Exploring Nominal Coercion in Semantic Spaces with Static and Contextualized Word Embeddings

Chenxin Liu and **Emmanuele Chersoni**
The Hong Kong Polytechnic University
Department of Chinese and Bilingual Studies
Yuk Choi Road 11, Hung Hom, Kowloon, Hong Kong (China)
chenxin.liu@connect.polyu.edu.hk, emmanuelechersoni@gmail.com

Abstract

The distinction between *mass* nouns and *count* nouns has a long history in formal semantics, and linguists have been trying to identify the semantic properties defining the two classes. However, they also recognized that both can undergo meaning shifts and be used in contexts of a different type, via *nominal coercion*.

In this paper, we present an approach to measure the meaning shift in count-mass coercion in English that makes use of static and contextualized word embedding distance.

Our results show that the coercion shifts are detected only by a small subset of the traditional word embedding models, and that the shifts detected by the contextualized embedding of BERT are more pronounced for mass nouns.

1 Introduction

The literature in formal semantics has debated for long on the distinction between *count* nouns and *mass* nouns, which has often been described as an opposition between discrete, countable objects and substances that cannot instead be divided into subunits. A notorious formal characterization of this intuition is provided by Link (1983): mass nouns like *wine* are *non-quantized*, in the sense that each subpart of *wine* will still count as *wine*; on the other hand, count nouns like *cat* are *quantized*, because if you take a subpart of a *cat*, it will not count as a *cat* (Cheng, 1973). According to such view, in other words, the two types of nouns denote in different domains with different properties.

Chomsky (1965) proposed instead a lexicalist perspective on the problem, where nouns are marked with a binary feature ± COUNT determining the kind of syntactic context (mass or count) in which they can appear. Although the approaches adopt different criteria for defining the "countability" of the nouns, they both predict that count nouns will (mostly) appear in count contexts, and mass nouns will (mostly) appear in mass contexts.

However, cases like the following are extremely frequent in natural language:

1. There is *rabbit* in my soup. (count to mass)
2. Two *wines* at table four! (mass to count)

In 1., the count noun *rabbit* is interpreted as *rabbit meat*, while in 2. the plural form of the mass noun *wine* means *glasses of wine*. Both cases are examples of **coercion**, a semantic phenomenon occurring when the standard interpretation of an expression (in our case, the noun) yields an impossible conceptual representation (e.g. in 1. a rabbit swimming in the soup) (Wiese and Maling, 2005); consequently, a more plausible interpretation is retrieved by "enriching" the semantic representation with concepts that are associated to the standard interpretation of the target expression (*enriched composition*; see Jackendoff (1997)). The focus of this paper is specifically on nominal coercion of mass and count nouns.

Since it is rare to find nouns that occur exclusively in either mass or count contexts, it makes more sense to talk about predominantly count and predominantly mass nouns. Chierchia (2010) describes the idea of mass-count *elasticity*, meaning that any noun can be in principle mass or count, its status being determined at the level of the nominal phrase. When we say "predominantly" mass or count noun, therefore, we mean that a noun has the tendency to occur more frequently in one of the two context types. On such basis, the count-mass distinction can be intuitively seen as a continuum, with the nouns traditionally described in the literature being closer to one the two extremes (Katz and Zamparelli, 2012).

In this work, we investigate to what extent modern *Distributional Semantic Models* -which are nowadays the standard for the representation of lexical meaning in NLP- encode the meaning shifts caused by mass-count coercion. We run two different experiments, making use respectively of

Proceedings of the Workshop on Cognitive Aspects of the Lexicon, pages 49–57
November 20, 2022. ©2022 Association for Computational Linguistics
https://doi.org/10.26615/978-954-452-056-4_007

static and contextualized word embedding models to identify the meaning shifts, and we study some of the potential factors that might influence the extent to which a noun is shifting.

2 Related Work

Modern NLP widely adopts Distributional Semantic Models (DSMs) for the representation of lexical meaning, using vectors that are based on the co-occurrences patterns of words in large text corpora. Vector representations are usually compared using the cosine of the angle between them, and the smaller the angle between two words, the closer their meanings will be (Turney and Pantel, 2010).

The literature on DSMs has identified three generations of vector spaces (Lenci et al., 2022). The first generation is typically referred to as *count models* (Baroni et al., 2014), because the spaces are obtained from the extraction of co-occurrences between the target words and the linguistic contexts that are deemed relevant, then the co-occurrences are weighted via associations measures (Landauer and Dumais, 1997; Baroni and Lenci, 2010; Bullinaria and Levy, 2012).

