Why is *penguin* more similar to *polar bear* than to *sea gull*? Analyzing conceptual knowledge in distributional models

Pia Sommerauer
Computational Lexicology and Terminology Lab
Vrije Universiteit Amsterdam
De Boelelaan 1105 Amsterdam, The Netherlands
pia.sommerauer@vu.nl

Abstract

What do powerful models of word meaning created from distributional data (e.g. Word2vec (Mikolov et al., 2013) BERT (Devlin et al., 2019) and ELMO (Peters et al., 2018)) represent? What causes words to be similar in the semantic space? What type of information is lacking? This thesis proposal presents a framework for investigating the information encoded in distributional semantic models. Several analysis methods have been suggested, but they have been shown to be limited and are not well understood. This approach pairs observations made on actual corpora with insights obtained from data manipulation experiments. The expected outcome is a better understanding of (1) the semantic information we can infer purely based on linguistic co-occurrence patterns and (2) the potential of distributional semantic models to pick up linguistic evidence.

1 Introduction

Distributional semantic representations capture semantic similarity and relatedness and, perhaps more importantly, enable machine learning-based Natural Language Processing models to abstract over lexical representations. But what type of semantic information do they contain? Could distributional models show that the concepts *lemon* and *moon* share shape and color, but differ with respect to almost everything else? Understanding what semantic knowledge is represented in embeddings can not only help us improve those representations but also shed light on questions about lexical representation raised in cognitive linguistics (e.g. the suitability of embeddings for models of metaphor interpretation (Utsumi, 2011)). Understanding the way components of meaning are represented could eventually enable us to use data-derived, distributional representations for lexical reasoning.

While exiting model analysis methods (Hupkes et al., 2018; Belinkov and Glass, 2019; Saphra and Lopez, 2018) have yielded initial insights, they are still limited when applied to distributional word representations. Gaining insights into semantic representations derived from massive amounts of textual data thus entails answering two core questions: (1) What information about concepts can we find in the linguistic data and how does it relate to people's knowledge about concepts? (2) What linguistic information in the data can be picked up by a distributional semantic model and how is it represented? Answering these questions entails the following four steps:

1. Formulate linguistic hypotheses about what kind of knowledge about concepts we expect to be reflected by linguistic corpora based on theoretical and experimental research.
2. Build a corpus of human judgments reflecting human knowledge about concepts suitable to test the hypotheses.
3. Investigate the potential of distributional models and model analysis methods by simulating different types of linguistic evidence of semantic properties in text corpora.
4. Test hypotheses about what is represented in distributional models and data and interpret the results with respect to the potential of distributional models and analysis methods.

The core questions of this research proposal and their interaction are illustrated in Figure 1. The remainder of this paper is structured as follows: After discussing related work in Section 2, I present linguistic hypotheses in Section 3. The corpus of human judgments of property-concept pairs for testing these hypotheses is presented in Section 4. Section 5 outlines model analysis methods and simulation experiments, followed by a conclusion and reflection on possible outcomes in Section 6.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 134–142
July 5 - July 10, 2020. ©2020 Association for Computational Linguistics

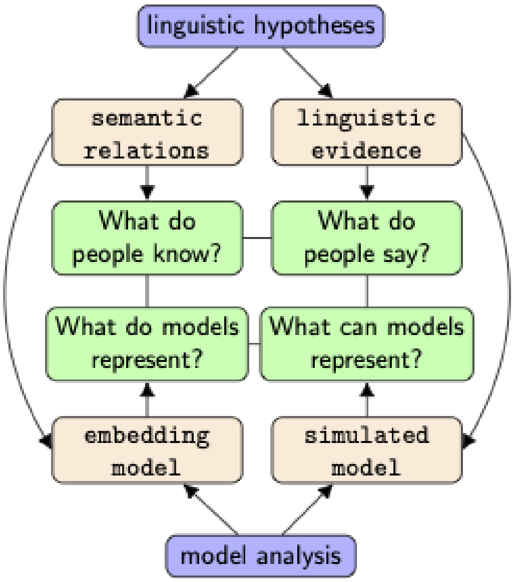

Figure 1: Framework for investigating conceptual knowledge in distributional models from two perspectives: (1) linguistic hypotheses about semantic knowledge and textual evidence and (2) the potential of model analysis methods and models. The questions are approached through model analysis methods on real and simulated data.

