498*	*0.556*	*0.764*
VERB	0.694	0.572	0.627	0.798	EBL*	*0.833*	*0.690*	*0.755*	*0.871*
OBJECT	0.636	0.529	0.577	0.770	R&H 2010	-	-	-	*0.961*
MERGED	0.708	0.562	0.627	0.801	R&H 2011	-	-	-	0.812
SVD	0.673	0.253	0.368	0.740	SKIP-GRAM	0.682	0.511	0.584	0.781

Table 3: Coercion identification scores on test data, based on a logistic regression on various dimension selection techniques in a 5x5 word co-occurrence window, 200 dimensional model built from training data, as well as scores for baselines. Methods using information about the identity of words priorly observed in selectional or coercive relationships are reported in italics.

both cases, we consider cosine distance between the word-vectors in the spaces as the singular metric of relationships between words, in line with results reported through the NLP literature.

The method described by Roberts and Harabagiu (2010) learns classes for nouns based on analysis of entailment relationships within WordNet. Combined with a statistical analysis of word and named entity co-occurrences, this approach essentially seeks to recapitulate the semantic class information available in knowledge bases in order to identify instances where coercion is indicated by verb-object class mismatches. We take as our main point of comparison the results reported on this dataset by Roberts and Harabagiu (2011), who develop a probabilistic model for coercion detection based within the latent Dirichlet allocation paradigm (Blei et al., 2003). In this later work the authors establish probability distributions for classes that can be taken as an argument by a verb V, and likewise for classes that can be assigned to an object N, and then calculate the summation of the joint probabilities of V taking a word of the same class as N as an argument, learning a threshold below which the value of this summation indicates coercion. The distributions themselves are learned through observations of predicate-argument pairings in a large-scale textual corpus, and so one might argue that here, again, there is an element of example based learning.

To briefly compare our different dimension selection techniques, the INDY technique seems to do the best job of capturing the semantic interaction between verb-object pair under analysis: the way that these terms intermingle across independently salient co-occurrence dimensions is most predictive of the alignment of semantic classes, while delineating subspaces based on joint or semi-joint co-occurrence profiles through the JOINT and ZIPPED techniques is less informative. In general the tendency towards stronger precision versus recall results indicates a tendency of our regression model to learn caution in predicting the minority class, an observation which may indicate future directions for experimenting with modelling techniques. It's also interesting to note that the VERB technique, focusing on the co-occurrence profile of the predicate in a sentence, outperforms the argument-oriented OBJECT technique, arguably supporting the hypothesis outlined in Section 2 that certain verbs tend to be more coercive than others. In terms of comparison with results from elsewhere, we significantly outperform baselines and event the EBL technique on all counts, and do slightly better than Roberts and Harabagiu (2011) on accuracy (f-scores weren't provided by those authors).

In terms of comparing with the fully recorded statistics for the abstract distributional approaches, it is interesting to note that, like with our context sensitive models, the static models also achieve higher precision than recall. In fact, the effect is even more obvious here, leading to relatively low f-scores as lower recall drags down the harmonic mean of model performance: combined with fairly high accuracy scores, this suggests that these models are learning a conservative strategy of favouring the more likely classification of selection over coercion. The stronger performance of the neural network skip-gram model over the SVD model is in line with the impressive results the `word2vec` paradigm has achieved in tests across the field, though Levy and Goldberg (2014) have made an interesting case for the commensurability of neural network and matrix factorisation techniques, attributing apparent differences in performance to the effects of the tuning of the many parameters associated with these types of models.

	prec	rec	f-score	acc
INDY 5X5 200	0.689	0.561	0.618	0.716
MINORITY	0.410	1.000	0.582	0.410
MAJORITY	0.000	0.000	0.000	0.590

Table 4: Coercion identification scores based on a logistic regression on the INDY selection technique in a 5x5 word co-occurrence window, 200 dimensional model, as well as scores for baselines, when the model was tested on words which were never seen in the training phase.

Regardless, the results of our experiment present context sensitive approaches in a relatively favourable light compared to two other general approaches to lexical semantic modelling.

