*Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 171 181,
Dublin, Ireland, August 23-24 2014.

al., 2009; Poesio et al., 2010).

Despite their theoretical and practical import, modification effects have been largely overlooked in computational linguistics, with the notable exception of Boleda et al. (2012; 2013), who only focused on the extreme case of intensional adjectives, studied a limited number of modifiers, and did not attempt to capture the graded nature of modification (a *dead parrot* is not a prototypical *animal*, but a *toy parrot* is not an *animal* at all).

This paper aims to stimulate computational research into modifier effects on phrase meaning in two ways. First, we introduce a new, large, publicly available data set of modifier-head phrases annotated with four kinds of modification-related subject ratings: whether the concept denoted by the phrase is an instance of the concept denoted by its head (is a *dead parrot* still a *parrot*?), to what extent it is a member of one of the larger categories the head belongs to (is it still a *pet*?), and typicality ratings for the same questions (how typical is a *dead parrot* as a *parrot*? and as a *pet*?).

Second, we present a first attempt to model the collected judgments computationally. We choose distributional semantics (Erk, 2012) as our frame of reference, as it produces continuous similarity scores, in line with the graded nature of the modification effects we are investigating. In particular, we look at the *compositional* extension of distributional semantics (Baroni, 2013), because we need representations not only for words, but also phrases, and we adopt the *asymmetric* similarity measures developed in the literature on lexical entailment (Kotlerman et al., 2010; Lenci and Benotto, 2012), because we are interested in an asymmetric relation (to what extent the concept denoted by the phrase is a good instance of the target class, and not *vice versa*). As far as we know, this is the first time these asymmetric measures are applied to composed representations (Baroni et al. (2012) experimented with entailment measures applied to phrase representations directly harvested from corpora, and not derived compositionally). We are thus also providing a novel evaluation of compositional models and asymmetric measures on a challenging task where they could potentially be very useful.[2]

[2]Connell and Ramscar (2001) showed good correlation of similarity scores produced by the LSA distributional semantic model with human category typicality judgments, however they did not consider phrases nor adopted an asymmetric measure to take directionality into account.

2 The Norwegian Blue Parrot data set

We introduce *Norwegian Blue Parrot* (NBP),[3] a new, large data set to explore modification effects. Given a **h**ead noun h and a **m**odifier adjective or noun m, NBP contains average membership and typicality ratings for the phrase mh both as an instance of h and as an instance of c (a broader **c**ategory h belongs to). As a control, we also present ratings for unmodified h as an instance of c (we will use them below to test similarity measures on their ability to capture the direction of the membership relation, and to zero in on the effect of modification vs. more general membership/typicality effects). We include, and indeed focus on, relations with broader categories because they are more prone to modification effects: Intuitively, a *dead parrot* is still a *parrot*, but it is, at the very least, an atypical *pet*. The statistics in Table 1, discussed below, confirm our intuition that subjects are more likely to assign lower scores with respect to a broader category than to the head category itself (although this is, no doubt, in part by construction, since we started constructing the dataset by mining examples where mh is atypical of c, not h). We collect both membership and typicality ratings because we expect them to have different implications for sound entailment. If x is not a member of class y, then x obviously does not entail y. However, if x is an atypical y, entailment still holds, but some typical properties of y might not carry over (e.g., in an anaphora resolution setting, we might still consider co-indexing *dead parrot* with *animal*, but not with *breathing creature*, despite the fact that *breathing* is a highly characteristic property of *animals*).