2 Related work

Several studies investigate the relation between semantic features recorded in feature norm datasets (McRae et al., 2005; Devereux et al., 2014; Vinson and Vigliocco, 2008; Vigliocco et al., 2004) and embedding vectors (Fagarasan et al., 2015; Tsvetkov et al., 2015, 2016; Herbelot and Vecchi, 2015; Herbelot, 2013; Riordan and Jones, 2011; Glenberg and Robertson, 2000; Derby et al., 2018; Forbes et al., 2019; Rubinstein et al., 2015). These studies indicate that (at least partial) mappings between distributional and conceptual spaces are possible and that conceptual knowledge can complement distributional representations. Erk (2016) shows that distributional similarity can indicate property-overlap. Gupta et al. (2015) show that attributes of the type of knowledge recorded in knowledge bases can, to some extent, be learned from word embeddings. Herbelot (2013) hypothesizes that Gricean maxims determine what is mentioned in text, based on limited datasets. These studies provide partial evidence for conceptual knowledge in distributional data, but they do not provide a systematic account of the underlying factors at play.

A major reason for this gap is the difficulty of interpreting representations resulting from machine learning models. Diagnostic classification has proven successful in the analysis of such representations (Hupkes et al., 2018; Belinkov and Glass, 2019) and word embedding representations (Yaghoobzadeh and Schütze, 2016; Sommerauer and Fokkens, 2018; Yaghoobzadeh et al., 2019). However, the results of these experiments provide limited insights.

Unverified negative examples. For instance, in the CSLB norms (Devereux et al., 2014), **has_legs** is listed for several birds, but not for *owl*, *duck*, and *eagle*. This introduces noise to already rather small datasets used to investigate property knowledge in distributional data (Derby et al., 2018). Yaghoobzadeh et al. (2019) apply diagnostic classification to investigate semantic classes using a large, automatically generated dataset derived from Wikipedia, which is likely to contain noise. Sommerauer and Fokkens (2018) and Herbelot and Vecchi (2015) have provided small sets of verified examples to combat this issue.

Distribution of examples. A classifier is likely to be able to separate words which are located in entirely different areas of the semantic space, but this does not mean it has recognized a specific property. For instance, the ability to separate red fruits (e.g. *strawberry*) from furniture (e.g. *table*) does not indicate that the property **red** was recognized. Sommerauer and Fokkens (2018) provide a small, qualitative analysis with respect to example distribution, but to the best of my knowledge, this has not been investigated systematically. Rubinstein et al. (2015) show that taxonomic properties yield higher performance in diagnostic classification experiments than (mostly physical) attributes. A possible explanation for this could be that taxonomic properties (e.g. **is_animal** are much easier to detect because of many correlating properties resulting in high general similarity in the semantic space.

Interpretation of performance. Saphra and Lopez (2018) point out that diagnostic classifiers can achieve high performance purely based in noise in the data instead of meaningful signals (Zhang and Bowman, 2018; Wieting and Kiela, 2018). To the best of my knowledge, this has not been taken into account yet in studies on embeddings.

The research proposed here is the first attempt to combine a systematic analysis in terms of linguistic hypotheses with with a methodological investigation addressing these limitations.

3 Linguistic hypotheses

This sections presents hypotheses about (a) *what* aspects of conceptual information people mention in texts (Section 3.1) and (b) *how* they mention it (Section 3.2).

3.1 Semantic relations

I define semantic relations representative of four major factors: impliedness, typicality, affordedness and variability. The factors are based on theoretical and empirical accounts in cognitive and computational linguistics (Grice, 1975; Gibson, 1954; Glenberg, 1997; Dale and Reiter, 1995; Sommerauer et al., 2019). The relations are used to label a corpus of property-concept pairs. To test the hypotheses by means of model analysis methods, it is necessary to have reliable information about negative examples of properties. I distinguish several negative relations (e.g. it can be impossible or unusual that a property applies to instances of a concept) to facilitate the annotation task.