A second family of models emerged in the early 2010s and became known as *word embeddings* or *prediction-based* models (Mikolov et al., 2013; Bojanowski et al., 2017). In such models, the learning of word vectors is generally framed as a supervised task: a neural network is trained to predict words given other context words, and the vectors are learned as parameters. Words that tend to co-occur will have similar vector representations.

However, a common feature of both families is that they produce *static* vector representations, in the sense that each word gets represented as a single vector, which makes it difficult to handle cases of ambiguity and polysemy. The most recent generation of distributional vectors is said instead to be *contextualized*, because word representations are generated in context on the basis of the activation states of a neural language model (Peters et al., 2018; Devlin et al., 2019). One of the advantages of models like BERT (Devlin et al., 2019) is that they allow generating a specific word embedding for each context in which target words occur, making them an interesting option for modeling contextual phenomena such as nominal coercion.

Concerning the modeling work on nominal coercion in Distributional Semantics, Katz and Zamparelli (2012) were the first, to our knowledge, to use DSMs to investigate the phenomenon. They considered pluralisation as a proxy of count usage, and built a traditional count model with separate vector representations for the singular and the plural of a list of candidate mass and count nouns. Consistently with their initial hypothesis, they found that the vector similarity between singular and plural is higher for count nouns than for mass nouns, since the latter undergo a meaning shift when they are pluralized (cf. example 2 in Section 1). Hürlimann et al. (2014) later analyzed the factors affecting the similarity scores in the data by Katz and Zamparelli (2012), reporting that abstract and highly polysemous nouns undergo greater semantic shifts as a consequence of pluralization.

Both these works are close in spirit to our research: in our first experiment, we will use several types of word embedding models to compare the distances between singular and plural forms of mass and count nouns; in our second experiment, we will use the contextualized vectors of BERT to observe how coercion changes the semantic representations of the nouns in mass and count contexts, which we automatically extract from the British National Corpus (Leech, 1992). To our knowledge, this is the first study specifically on mass-count nominal coercion including both static and contextualized embedding models, although other types of coercion have previously been investigated in the literature on DSMs, e.g. complement coercion (Zarcone and Padó, 2011; Chersoni et al., 2017; Rambelli et al., 2020; Chersoni et al., 2021; Ye et al., 2022) or classical metonymies (CONTAINER-FOR-CONTENT, PRODUCER FOR PRODUCT etc.) (Pedinotti and Lenci, 2020).

3 Experiment 1: Comparing the Singular-Plural Similarity in Static Embedding Spaces

In our first experiment, we follow Katz and Zamparelli (2012) in considering pluralisation as a reliable proxy of count usage and we compare the distributional representations of singular and plural forms of candidate mass and count nouns across the most popular word embedding spaces in the literature. If a model is able to detect the coercion meaning shift, then we expect to see that the average semantic similarity between singular and plural forms is *lower* for the mass nouns (see example 2 in Section 1). We use a list of predominantly count and predom-