In order to make sure that NBP would contain a fair number of examples affected by strong modification effects, we first came up with a set of $\langle m, h, c\rangle$ tuples where, according to our own intuition, m makes h fairly atypical as an instance of c. For example, a *bottle* is a piece of *drinkware*. If we add the modifier *perfume*, we expect that, while subjects might still agree that a *perfume bottle* is a *bottle*, they should generally disagree on the statement that a *perfume bottle* belongs to the *drinkware* category. We refer to tuples of this sort (e.g., $\langle perfume, bottle, drinkware\rangle$) as *distorted* tuples in what follows.[4]

[3]Available from `http://clic.cimec.unitn.it/composes/`

[4]When creating the tuples, we also used some adjectives

We then constructed a number of tuples that should not display a strong modification effect. In particular, in order to insure that any atypical rating we obtained on the distorted tuples could not be explained away by characteristics of m or h alone (rather than by their combination), for each distorted tuple we constructed a few more tuples with the same h and c but a different m, that we did not expect to be strongly distorting (e.g., $\langle plastic, bottle, drinkware \rangle$). Similarly, for each distorted tuple we generated a few more with the same m, but combined with (the same or different) h and c on which the m should not exert a strong effect ($\langle perfume, bottle, container \rangle$). In total, NBP is based on 489 distorted tuples and 1938 more matching tuples.

We constructed NBP to insure that it would contain many tuples displaying strong modification effects, and highly comparable tuples that do not feature such effects. An alternative approach would have been to rate phrases that were randomly selected from a corpus. This would have led to a dataset reflecting a more realistic distribution of modification effects, but it would not have guaranteed, for the same number of pairs, a fair amount of distorted tuples and comparable controls. We leave the study of the natural distribution of modification strength in text to further work.

To find inspiration for the tuples, we looked into various databases containing concepts organized by category, namely BLESS (Baroni and Lenci, 2011), ConceptNet (Speer and Havasi, 2013) and WordNet (Fellbaum, 1998). We insured that all words in our tuples occurred at least 200 times in the large corpus we describe below (phrases were not filtered by frequency, due to data sparseness). Finally, when looking for tuples matching the distorted ones, we made sure that the mh phrases in the new tuples have similar Pointwise Mutual Information to the corresponding phrases in the distorted tuple (or, where the latter were not attested in the corpus, similar m and h frequencies). Finding meaningful combinations among unattested or infrequent phrases was not an easy task and there was not always a perfect candidate. However, the phrases selected in this way yielded challenging items for which there is little or no direct corpus evidence, so that compositional models are required to account for them.

that have been traditionally labeled as intensional by semanticists: *artificial*, *toy*, *former*.

From each source tuple (e.g., $\langle plastic, bottle, drinkware \rangle$), we generated 3 instance-class combinations to be rated: $mh \rightarrow c$ (*plastic bottle* → *drinkware*), $mh \rightarrow h$ (*plastic bottle* → *bottle*), $h \rightarrow c$ (*bottle* → *drinkware*), for a total of 5,849 pairs, that constitute the final NBP data set (2,417 $mh \rightarrow c$ pairs, 2,115 $mh \rightarrow h$ pairs and 1,317 $h \rightarrow c$ pairs).[5]

For each of these pairs, we collected both membership and typicality ratings through two surveys on the CrowdFlower platform.[6] Subjects came exclusively from English speaking countries and no special qualifications were required from them. Membership ratings were collected by asking subjects whether the instance is a member of the class (formulated as a yes/no question). In a separate study, we asked subjects to rate how typical the instance is as member of the class on a 7-point scale. For both questions, we collected 10 judgments per pair and report their averages in NBP. For both surveys, we added 48 control pairs with an expected answer (yes/no for membership, high/low range for typicality), that the subjects had to provide in order for their ratings to be included in the final set ("gold standard" items in crowd-sourcing parlance). These controls included highly prototypical pairs (*dog* → *animal*), possibly with stereotypical modifiers (*beautiful rose* → *flower*), and unrelated pairs (*biology* → *dance*), also possibly under modification (*popular magazine* → *animal*).

We asked for binary rather than graded membership judgments because these are more in line with commonsense intuitions about category membership (we might naturally speak of *sparrows* being more typical birds than *penguins*, but it is strange to say that they are "more birds"). The standard view in the psychology of concepts (Hampton, 1991) is that membership judgments are the product of a hard threshold we impose on the typicality scale (x is not y if the typicality of x as y is below a certain, subject-dependent threshold), although under certain experimental conditions subjects can also conceptualize membership as a graded property (Kalish, 1995).