Impliedness. Most conceptual knowledge can be seen as highly implied. Mentioning it would constitute a violation of the Gricean maxim of quantity. This is likely to be particularly relevant for properties which are inherited from lexical categories. For instance, the knowledge that a dog is an animate being with a heartbeat is unlikely to be mentioned explicitly. This tendency could be connected to claims about lexical retrieval (Collins and Quillian, 1970). Whether this is indeed the case is a question for further research.

Typicality. Corpus research has shown that people tend to express property-concept relations explicitly for cases in which a concept is a particularly good example of a property (Veale and Hao, 2007; Veale, 2011, 2013). For instance, colors tend to be described in terms of things which illustrate them particularly well (e.g. *as white as snow*, *as red as blood*, *as black as ebony wood*[1]). In contrast, properties which are typical of a concept (and evoked in many participants in elicitation tasks such as the CSLB norms (Devereux et al., 2014)) are most likely strongly implied conceptual knowledge and not mentioned explicitly (e.g. **green** - *broccoli*).

Affordedness. According to research in cognitive linguistics, a central component of semantic knowledge consists of the actions which are available to a person in a particular situation (called afforded actions) (Gibson, 1954; Glenberg, 1997; Glenberg and Robertson, 2000). For instance, you can do several things with a rock, such as throw or drop it (Fulda et al., 2017). Many texts refer to events, which consist of actions involving participants. From this perspective, it is very likely that activities in which instances of concepts are involved are also mentioned in natural language. Glenberg and Robertson (2000) show that distributional models give good indications of activities usually associated with concepts, but cannot distinguish possible but unusual from impossible activities. Fulda et al. (2017) show that embedding models are helpful in affordance extraction. It can thus be expected that frequently performed actions are mentioned in text and can give indications about other properties (e.g. round objects such as bowling balls tend to roll). Possible but unusual activities are unlikely to be mentioned consistently.

Variability. Instances of concepts can vary with respect to a particular property. For instance, bell peppers can be red, green or yellow. Since neither of the colors is implied by the concept, information about it is more likely to be mentioned. In some cases, the property can even indicate an important distinction between different sub-concepts (e.g. brown, black and grey can distinguish different types of bears). In such cases, important and potentially distinguishing information is expressed via the property.

Negative relations. Several relations with no or only a loose association between property and concept can be distinguished. Linguistic corpora are unlikely to contain consistent evidence of such cases. The main reason for defining different types of negative relations is to facilitate the annotation task. Furthermore, they can be informative for further analysis. The relations include: properties which apply to concepts in rare cases, properties which can apply in unusual (such as fictional) cases and impossible combinations. We also include properties which can apply in creative, figurative expressions.

3.2 Linguistic evidence

Linguistic evidence of a semantic property can appear in different forms:

Direct. A property is expressed by its corresponding lexical form. For instance, a direct expression of the semantic property **red** is the adjective *red* and its morphological variants (if they exist),

[1] `https://www.pitt.edu/~dash/grimm053.html` (last accessed 2020-02-18)

for instance *reddish*.

Indirect. Semantic properties can be expressed indirectly in terms of a logical consequence or behavior that is tied to a property. For instance, things which have the semantic property **round** usually roll. Words such as *roll* and their morphological variants act as indirect evidence.

Property-preserving. Words can express properties which partially overlap with the semantic property in question. For instance, the semantic property **swim** can be expressed by *float* or *glide* in some contexts. Those expressions can also express other semantic properties and are thus not exclusively tied to the target property.

Related. Semantic properties can be related to other properties of concepts. For instance, the semantic property **swim** is closely related to different kinds of water, such as *sea*, *river* or *pond* and possibly also *beach* or *sand*. These expressions are related to a wide variety of properties and most certainly not exclusively tied to instances of concepts which swim.

Correlation. Properties which are not expressed can correlate strongly with an entire category of concepts. For instance, all birds **lay eggs**. While this is something *chickens* usually do/are used for, the activity is less prominent for *canaries* and thus unlikely to be mentioned in texts. However, it is likely that something like **belonging to the category of birds** is apparent from linguistic context, as indicated by Hearst patterns (Hearst, 1992) and research about predicting hyponymy relations from embeddings (Fu et al., 2014). Thus, the close connection between category and property may result in a form of linguistic evidence indicating a category which is very closely tied to a semantic property.