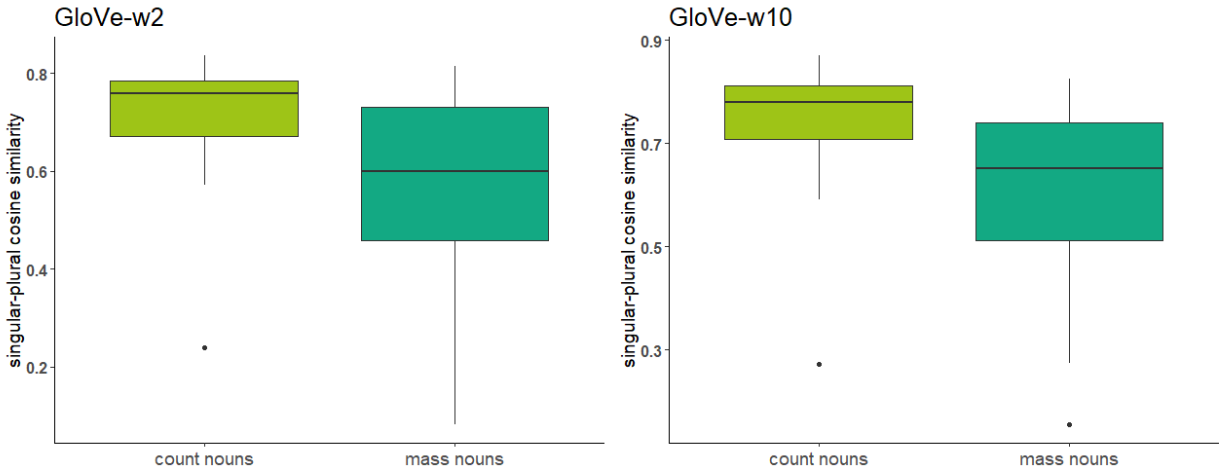

Figure 1: Cosine similarity scores for the singular-plural comparison of count and mass nouns in the GloVe-w2 (left) and in the GloVe-w10 model (right).

inantly mass nouns introduced in the same paper, identified via the selection of syntactic contexts:

- 1a. candidate mass nouns: *information, time, money, detail, space, fun, attention, info, part, work, interest, evidence, experience, energy, power, water, room, recipe, use, opportunity, effort, emphasis, support, research, trouble*;
- 1b. candidate count nouns: *time, year, day, way, person, place, bit, week, man, opportunity, problem, lot, thing, role, company, basis, child, look, one, report, month, book, area, approach, hour*.

The vectors of the target nouns and their corresponding plural forms are firstly extracted from different word embedding spaces. Our pool of models includes the following vector spaces: CBOW vectors (Mikolov et al., 2013), one model with window size 2 (**CBOW-w2**) and one with window size 10 (**CBOW-w10**); Skip-Gram vectors (Mikolov et al., 2013), one model with window size 2 (**SGNS-w2**) and one with window size 10 (**SGNS-w10**); GloVe vectors (Pennington et al., 2014), one model with window size 2 (**GloVe-w2**) and one with window size 10 (**GloVe-w10**); FastText vectors (Bojanowski et al., 2017), one model with window size 2 (**FastText-w2**) and one with window size 10 (**FastText-w10**). Finally, we also include two variants of the Skip Gram where the contexts are selected via syntactic dependency with the target word (Levy and Goldberg, 2014; Lenci et al., 2022), one with untyped dependencies (**SGNS-synf**) and one with typed dependencies (**SGNS-synt**) (e.g. in the first case, given the target *dog* and the context *big dog*, the model will just use the syntactic neighbor *big* as a context, while the second will also include the type of syntactic relation linking the two words, i.e. adjectival modifier). All models have been trained with default hyperparameters on a concatenation of the UkWac (Baroni et al., 2009), of the British National Corpus and of a 2018 dump of Wikipedia [1], and the semantic similarity is estimated via the classical cosine metric.

Model	Avg. mass	Avg. count	p
CBOW-w2	0.59	0.60	
CBOW-w10	0.54	0.56	
FastText-w2	0.66	0.68	
FastText-w10	0.69	0.74	$*$
GloVe-w2	0.56	0.71	$***$
GloVe-w10	0.60	0.75	$***$
SGNS-w2	0.64	0.66	
SGNS-w10	0.62	0.67	$*$
SGNS-synf	0.66	0.66	
SGNS-synt	0.67	0.65	

Table 1: Average of cosine similarity scores between singular and plural forms for each vector space, and p-values computed on the scores of mass and count nouns. Significant differences are reported as follows: $p < 0.05*$, $p < 0.01 **$, $p < 0.001 ***$.

We report the average similarity scores between singular and plural forms for both mass and count nouns in Table 1, and we use the Wilcoxon rank sum test to identify significant differences between the two groups. While for most models the scores are very close, 4 of them manage to identify a sig-

[1] The corpus was POS-tagged and parsed and contains syntactic annotations in the Universal Dependencies format (Nivre et al., 2016; De Marneffe et al., 2021).

nificant difference and in all cases the similarity is lower for the mass nouns. GloVe models are the ones finding the biggest differences, with the scores of mass nouns being significantly lower (see the boxplots in Figure 1). Interestingly, among the embedding models, GloVe is the only one belonging to the more traditional count-based types, and thus more similar to the ones used in the studies of Katz and Zamparelli (2012) and Hürlimann et al. (2014). This may suggest that the GloVe training method, based on global co-occurrence statistics, is a better fit for capturing fine-grained semantic differences than the vectors derived from the Word2Vec family, which are all trained on separate local context windows. Additionally, larger differences are found by the models with a larger window, suggesting that semantic shifts are better captured by vector spaces modeling topic/domain similarity (Turney, 2012). On the other hand, vector spaces modeling local contextual co-occurrences fail to find any difference between mass and count nouns.