Testing on Unseen Examples In order to test the generalisability of our approach, we reshuffled the data in such a way that the model could be trained on one set of verb-object pairs and then could be tested on a different set of word pairs where neither the verbs nor the objects had been observed in any of the pairings throughout the training data. The new arrangement of the data would be uninterpretable to the EBL techniques and the method of Roberts and Harabagiu (2010), all of which rely on prior observations of the words being analysed tagged for either selection or coercion. We found that by taking all objects paired with forms of the verbs *arrive*, *cancel*, and *deny* that weren't also paired with forms of the verbs *finish* and *hear* as training data, and then considering all pairings involving *cancel* and *deny* as test data, we could reshuffle the data such that we have 895 training sentences, 191 of which are instances of coercion, and 865 test sentences, 355 of which are instances of coercion. In order to maintain the generality of our results, we once again normalise the test data based on the mean and standard deviation of the training data.

Results for this version of the test are reported in Table 4. Accuracy scores are affected negatively by this data reshuffling, though this decrease should be understood in the context of the new balance of non-coercion and coercion in the data, likewise reflected in the new baselines—and in fact the improvement from our method over the majority class accuracy score is at least as substantial here. More notably, f-scores are also negatively impacted, but the effect here is considerably more marginal. From this we can infer that, in the case of classifying data on completely unseen word pairs, the model to some extent learns to usually err on the side of guessing for the majority class of argument type selection over coercion, but the gains in identifying coercion over the minority class baseline are still significant, and accuracy is likewise substantially improved from both baselines. In other words, in the case of the INDY subspace projection technique, the model seems to generalise very nicely.

5 Analysis: Interpreting the Geometry from Selection to Coercion

Table 5 presents the coefficients corresponding to geometric features learned by our logistic regression on the 5x5 co-occurrence window, 200 dimensional projections using the INDY method to analyse verb-object pair input, in this case without taking sentential context into account, as concatenating different contexts would complicate the visual analysis of the geometry of the subspace. An examination of these coefficients reveals the geometric tendencies that correspond to the slide from selection to coercion. One interesting outcome of the projection of coercion classification onto a logistic curve is the implication that coercion is a gradable as opposed to a binary phenomenon, something which is not necessarily taken for granted in the theoretical literature. Our regression is modelled to associate coercion with positive values and selection with negative values, so a positive coefficient indicates a positive correlation with the tendency towards coercion in a given subspace. The angular values used in the model are cosines, so a positive correlation here indicates a move towards coercion as the angle between two vectors becomes smaller.

The mean of the distances from the verb and object word-vectors to the maximal point (ie, the average length of $\overline{VA}$ and $\overline{NA}$) has a strong negative correlation with coercion, which, along with the positive

DISTANCES & ANGLES				MEANS & RATIOS			
$\overline{VN}$	$\overline{V'N'}$	$\overline{VO}$	$\overline{NO}$	$\mu(\overline{VO}, \overline{NO})$	$\mu(\overline{V'M'}, \overline{N'M'})$	$\mu(\overline{VM}, \overline{NM})$	$\mu(\overline{VA}, \overline{NA})$
-0.131	0.757	0.090	0.594	0.558	0.680	0.032	-0.949
$\angle VON$	$\angle V'M'N'$	$\angle VMN$	$\angle VAN$	$\overline{VO} : \overline{NO}$	$\overline{V'M'} : \overline{N'M'}$	$\overline{VM} : \overline{NM}$	$\overline{VA} : \overline{NA}$
-0.594	-0.824	-0.018	0.980	0.379	0.298	0.367	-0.191

Table 5: Coefficients assigned to various geometric features based on a logistic regression of a 5x5 word co-occurrence window, 200 dimensional INDY type space.

correlation with the cosine of $\angle VAN$, suggests a tendency for the verb and object word-vectors to move outwards and away from each other even as N moves towards M in increasingly coercive contexts. This trend suggests something about the overall dimensional profiles selected in more coercive cases: as M moves away from the central region of the space and the distance of N from the origin increases, we get a picture of a set of co-occurrence dimensions with less aligned distributions between verbs and objects, indicating lower overall frequencies and a propensity for co-occurrence with other likewise less frequent terms. In other words, in the case of nouns in particular, we find that less frequent, less ambiguous, more specialised nouns are also more prone to coercion.