Membership and typicality ratings, especially in borderline cases such as those we constructed, are the output of complex cognitive processes where large inter-subject differences are expected,

[5] There is a larger number of $mh \rightarrow c$ pairs because different tuples can lead to the same $mh \rightarrow h$ or $h \rightarrow c$ combinations.

[6] `http://crowdflower.com/`

measure	$mh \rightarrow c$	$mh \rightarrow h$	$h \rightarrow c$	tot.
memb.	0.84 (0.2)	0.97 (0.1)	0.88 (0.2)	0.89 (0.2)
typ.	5.45 (1.1)	6.29 (0.6)	5.81 (1.0)	5.84 (1.0)

Table 1: NBP summary statistics: Mean average ratings and their standard deviations across pairs, itemized by instance-class type and in total. Membership values range from 0 to 1, typicality values from 1 to 7.

so it doesn't make sense to worry about "inter-annotator agreement" in this context. Still, several sanity checks indicate that, overall, our subjects understood our questions as we meant them, and behaved in a reasonably coherent manner. First, both average membership and typicality, ratings are significantly lower ($p < 0.001$) for the $mh \rightarrow c$ pairs deriving from those tuples that we manually labeled as distorted than for the non-distorted ones. Moreover, for membership, in 86% of the cases at least 8 over 10 subjects gave the same response. For typicality, the observed average rating standard deviation across pairs (1.2) is significantly below what expected by chance ($p < 0.05$), based on a simulated random rating distribution. Membership and typicality ratings are highly correlated, but not identical ($r = 0.76$)

Table 1 reports mean membership and typicality scores in NBP. Both ratings are negatively skewed, that is, subjects had the tendency to respond assertively to the membership question and to give high typicality scores. This is not surprising: Because of the way NBP was constructed, there are about 4 tuples with no expected strong modification effect for each distorted tuple. Furthermore, except for the negative control items (not entered in NBP), our questions did not feature cases where a negative/low response would be entirely straightforward (of the "is a cat a building?" kind). We observe moreover that, in accordance with the intuition we discussed at the beginning of this section, the ratings are extremely high when the class is identical to the phrase head. On the other hand, the $mh \rightarrow c$ condition displays, as expected, the lowest averages, suggesting that this will be the most interesting type to model experimentally.

Table 2 presents a few example entries from NBP. The first block of the table illustrates cases with the highest possible membership and typicality scores. At the other extreme, the second block contains examples with very low membership and typicality. Interestingly, there are also cases, such as the ones in the third block of the table, where all subjects agreed on class membership, but the typicality scores are relatively low (we did not find clear cases of the opposite pattern, and indeed we would have been surprised to find highly typical instances of a class not being treated as members of the class).

instance	*class*	*memb.*	*typ.*
top membership, top typicality			
gourmet soup	food	1.00	7.00
huge tiger	predator	1.00	7.00
sugared soda	drink	1.00	7.00
live fish	animal	1.00	7.00
Thai rice	rice	1.00	7.00
silver spoon	spoon	1.00	7.00
low membership, low typicality			
fatal shooting	sport	0.20	1.40
human egg	food	0.40	1.50
perfume bottle	drinkware	0.10	1.30
explosive vest	commodity	0.30	1.90
lemon water	chemical	0.20	1.60
creamy rice	bean	0.20	1.30
top membership, (relatively) low typicality			
sick tuna	tuna	1.00	3.20
explosive vest	vest	1.00	3.50
perforated sieve	tool	1.00	4.20
bottled oxygen	substance	1.00	4.30
grilled trout	creature	1.00	4.40
educational toy	amusement	1.00	4.50

Table 2: Instance-class pairs illustrating various combinations of membership and typicality ratings in NBP.