4 Experiment 2: Modeling Mass/Count Coercion with BERT

In our second experiment, we extract sentences in which our mass and count nouns occur from the British National Corpus, and we use the patterns described in Katz and Zamparelli (2012) to divide them into mass contexts and count contexts, and then we use the BERT model to compare their contextualized representations. Using BERT allows us to take into account a wider variety of contexts rather than just using a pair of vectors for the singular and plural forms. In this case, we expect that both types of nouns, when they occur in different context types, will have a lower semantic similarity, because both of them will be undergoing semantic shifts (mass to count or count to mass).

The selected patterns are the following:

- 2a. mass: i) singular nouns immediately be preceded by *lots, plenty of, much, more, less, enough, most, sufficient, considerable, boundless, ample,* or *limited* that are not preceded by *a(n)*; ii) singular nouns directly following a verb;

- 2b. count: i) singular nouns immediately be preceded by *a, an, one, every, first, each, another*; ii) plural nouns.

In both contexts, the nouns are excluded if followed by another noun, adjective, or participle to avoid selecting noun-noun compounds. In mass contexts, we also exclude the cases where the target nouns directly follow a participle to prevent misclassification of the participle noun phrases, e.g. *the baked cake*. To increase the reliability of the sentences for the experiment, we manually filter the sentences containing cases of idiomatic usages, e.g. *day by day*. As a result, we extract a total of 614512 sentences. Only the candidate mass or count nouns occurring at least 20 times in both mass and count contexts are considered. Generally, count nouns have a higher average frequency in both contexts, and both count and mass nouns have a higher average frequency in count contexts. The frequency of count nouns in count context ranges from 49218 (*time*) to 6148 (*role*), with a mean of 27787.47, while the frequency in mass context ranges from 8211(*time*) to 20 (*role*), with a mean of 1245.53. For mass nouns, the width of frequency in mass context is from 9259 (*part*) to 45 (*recipe*) and the average frequency is 2110.43, whereas the mass nouns in count contexts have an average frequency of 6414.19, a maximum of 49598 (*time*) and a minimum of 34 (*information*). Notice that nouns can, in principle, occur both as count and as mass nouns, and their frequencies are computed separately as they have been extracted with different patterns. Among the target nouns, *time* and *opportunity* appear as both candidate count noun and candidate mass noun. Although the nouns may be argued as ambiguous, the syntactic patterns used to extract them are unambiguous and can correctly reflect their usage in the count contexts and mass contexts. [2] Therefore, they could still be included to compare the meaning shift a noun undergoes in the transition from the 'standard' context to coerced context.

Noun	Context	Avg. freq	Avg. freq	Min. freq
Count	Count	27787.47	49218	6148
Count	Mass	1245.53	8211	20
Mass	Mass	2110.43	9259	45
Mass	Count	6414.49	48598	34

Table 2: Statistics for the context extraction from the British National Corpus: average, max and min frequency for each noun-context type.

Then we use the BERT-BASE-UNCASED model and the MINICONS Python library (Misra, 2022) [3] to generate semantic representations of the target

[2] We thank the first anonymous reviewer for pointing out this issue.

[3] https://github.com/kanishkamisra/minicons

nouns in context: the idea is to measure the similarity scores of each (mass or count) noun to itself for randomly sampled sentences. We carry out the sampling either i) by selecting context pairs where the target noun occurs in both cases in its mass, or in its count contexts (within the same context type, which could be either count or mass); or ii) by selecting context pairs where the target noun occurs once in a mass context and once in a count context (*between* context types).

This means that each noun type will have its occurrences sampled in three different ways:

1. all context pairs sampled from its own type (mass nouns in mass contexts, count nouns in count contexts);
2. all context pairs sampled from the other type (mass nouns in count contexts, count nouns in mass contexts);
3. the context pair composed by one mass context and one count context.

The similarity comparison between 1) and 3) is the most relevant one for our study: we expect that similarities in 3) to be much lower than in 1), to an extent proportional to the meaning shift that the noun is undergoing. For each noun, we repeat the sampling 10 times from each group, and for each time we randomly extract 10 different context pairs to generate the vectors.