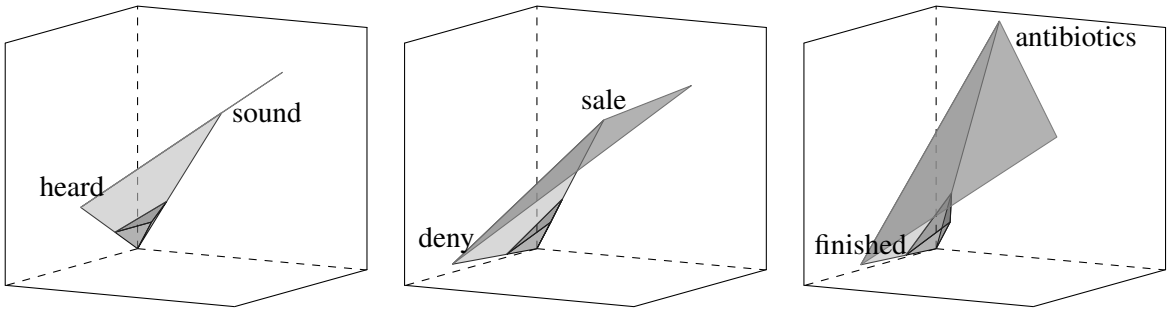
There is, conversely, a strongly positive correlation between the cosine of the normalised word-vectors at the vertex of the mean vector $\angle V'M'N'$, accompanied by a positive correlation with the distance $\overline{V'N'}$ and, at the same time, the average values of $\overline{V'M'}$ and $\overline{N'M'}$, indicating a broadening and a move again away from one another and also in this case away from the mean-adjusted centre of the subspace as the semantic context of the usage becomes more coercive. These statistics regarding effectively angular relationships between normalised vectors suggest a dimension-by-dimension divergence in the relative values of the analysed word-vectors as the INDY method selects increasingly uncorrelated dimensions for increasingly coercive semantic relationships, without necessarily saying anything about the overall trend of the length of the word-vectors in the overall subspace.

The negative correlation with the cosine $\angle VON$ tells a similar story: more coercive words tend to select co-occurrence subspaces in which the orientation of the corresponding word-vectors are less aligned. It is interesting to note, however, a likewise negative, albeit relatively minor, correlation with the actual distance between the vectors $\overline{VN}$. The immediate implication of an angle between vectors increasing even as the distance between them decreases is that the lengths of the word-vectors are shrinking, but this assumption is actually contradicted by the positive correlation with $\mu(\overline{VO}, \overline{NO})$, the average length of the word-vectors. Instead, it appears that the relative lengths of the word-vectors are actually growing closer to one another in coercive instances, moving towards a point where one vector is more optimally close to the other for a given angle. This suggests that more coercive subspaces are actually defined by dimensions for which the words in question have more distinctive profiles, and once again implies that nouns more susceptible to coercion tend to be more specialised and less ambiguous, in turn contributing a set of dimensions that are more conceptually specific, with a sparser distribution of higher PMI values by way of their half of the INDY dimensional selection process.[9]

Figure 2 illustrates cases from three points along the spectrum from selection to coercion, based on an analysis of just the verb-object pair modelled in a 5x5 co-occurrence window 200 dimensional space. Each of the subfigures shows each word-vector concerned projected into a three-dimensional space, along with the intersects of the normalised word vectors, V' and N', in their relationship to the normalised mean point M' and the maximal point A. These examples are extracted from the testing data based on the logistic regression method described above, and the geometries and the figures above have been projected to preserve the most predictive relationships $\angle VAN$, $\mu(\overline{VA}, \overline{NA})$, and $\angle V'M'N'$ while also maintaining the distances of V, N, and A from the origin, taking M' as central to the space.

The example where coercion is considered to be absent, the pairing *heard sound*, is unambiguously

[9]It should be noted that there could also be a degree of collinearity at play here, and there is grounds for experimenting with regularisation strengths and techniques in future work, as well as the application of a feature selection process involving something like a variance inflation factor (O'Brien, 2007).



(a) selectional (b) ambiguous (c) coercive

Figure 2: Spectrum of Coercion: Verb-object pairs deemed most selectional, ambiguous, and coercive are projected into subspaces using the INDY technique, with key geometric features preserved here.

an instance of a sound verb selecting a sound argument. Other instances at this end of the spectrum as construed by our model include the likewise straightforward *finished event* and *arrived port*. In the neutral area, effectively defined as the pairs whose geometric features are closest to 0.5 when passed through the softmax function, we observe the pairing *deny sale*, and note that *deny* here has an ambiguous interpretation: it could indicate the refutation of the information associated with the event of a sale, or it could alternatively denote the prevention of the same event. In this region of the model's output we also find instances where the object in the pairing offers an ambiguous interpretation, such as *heard chink* (is a *chink* a sound or a small gap?), *arrived flat* (flat could actually be interpreted as an adjunct), and *cancel classes* (*class* is in itself a very ambiguous noun, though arguably somewhat specified by the context of *cancel* in this case). Finally, at the coercive end of the model's output, we have examples such as *finished antibiotics*, *hears vowel*, and *denies rift* where the object is clearly taking on the type of the argument implied by the verb (and it's notable that the highly coercive verb *finish* figures prominently in this region of the output).