Some examples in Table 2 illustrate an important design choice we made in constructing NBP, namely, to ignore the issue of whether potential modification effects are actually due to the modifier and the category pertaining to different *word senses* of the head term. One might argue, for example, that *egg* has a *food* sense and a *reproductive vessel* sense. The *human* modifier picks the second sense, and so, obviously, *human eggs* are judged as bad instances of *food*. While we see the point of this objection, we think it's impossible to draw a clear-cut distinction between discrete word senses (even in the rather extreme egg case, the eggs we eat are reproductive vessels from a chicken point of view!). This has been long recognized in the linguistic and cognitive literature (Kilgarriff, 1997; Murphy, 2002),

and even by the computational word sense disambiguation community, that is currently addressing the continuous nature of polysemy by shifting to the lexical-substitution-in-context task (McCarthy and Navigli, 2009). Context provides fundamental cues to disambiguating polysemous words, and noun modifiers typically act as important disambiguating contexts for the nouns. Thus, we think that it is more productive for computational systems to handle modifier-triggered disambiguation as a special case of the more general class of modification effects, than to engage in the quixotic pursuit to determine, *a priori*, what's the boundary between a word-sense and a "pure" modification effect. Note in Table 2 that *grilled trout* was unanimously rated by subjects as an instance of the *creature* category, despite the fact that the cooking-related *grilled* modifier cues a classic shift from an *animal* (and thus *creature*) sense to *food* (Copestake and Briscoe, 1995). Examples like this suggest that our agnosticism is warranted.

3 Methods

3.1 Composition models

We experiment with many ways to derive a phrase vector by combining the vectors of its constituents. Mitchell and Lapata (2010) proposed a set of simple models in which each component of the phrase vector is a function of the corresponding components of the constituent vectors. Given vectors $\vec{a}$ and $\vec{b}$, the weighted additive model (**wadd**) returns their weighted sum: $\vec{p} = w_1\vec{a} + w_2\vec{b}$. In the dilation model (**dil**), the output vector is obtained by decomposing one of the input vectors, say $\vec{b}$, into a vector parallel to $\vec{a}$ and its orthogonal counterpart, and then dilating only the parallel vector by a factor λ before re-combining. The corresponding formula is: $(\vec{a}\cdot\vec{a})\vec{b} + (\lambda - 1)(\vec{a}\cdot\vec{b})\vec{a}$. In our experiments, we stretch the head vector in the direction of the modifier (i.e., $\vec{a}$ is the modifier, $\vec{b}$ is the head). In the multiplicative model (**mult**), vectors are combined by component-wise multiplication, such that each phrase component p_i is given by: $p_i = a_i b_i$.

Guevara (2010) and Zanzotto et al. (2010) propose a full form of the additive model (**fulladd**), where the two constituent vectors are multiplied by weight matrices before being added, so that each phrase component is a weighted sum of *all* constituent components: $\vec{p} = W_1\vec{a} + W_2\vec{b}$.

Finally, the lexical function (**lexfunc**) model of Baroni and Zamparelli (2010) and Coecke et al. (2010) takes inspiration from formal semantics to characterize composition as function application. In particular, in modifier-head phrases, the modifier is treated as a linear function operating on the head vector. Given that linear functions can be expressed by matrices and their application by matrix-by-vector multiplication, the modifier is represented by a matrix A to be multiplied with the modifier vector $\vec{b}$, so that: $\vec{p} = A\vec{b}$.

We use the DISSECT toolkit[7] to estimate the parameters of the composition methods and derive phrase vectors. In particular, DISSECT finds optimal parameter settings by learning to approximate corpus-extracted phrase vector examples with least-squares methods (Dinu et al., 2013). We use as training examples all the modifier-head phrases that contain a modifier of interest and occur at least 50 times in our source corpus (see Section 3.3 below).

3.2 Asymmetric similarity measures

Several measures to identify word pairs that stand in an instance-class relationship by comparing their vectors have been proposed in the recent distributional semantics literature (Kotlerman et al., 2010; Lenci and Benotto, 2012; Weeds et al., 2004).[8] While the task of deciding if u is in class v is typically framed (also by distributional semanticists) in binary, yes-or-no terms, all proposed measures return a continuous numerical score.[9] Consequently, we conjecture that they might be well-suited to capture the graded notions of class membership and typicality we recorded in NBP.[10]

In what follows, we use $w_x(f)$ to denote the weight (value) of feature (dimension) f in the distributional vector of term x. F_x denotes the set of features (dimensions) in the vector of x such that $w_x(f) > t$, where t is a predefined threshold to decide whether a feature is active.[11] Importantly,

[7] http://clic.cimec.unitn.it/composes/toolkit/

[8] We speak of "instance-class relations" in a very broad and loose sense, to encompass classic relations such as hyponymy but also the fuzzier notion of lexical entailment.