Notice that, differently from a big part of the literature, we use Spearman's rank correlation and not the cosine as a similarity metric for BERT vectors. Our choice is motivated by recent findings about the anisotropy of contextualized vector spaces, where a small number of 'rogue' dimensions dominate the cosine similarity scores (Ethayarajh, 2019; Timkey and van Schijndel, 2021). Timkey and van Schijndel (2021) showed that using postprocessing techniques like normalization or rank-based metrics such as Spearman's rank led to much better correlations with human similarity judgments. Moreover, rank-based metrics have been previously proven to be more robust than cosine in several similarity-related tasks (Santus et al., 2016a,b, 2017, 2018; Zhelezniak et al., 2019).

The results of Spearman's rank correlation experiment are reported in Table 3. The average correlation of context pairs where the target noun occurs in its typical kind of context (i.e. count nouns in count contexts, mass nouns in the mass ones) reflects how semantically similar the target noun is to itself when used in the 'standard' meaning, while the average correlation across different context types reflects the similarity between the standard and the coerced meaning. Therefore, the difference between the two correlations should quantify the meaning shift of the target noun when nominal coercion is imposed on the standard interpretation.

Noun	**Context**	**Avg. corr**
Count	Count	0.455
Count	Mass	0.466
Count	Both	0.360
Mass	Mass	0.550
Mass	Count	0.476
Mass	Both	0.391

Table 3: Average Spearman's rank correlation scores for each noun type under the six different sampling conditions.

Let us illustrate the statement with an example for the predominantly count noun *problem* and an example for the predominantly mass noun *water*, respectively.

s1. more, i believe, than would be acceptable to people, so that nuclear power in itself will never be the solution to our energy ***problems***. (count context)

s2. not surprisingly these devices are distributed with little or no instruction on correct use — thus increasing women's health ***problems***. (count context)

s3. current models seem to be auto-sensing, so there shouldn't be much ***problem***. (mass context)

The noun *problems* in s1 and s2 refer in both cases to a specific issue that needs to be resolved, while *problem* in s3 seems to be more generic and more similar to *trouble*. Accordingly, the correlation of s1 and s2 is 0.55, and the correlation of s1 and s3 is 0.33. The correlation difference between the s1-s2 pair and the s1-s3 pair should reflect the meaning shift that *problem* undergoes, changing from its standard "count" meaning to its coerced interpretation in a mass context.

s4. you can drink ***water*** freely during the course of the diet. (mass context)

s5. in the year to august 1992, the works used 22 per cent less ***water***, 18 per cent less nitrogen, 11 per cent less steam and nine per cent less electricity. (mass context)

s6. to the east of Venice lies Lido di Jesolo and Caorle, with miles of golden sand lapped by the warm ***waters*** of the Adriatic. (count context)

The noun *water* in s4 and s5 is used in mass contexts and it has the standard meaning of i) *water* as

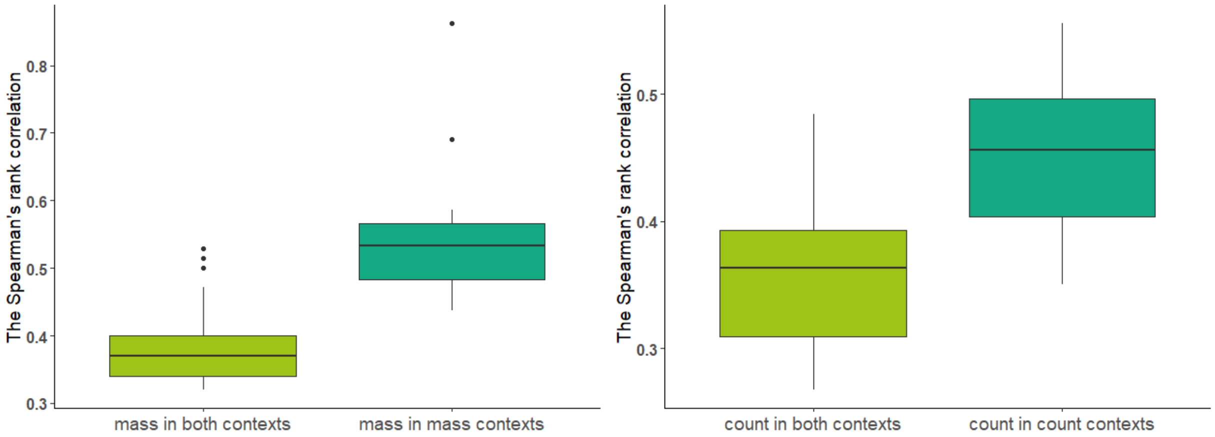

Figure 2: Spearman's rank correlations for mass (left) and count (right) nouns computed with BERT-BASE-UNCASED. The dark green box shows the correlations for the same group (mass in mass, count in count) context sampling, the light green box shows the correlations with one mass and one count context.