Geometrically speaking, what we observe as we move from the selectional to the coercive is first a broadening of the region defined by our model, and then a gradual listing as the noun typically becomes prevalent through the co-occurrence dimensions it contributes to the projection. We can detect a move into a less semantically coherent subspace as we discover less overlap between the dimensions that are salient to each of the terms under analysis. The decrease in the angle at the vertex of the maximal point A, and the corresponding increase in the angle at the normalised mean point M', is a perhaps slightly surprising but also rewarding and ultimately understandable feature of this approach. Another point of note is the relative lack of correlation with the actual distance $\overline{VN}$ between the word-vectors and coercion, which, in conjunction with the somewhat strong negative correlation between $\angle VON$ and coerciveness, suggests that the actual Euclidean relationship of the word-vectors is less semantically indicative than various other geometric features of these subspaces. It's also worth mentioning that the length of the object vector tends to increase towards coercion, indicating an increasing dominance of the argument over semantically contextualised subspaces, whereas the length of the verb is somewhat neutral across the spectrum.

6 Conclusion: Strong Results Using Minimal Data

We have proposed a new approach to the identification of semantic type coercion, achieving state-of-the-art results by using the context of both verb-object pairs and their sentential situation to projection semantically productive geometries. Moreover, we have demonstrated the generalisability of this approach, applying it to a more challenging experimental set-up based on a reshuffling of the data provided for the original task.

The work presented here is clearly an introduction to a novel approach to distributional semantics, motivated by theoretical insight. There are a variety of model parameters which merit further exploration:

the dimensionality of our subspaces, for instance, and the co-occurrence window size used to build our base space, not to mention the fundamental issue of corpus selection. There is also the question of the statistics which we use to calculate the scalars of our base spaces. PMI is a well known option, with the variant presented here being adapted to the fit the selectional requirements of our approach, but there are other methods worth considering as well (Bullinaria and Levy, 2007, offer an overview). Following on this is the question of the calculations used to make our dimensional selections. While we have made the assumption that dimensions with high PMI values for either or both terms being analysed will be good candidates for defining a subspace in which to compare the semantic relationship between the terms, it may be the case that some more subtle aspect of the relationship between the terms along a given dimension – their relative situation in relation to the mean value of the dimension, for instance – could indicate an even more productive projection from our base spaces. Indeed, it could turn out that there are features of dimensions themselves, such as variance, the clustering of values, of just the number of non-zero values, that might suggest a dimension is simply *ibso facto* better suited for providing a basis for a geometric analysis.

Returning to the theoretical overview of coercion offered in Section 2, we can now posit that there are interactions between the co-occurrence profiles of verbs, their arguments, and the overall sentential context in which they occur that induce geometries relating to the match or mismatch in the semantic class of the words being modelled. The tendency towards coercion can be captured in terms of a general widening and decentralising of the region of points associated with the words and the overall statistical features of the dimensions that they select. We have not attempted to make any headway on the interpretation of coercive usage through the identification of specific classes here, but the groundwork for a geometric, computational approach to this more involved semantic analysis has been laid.

We also note that our methodology does not make use of the identification of dependency relationships between the words in the sentences used for training and testing, or on any sort of parsing of the underlying corpus used to build our base model. It would be reasonable to conjecture that such steps might further enhance the models' already strong performances, as we would be building precisely the type of information used for the identification of the selected semantic class into the models' processes. But on the other hand, we argue that the fact that we can extrapolate such semantically productive geometries from such basic data indicates the power of this approach, not only in terms of its generalisability beyond the data observed in the process of training for coercion identification, but also potentially towards a wider range of semantic tasks involving more generally ambiguous language and compositionality.

Acknowledgement

Stephen McGregor's research has been supported by EPSRC grant EP/L50483X/1.

References

Agres, K., S. McGregor, M. Purver, and G. Wiggins (2015). Conceptualising creativity: From distributional semantics to conceptual spaces. In *Proceedings of the 6th International Conference on Computational Creativity*, Park City, UT.