Property-category. Expressions of properties belonging to the same category (e.g. *red*, *yellow* and *green* express colors) in the context of a concept can indicate an entire property-category. This is likely to be the case if instances of a concept can have one of a variety of properties that belong to the same category (e.g. color) and the properties occur with similar frequencies (e.g. white, red, blue (etc.) t-shirts).

Table 1 shows the specific semantic relations with respect to the (sub-)set of instances of a concept they apply to and the type of corpus evidence we expect to find for property-concept pair.

4 Dataset design and crowd annotation

The dataset for this thesis should contain concept-property pairs annotated with the fine-grained semantic relations introduced in Section 3.1. The dataset should contain (1) enough positive and negative examples of a property to allow for diagnostic experiments and (2) positive and negative examples which cannot easily be separated based on general similarity in the semantic space (Sommerauer and Fokkens, 2018; Sommerauer et al., 2019).

To address these aspects, the property-concept pairs were collected following the strategy outlined in Sommerauer et al. (2019). Firstly, properties which are expected to apply to concepts across different semantic categories were selected (e.g. colors). Secondly, existing resources (the CSLB feature norms (Devereux et al., 2014) (an extended and improved version of the norms collected by (McRae et al., 2005)), but also WordNet (Miller, 1995; Fellbaum, 2010), ConceptNet (Speer and Havasi, 2012) and stereotype data (Veale, 2013) were used to collect positive and negative example candidates for these properties. Where possible, candidates were selected from diverse semantic categories. The candidates were extended by using a large-scale distributional model (GoogleNews Word2vec model[2]).

The candidate pairs are labeled with semantic relations in a crowd task. Crowd workers are presented with natural language statements about a specific pair illustrating a semantic relation and asked to indicate whether they agree or disagree.[3] Test runs indicate that workers can complete around 70 questions in about 10 minutes.[4]

Each property-concept pair should have at least one relation which is perceived as appropriate by most participants (and is thus labeled with 'agree').[5] However, it has been shown that ambiguity is inherent to many semantic annotation tasks (Dumitrache et al., 2018), leading to disagreements. Disagreement in this task is likely to arise

[2] Downloaded from https://code.google.com/archive/p/word2vec/

[3] An example of such a statement illustrating typical_of_concept would be: *"Fly" is one for the first things which come to mind when I hear "stork' because flying is one of the typical movements of (a/an) stork'.* The full set of statements can be found at https://github.com/cltl/SPT_annotation

[4] The task was set up using the Lingoturk framework (Pusse et al., 2016) and is being distributed via the platform Prolific https://www.prolific.co/.

[5] More than one relation can apply (e.g. both typicality relations).

set of instances	factor	relation	evidence
most - all	impliedness	implied	correlation
	typicality	typical_of_concept	sparse - none
		typical_of_property	direct, property-preserving, related
	affordedness	affording_activity	indirect, property-preserving, related
		afforded_usual	direct, property-preserving, related
		afforded_unusual	sparse to none
some	variability	variability_limited	direct, property-category
		variability_open	property-category
few-none	negative cases	rare	sparse - none
		unusual	sparse - none
		impossible	sparse - none
		creative	sparse - none

Table 1: Summary of linguistic hypotheses about semantic relations and types of evidence.

.

from two factors: (1) ambiguity in the interpretation of the concept, property, relation or combination and (2) different levels of knowledge about the world. Disagreement caused by interpretation differences is particularly relevant for this dataset, as this is can indicate polysemy, which has been shown to have an impact on embedding representations (Del Tredici and Bel, 2015; Yaghoobzadeh et al., 2019, e.g.). It is, however, still an open question how exactly it relates to the representation of semantic properties. Table 2 shows the answers collected for a clear pair, an ambiguous pair and an ambiguous pair additionally perceived as difficult.

relation	p1	p2	p3
typical_of_property	10	3	3
typical_of_concept	10	5	5
affording_activity	10	4	5
implied_category	8	6	4
variability_limited	7	7	3
variability_open	2	3	7
rare	1	2	5
unusual	0	4	4
impossible	0	3	0
creative	0	4	3

Table 2: Number of annotators (out of 10) who selected ‘agree’ for a semantic relation shown for three pairs of varying difficulties: sweet-honey (p1) (clear), made of wood - beam (p2) (ambiguous) and hot-chutney (p3) (not well known according to a worker).

.