a liquid that can be drunk (s4); and ii) the amount of *water* usage in a hydraulic system (s5). On the other hand, in the last context (s6) *waters* is rather referring to a specific geographical/territorial unit. The correlation of the s4-s5 pair is 0.59, while the correlation of the s4-s6 pair is 0.41, and the correlation difference between the two pairs should reflect the meaning shift from mass to count usage.

It is immediately evident from Table that 3 count nouns generally have a lower average Spearman correlation score than mass nouns in either count or mass contexts, suggesting that the cluster of count nouns is less compact and their meanings are more varied and scattered across the semantic space. It should be noticed that many of the count nouns are highly frequent (e.g. *child, thing, way, man, one, place, time, day* all have more than 10K occurrences, more than any mass noun in our data), therefore they might display much more contextual variation in their usage, which could explain the relatively low similarity value. Indeed, the similarity of the count nouns when the contexts are sampled from the count or the mass groups does not differ significantly, with the latter being even slightly higher.

For both count and mass nouns, as we expected, we observe a lower similarity when the contexts are sampled from both groups (see also Figure 2): the average difference with the sampling within the same group is, respectively, of 0.1 and 0.16 correlation points, confirming the finding that mass nouns typically undergo a greater meaning shift (Hürlimann et al., 2014) even when patterns beyond pluralization are taken into account.

It is also noticeable that mass nouns have a relatively high similarity when they are sampled in count contexts, which could be explained by the fact that many of these nouns have systematic secondary meanings that are more compatible with a count usage (nouns denoting drinkable liquids are typically undergoing a shift from the liquid to the container, e.g. *beers* → *pints of beer*, or from the liquid to the variety, e.g. *wines* → *varieties of wine*). In sum, the results of our experiment provide further support to the view that the mass/count distinction should be seen as a continuum, and that the syntactic context is the strongest cue to the type of denotation (Chierchia, 2010). Moreover, even with more varied mass and count contexts than the ones used in previous studies (Katz and Zamparelli, 2012; Hürlimann et al., 2014), we also find that coercion makes mass nouns undergo a greater semantic shift than count nouns.

We also analyze some of the factors mentioned as relevant by Hürlimann et al. (2014) to predict the meaning shift of the nouns: frequency, polysemy, and concreteness. For polysemy, we simply use the WordNet synsets (Fellbaum, 2010) of a noun as an indicator of the number of word senses, while for concreteness we use the values from the English norms by Brysbaert et al. (2014).

Unfortunately, we do not find any significant correlation between the average differences in the Spearman correlations that we computed with BERT and the above-mentioned factors, probably because of the small size of our set of nouns. Table 4 presents the statistics for the top-5 most shifting count and mass nouns (i.e. the nouns with the

Noun	Freq.	WordNet Synsets	Concreteness
company	18792	9	4.11
child	43018	4	4.78
thing	47655	12	3.17
way	29644	12	2.34
area	26273	6	3.72
information	3199	5	2.87
attention	2323	6	2.30
trouble	1051	6	2.25
support	308	11	2.83
money	4947	3	4.54

Table 4: Frequency, synsets and concreteness for the top-5 most shifting count (**bold**) and mass (*italic*) nouns.

highest average correlation difference). Despite the lack of significance of the correlation scores, it can still be noticed that: i) regarding polysemy, the most shifting nouns tend to have a relatively high number of word senses; ii) as for concreteness, the most shifting count nouns have relatively high values, while the most shifting mass nouns tend to denote more abstract entities. More studies with a larger set of predominantly mass and count nouns are needed to confirm the finding.

4.1 A Final Note about Polysemy

With reference to our *rabbit meat* example (see Section 1) and as a general methodological consideration, Reviewer 2 points out that the, given the polysemy of the word *rabbit*, which is also attested in dictionaries, this example cannot be considered as a case of coercion, but it just corresponds to a different word sense. As a consequence, the polysemy of the target nouns should be identified in advance, because otherwise we risk to confuse coercion with occurrences of different word senses.