Agres, K. R., S. McGregor, K. Rataj, M. Purver, and G. A. Wiggins (2016). Modeling metaphor perception with distributional semantics vector space models. In *Proceedings of the Workshop on Computational Creativity, Concept Invention, and General Intelligence*.

Asher, N. (2011). *Lexical meaning in context: A web of words*. Cambridge University Press.

Baroni, M. and A. Lenci (2010). Distributional memory: A general framework for distributional semantics. *Computational Linguistics 36*(4).

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Bullinaria, J. A. and J. P. Levy (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods 39*(3), 510–526.

Clark, S. (2015). Vector space models of lexical meaning. In S. Lappin and C. Fox (Eds.), *The Handbook of Contemporary Semantic Theory*. Wiley-Blackwell.

Copestake, A. and T. Briscoe (1995). Semi-productive polysemy and sense extension. *Journal of semantics 12*(1), 15–67.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990). Indexing by latent semantic analysis. *Journal for the American Society for Information Science 41*(6), 391–407.

Fauconnier, G. and M. Turner (2003). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York, NY: BasicBooks.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and Semantics Volume 3: Speech Acts*, pp. 41–58. New York: Academic Press.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. The MIT Press.

Jezek, E. and V. Quochi (2010). Capturing coercions in texts: a first annotation exercise. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, pp. 1464–1471.

Lakoff, G. and M. Johnson (1980). *Metaphors We Live By*. University of Chicago Press.

Langacker, R. (1991). *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Berlin: Mouton de Gruyter.

Lapata, M. and A. Lascarides (2003). A probabilistic account of logical metonymy. *Computational Linguistics 29*(2), 261–315.

Lebret, R. and R. Collobert (2014). Word embeddings through hellinger pca. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 482–490.

Levy, O. and Y. Goldberg (2014). Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27*, pp. 2177–2185. Curran Associates, Inc.

Markert, K. and M. Nissim (2009). Data and models for metonymy resolution. *Language Resources and Evaluation 43*(2), 123–138.

McGregor, S., K. Agres, M. Purver, and G. Wiggins (2015). From distributional semantics to conceptual spaces: A novel computational method for concept creation. *Journal of Artificial General Intelligence*.

Michaelis, L. A. (2004). Type shifting in construction grammar: An integrated approach to aspectual coercion. *Cognitive linguistics 15*(1), 1–68.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. In *Proceedings of ICLR Workshop*.

Moens, M. and M. Steedman (1988). Temporal ontology and temporal reference. *Computational linguistics 14*(2), 15–28.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity 41*(5), 673–690.

Pennington, J., R. Socher, and C. D. Manning (2014). Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*.

Pustejovsky, J. (1991). The generative lexicon. *Computational linguistics 17*(4), 409–441.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Pustejovsky, J. (2011). Coercion in a general theory of argument selection. *Linguistics 49*(6), 1401–1431.

Pustejovsky, J., P. Hanks, and A. Rumshisky (2004). Automated induction of sense in context. In *Proceedings of the 20th international conference on Computational Linguistics*, pp. 924–931.

Pustejovsky, J. and E. Jezek (2008). Semantic coercion in language: Beyond distributional analysis. *Italian Journal of Linguistics 20*(1), 175–208.

Pustejovsky, J., A. Rumshisky, A. Plotnick, E. Jezek, O. Batiukova, and V. Quochi (2010). Semeval-2010 task 7: Argument selection and coercion. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 27–32.

Roberts, K. and S. M. Harabagiu (2010). UTDMet: Combining WordNet and corpus data for argument coercion detection. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 252–255.

Roberts, K. and S. M. Harabagiu (2011). Unsupervised learning of selectional restrictions and detection of argument coercions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 980–990.

Shutova, E. (2013). Metaphor identification as interpretation. In *Proceedings of *SEM 2013*.

Shutova, E., J. Kaplan, S. Teufel, and A. Korhonen (2013, July). A computational model of logical metonymy. *ACM Trans. Speech Lang. Process. 10*(3), 11:1–11:28.

Veale, T. and Y. Hao (2008). A fluid knowledge representation for understanding and generating creative metaphors. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, pp. 945–952.

Verspoor, C. M. (1997). *Contextually-dependent lexical semantics*. University of Edinburgh.

Wilson, D. and D. Sperber (2012). *Meaning and Relevance*. Cambridge University Press.