Inter-annotator agreement alone cannot be used to evaluate the quality of the dataset. Disagreement is not only an expected, but a desired and meaningful outcome. Instead, I consider the quality of the annotations from multiple perspectives: (1) As a basis for comparison, I apply IAA metrics to the entire dataset and portions of the dataset which I expect to trigger high or low agreement. These portions have been selected in advance. (2) I consider the quality of the workers in terms of whether they contradict themselves in their answers (e.g. label a single pair as typical *and* impossible). A low number of contradictions can be seen as an indication of a clear task. Workers with high contradiction rates can be excluded, which should increase the IAA on the remaining annotations.(3) I analyze the data with the crowd-truth framework (Dumitrache et al., 2018), which provides a fine-grained analysis of workers, annotation units and labels. (4) A subset of pairs is being annotated by trained experts. These annotations serve as a gold standard and can provide more insights into disagreements and worker behavior. They can help to reveal additional, possibly unexpected factors causing disagreement.

5 Method

Various analysis methods have been suggested to interpret latent representations resulting from machine learning (particularly deep learning) models. While they have yielded important insights, they still struggle with a number of limitations (Belinkov and Glass, 2019; Saphra and Lopez, 2018). I plan to approach these limitations by pairing analysis methods (described in Section 5.1) with data simulation experiments (described in Section 5.2). This combination is expected to yield insights into (1) the analysis methods and their potential and (2) the representation of linguistic evidence in a text corpus in distributional models.

5.1 Analyzing latent representations

I plan to use diagnostic classification (Hupkes et al., 2018; Belinkov and Glass, 2019) and SVCCA (Singular Vector Canonical Correlation Analysis) (Raghu et al., 2017). SVCCA has been suggested to address some of the limitations of diagnostic classification (Saphra and Lopez, 2018).

Both methods require a specific distribution of positive and negative examples. Distributional models place generally similar concepts in similar areas in the embedding space because they occur in similar contexts. This means that positive examples which are similar to one another, but dissimilar from the negative examples will be easily recognizable (e.g. **fly**: *seagull* vs *table*). Distinguishing them, however, does not mean that evidence of the particular property was discovered. If however, a diverse group of positive examples can be distinguished from negative examples similar to the positive ones (e.g. **fly**: *seagull* vs *penguin*), we conclude that the property has actually been identified with higher confidence. While this type of dataset control cannot eliminate all possible correlations, it is a first step towards more solid evidence.

5.2 Simulation experiments

The following questions should be answered before we can draw conclusions from the analysis of embedding models trained on natural language corpora:

1. How much evidence in the context of a concept is necessary to have an impact on the representation in an embedding model?
2. How do embedding models represent different kinds of evidence? Can they abstract over morphological variants or synonyms of a word?
3. What is the performance of a model analysis methods if there is very clear evidence of a property? What is the difference between embeddings with clear evidence and embeddings without clear evidence?

I approach these questions by introducing artificial evidence to text corpora and training distributional models on these corpora. In the case of distributional models and linguistic evidence, it is challenging to design small and controlled experiments, as the models rely on a substantial amount of data. Building an entirely artificial corpus (as for instance done by Yaghoobzadeh and Schütze (2016)) would entail the risk of losing information responsible for the general structure of a semantic space. Therefore, I will simulate textual evidence of a property by introducing artificial 'evidence words' to the contexts of a random set of words in an otherwise unchanged corpus. Embeddings resulting from this manipulated corpus can then be used to test how much evidence is sufficient for information to be recognized by analysis methods. I expect this approach to show how the performance of diagnostic methods relates to the presence or absence of textual evidence. These insights are crucial form the interpretation of analysis methods applied to a natural corpus.

6 Conclusion

This proposal presents a framework for investigating the semantic content of distributional word representations from two perspectives: Firstly, I propose to test linguistic hypotheses about what aspects of conceptual knowledge are represented in natural language. Secondly, I propose to interpret the results against the background of a methodological investigation of model analysis methods and the potential of distributional models.

The linguistic hypotheses to be tested may be falsified. While this would be a negative result, it is still a relevant insight and can be used as a basis for new predictions. Furthermore, it can be expected that the methodological insights gained in the simulation experiments can inform other approaches investigating non-transparent embedding representations and yield important insights about the behavior of distributional models.