Since we are adopting the perspective of distributional approaches, in our view the main issue is whether linguistic distributions are *determined* by the inventory of senses of a word, or they are *determining* what we conceive as their inventory of senses, in accordance to the so-called strong versions of the Distributional Hypothesis (Miller and Charles, 1991; Lenci, 2008). In cases such the above-mentioned one, coercion itself might be responsible the emergence of new meanings and senses. Keeping the *rabbit meat* example, one could imagine that, following the same pattern, the speakers of a language at some point could start using the name for its meat in similar mass contexts, and that would undoubtedly qualify as a case of coercion because the coerced meaning will be an innovation, and thus it would not be attested in any dictionary. Only when the usage of the name of the animal for its meat will have become frequent enough to be conventional, then dictionaries will start including it as a secondary sense.

This does not detract from the validity of the reviewer's objection. But we would like to clarify that, in our approach, we consider the word senses annotated in dictionaries and lexicographic resources as possibly consequential to shifts in linguistic distributions, and not the other way around.

5 Conclusion

In this paper, we have presented two experiments on modeling nominal coercion of mass and count nouns with two different typologies of Distributional Semantic Models. In the first experiment, we compared the vector representations of singular and plural mass/count nouns across several popular word embedding models. Perhaps surprisingly, we found that i) the count-based GloVe models and ii) the Word2Vec-like models with larger contextual windows were the most successful in identifying significant differences between singular and plural representations of mass nouns, whose meanings shifted more when they were pluralized, while the most of the other models did not detect any shift. We hypothesized, therefore, that such semantic shifts are better captured by semantic spaces that focus on modeling similarities of topic/domain, rather than similarity of co-occurrence in the same local contexts.

In the second experiment, we compared the vectors generated by BERT in different context types. We found that the self-similarity of the nouns sharply decreased when contexts of different types were sampled to generate the contextualized representations and that the shifts of predominantly mass nouns were more pronounced. Our qualitative analyses suggested that factors such as polysemy and concreteness of the nouns might play a role in predicting semantic shifts, although more studies with a larger set of nouns are necessary.

Another promising direction for future research would be using DSMs to model the effects of nominal coercion in human sentence processing, since psycholinguistic studies proved that, in several languages (e.g. English, German, Mandarin Chinese), coerced nouns lead to increased reading times and longer eye fixations (McElree et al., 2001; Traxler et al., 2002; Pylkkanen and McElree, 2006; Zarcone et al., 2017; Xue et al., 2021). In this sense, integrating DSMs-derived similarity metrics in the current computational models could lead to better estimation of reading difficulties induced by coercion operations.

Acknowledgements

This research was made possible by the start-up research fund (1-BD8S) at the Hong Kong Polytechnic University. We would also like to thank the two anonymous reviewers for their comments.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: a Collection of Very Large Linguistically Processed Web-crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don ' t Count, Predict! a Systematic Comparison of Context-counting vs. Context-predicting Semantic Vectors. In *Proceedings of ACL*.

Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A General Framework for Corpus-based Semantics. *Computational Linguistics*, 36(4):673–721.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas. *Behavior Research Methods*, 46(3):904–911.

John A Bullinaria and Joseph P Levy. 2012. Extracting Semantic Representations from Word Co-occurrence Statistics: Stop-lists, Stemming, and SVD. *Behavior Research Methods*, 44(3):890–907.

Chung-Ying Cheng. 1973. Response to Moravcsik. *Approaches to Natural Language*.

Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Logical Metonymy in a Distributional Model of Sentence Comprehension. In *Proceedings of *SEM*.

Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2021. Not All Arguments Are Processed Equally: A Distributional Model of Argument Complexity. *Language Resources and Evaluation*, 55(4):873–900.

Gennaro Chierchia. 2010. Mass Nouns, Vagueness and Semantic Variation. *Synthese*, 174(1):99–149.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Kawin Ethayarajh. 2019. How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of EMNLP*.

Christiane Fellbaum. 2010. Wordnet. In *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer.

Manuela Hürlimann, Raffaella Bernardi, and Denis Paperno. 2014. Nominal Coercion in Space: Mass/Count Nouns and Distributional Semantics. In *Proceedings of CLiC-it*.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. 28. MIT Press.

Graham Katz and Roberto Zamparelli. 2012. Quantifying Count/Mass Elasticity. In *Proceedings of the West Coast Conference on Formal Linguistics*.

Thomas K Landauer and Susan T Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211.

Geoffrey Neil Leech. 1992. 100 Million Words of English: The British National Corpus (BNC). *Language Research*.

Alessandro Lenci. 2008. Distributional Semantics in Linguistic and Cognitive Research. *Italian Journal of Linguistics*, 20(1):1–31.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A Comparative Evaluation and Analysis of Three Generations of Distributional Semantic Models. *Language Resources and Evaluation*, pages 1–45.

Omer Levy and Yoav Goldberg. 2014. Dependency-based Word Embeddings. In *Proceedings of ACL*.

Godehard Link. 1983. *The Logical Analysis of Plurals and Mass Terms: A Lattice-theoretical Approach*, volume 127. Blackwell Oxford.

Brian McElree, Matthew J Traxler, Martin J Pickering, Rachel E Seely, and Ray Jackendoff. 2001. Reading time evidence for enriched composition. *Cognition*, 78(1):B17–B25.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

George A Miller and Walter G Charles. 1991. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 6(1):1–28.

Kanishka Misra. 2022. minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint arXiv:2203.13112*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, and Natalia Silveira. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of LREC*.

Paolo Pedinotti and Alessandro Lenci. 2020. Don ’ t Invite BERT to Drink a Bottle: Modeling the Interpretation of Metonymies Using BERT and Distributional Representations. In *Proceedings of COLING*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of EMNLP*.

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*.

Liina Pylkkanen and Brian McElree. 2006. The Syntax-semantics Interface: On-line Composition of Sentence Meaning. *Handbook of Psycholinguistics*, pages 539–579.

Giulia Rambelli, Emmanuele Chersoni, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2020. Comparing Probabilistic, Distributional and Transformer-Based Models on Logical Metonymy Interpretation. In *Proceedings of AACL-IJCNLP*.

Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Measuring Thematic Fit with Distributional Feature Overlap. In *Proceedings of EMNLP*.

Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, Chu-Ren Huang, and Philippe Blache. 2016a. Testing APsyn against Vector Cosine on Similarity Estimation. In *Proceedings of PACLIC*.

Enrico Santus, Tin-Shing Chiu, Qin Lu, Alessandro Lenci, and Chu-Ren Huang. 2016b. What a Nerd! Beating Students and Vector Cosine in the ESL and TOEFL Datasets. In *Proceedings of LREC*.

Enrico Santus, Hongmin Wang, Emmanuele Chersoni, and Yue Zhang. 2018. A Rank-Based Similarity Metric for Word Embeddings. In *Proceedings of ACL*.

William Timkey and Marten van Schijndel. 2021. All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality. In *Proceedings of EMNLP*.

Matthew J Traxler, Martin J Pickering, and Brian McElree. 2002. Coercion in Sentence Processing: Evidence from Eye-movements and Self-paced Reading. *Journal of Memory and Language*, 47(4):530–547.

Peter D Turney. 2012. Domain and Function: A Dual-space Model of Semantic Relations and Compositions. *Journal of Artificial Intelligence Research*, 44:533–585.

Peter D Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Heike Wiese and Joan Maling. 2005. Beers, Kaffi, and Schnaps: Different Grammatical Options for Restaurant Talk Coercions in Three Germanic Languages. *Journal of Germanic Linguistics*, 17(1):1–38.

Wenting Xue, Meichun Liu, and Stephen Politzer-Ahles. 2021. Processing of Complement Coercion With Aspectual Verbs in Mandarin Chinese: Evidence From a Self-Paced Reading Study. *Frontiers in Psychology*.

Bingyang Ye, Jingxuan Tu, Elisabetta Jezek, and James Pustejovsky. 2022. Interpreting Logical Metonymy through Dense Paraphrasing. In *Proceedings of CogSci*.

Alessandra Zarcone, Ken McRae, Alessandro Lenci, and Sebastian Padó. 2017. Complement Coercion: The Joint Effects of Type and Typicality. *Frontiers in Psychology*.

Alessandra Zarcone and Sebastian Padó. 2011. Generalized Event Knowledge in Logical Metonymy Resolution. In *Proceedings of CogSci*.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Hammerla. 2019. Correlation Coefficients and Semantic Textual Similarity. In *Proceedings of NAACL*.