I expect that the corpus and insights gathered in this project can be complementary to resources capturing common-sense knowledge explicitly, such as Conceptnet (Speer et al., 2017) and common sense challenges (e.g. (Talmor et al., 2019)).

Acknowledgments

I would like to thank my supervisors Piek Vossen and Antske Fokkens for their support, feedback and guidance in developing this proposal. I would also like to thank Vivek Gupta for his helpful comments and the anonymous reviewers for their feedback. This research is funded by the PhD in the Humanities Grant provided by the Netherlands Organization of Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO) PGW.17.041.

References

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Allan M Collins and M Ross Quillian. 1970. Facilitating retrieval from semantic memory: The effect of repeating part of an inference. *Acta Psychologica*, 33:304–314.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Marco Del Tredici and Núria Bel. 2015. A word-embedding-based sense index for regular polysemy representation. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 70–78.

Steven Derby, Paul Miller, Brian Murphy, and Barry Devereux. 2018. Representation of word meaning in the intermediate projection layer of a neural language model. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 362–364.

Barry J. Devereux, Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46(4):1119–1127.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, 9:17–1.

Luana Fagarasan, Eva Maria Vecchi, and Stephen Clark. 2015. From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 52–57.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1199–1209.

Nancy Fulda, Daniel Ricks, Ben Murdoch, and David Wingate. 2017. What can you do with a rock? affordance extraction viaword embeddings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1039–1045.

James J Gibson. 1954. The visual perception of objective motion and subjective movement. *Psychological Review*, 61(5):304.

Arthur M Glenberg. 1997. What memory is for. *Behavioral and brain sciences*, 20(1):1–19.

Arthur M Glenberg and David A Robertson. 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language*, 43(3):379–401.

HP Grice. 1975. Logic and conversation. *Foundations of Cognitive Psychology*, page 719.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Aurélie Herbelot. 2013. What is in a text, what isn't, and what this has to do with lexical semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Short Papers*, pages 321–327.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, volume 13, pages 746–751.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Florian Pusse, Asad Sayeed, and Vera Demberg. 2016. Lingoturk: managing crowdsourced tasks for psycholinguistics. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 57–61.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6076–6085.

Brian Riordan and Michael N Jones. 2011. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):303–345.

Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. 2015. How well do distributional models capture different types of semantic knowledge? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 726–730.

Naomi Saphra and Adam Lopez. 2018. Understanding learning dynamics of language models with svcca. *Proceedings of NAACL-HLT*.

Pia Sommerauer and Antske Fokkens. 2018. Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2019. Towards interpretable, data-derived distributional semantic representations for reasoning: A dataset of properties and concepts. In *Wordnet Conference*, page 85.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL-HLT*, pages 4149–4158.

Yulia Tsvetkov, Manaal Faruqui, and Chris Dyer. 2016. Correlation-based intrinsic evaluation of word vector representations. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 111–115.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of EMNLP*.

Akira Utsumi. 2011. Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive science*, 35(2):251–296.

Tony Veale. 2011. Creative language retrieval: A robust hybrid of information retrieval and linguistic creativity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 278–287. Association for Computational Linguistics.

Tony Veale. 2013. The agile cliché: using flexible stereotypes as building blocks in the construction of an affective lexicon. In *New Trends of Research in Ontologies and Lexical Resources*, pages 257–275. Springer.

Tony Veale and Yanfen Hao. 2007. Learning to understand figurative language: from similes to metaphors to irony. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29.

Gabriella Vigliocco, David P Vinson, William Lewis, and Merrill F Garrett. 2004. Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive psychology*, 48(4):422–488.

David P. Vinson and Gabriella Vigliocco. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1):183–190.

John Wieting and Douwe Kiela. 2018. No training required: Exploring random encoders for sentence classification.

Yadollah Yaghoobzadeh, Katharina Kann, Timothy J Hazen, Eneko Agirre, and Hinrich Schütze. 2019. Probing for semantic classes: Diagnosing the meaning content of word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5740–5753.

Yadollah Yaghoobzadeh and Hinrich Schütze. 2016. Intrinsic subspace evaluation of word embedding representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 236–246.

Kelly W Zhang and Samuel R Bowman. 2018. Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis. *EMNLP 2018*, page 359.