[9] SVM classifiers have also been shown by Baroni et al. (2012) to be well-suited for entailment detection, but they do not naturally return continuous scores.

[10] Subjects had to answer a yes/no question concerning class membership, but by averaging their response we derive continuous membership scores.

[11] The obvious choice for t is 0. However, when working with the low-rank spaces described in Section 3.3 below, we set t to 0.1, since after SVD/NMF smoothing we observe

all measures assume non-negative values.

Most asymmetric measures proposed in the literature build upon the *distributional inclusion hypothesis*, stating that "if u is a semantically narrower term than v, then a significant number of salient distributional features of u is included in the feature vector of v as well" (Lenci and Benotto, 2012). In our terminology, u is the potential instance, and v is the class. We re-implement all the measures adopted by Lenci and Benotto, namely **weedsprec**, **cosweeds**, **clarkede** and **invcl** (see their paper for the original references):

$$weedsprec(u,v) = \frac{\sum_{f \in F_u \cap F_v} w_u(f)}{\sum_{f \in F_u} w_u(f)}$$

$$cosweeds(u,v) = \sqrt{weedsprec(u,v) \times cosine(u,v)}$$

$$clarkede(u,v) = \frac{\sum_{f \in F_u \cap F_v} \min(w_u(f), w_v(f))}{\sum_{f \in F_u} w_u(f)}$$

$$invcl(u,v) = \sqrt{clarkede(u,v) \times (1 - clarkede(u,v))}$$

The cosweeds formula combines weedsprec with the widely used symmetric *cosine* measure:

$$cosine(u,v) = \frac{\sum_{f \in F_u \cap F_v} w_u(f) \times w_v(f)}{\sqrt{\sum_{f \in F_u} w_u(f)^2} \times \sqrt{\sum_{f \in F_v} w_v(f)^2}}$$

Finally, we experiment with the carefully crafted **balapinc** measure of Kotlerman et al. (2010):

$$balapinc(u,v) = \sqrt{lin(u,v) \cdot apinc(u,v)}$$

where the *lin* term is computed as follows:

$$lin(u,v) = \frac{\sum_{f \in F_u \cap F_v} w_u(f) + w_v(f)}{\sum_{f \in F_u} w_u(f) + \sum_{f \in F_v} w_v(f)}$$

The balapinc score is the geometric average of a symmetric similarity measure (*lin*) and the strongly asymmetric *apinc* measure, that takes large values when dimensions with high values in the vector of the more specific term are also high in the vector of the more general term (refer to Kotlerman et al. (2010) for the apinc formula).

widespread low-frequency noise.

3.3 Distributional semantic spaces

We extract co-occurrence information from a corpus of about 2.8 billion words obtained by concatenating ukWaC,[12] Wikipedia[13] and the British National Corpus.[14] With DISSECT, we build co-occurrence vectors for the top 20K most frequent lemmas in the source corpus (plus any NBP term missing from this list). We treat the top 10K most frequent lemmas as context elements. We consider context windows of 2 and 20 words on the two sides of the targets. We weight the vectors by non-negative Pointwise Mutual Information and Local Mutual Information (Evert, 2005). We experiment with vectors in the resulting full-rank (10K-dimensional) semantic spaces as well as with vectors in spaces of ranks 100 and 300. Rank reduction is performed by applying the Singular Value Decomposition (Golub and Van Loan, 1996) or Non-negative Matrix Factorization (Lee and Seung, 2000). It is customary to represent the output of these operations directly in a dense low-dimensional space. However, the asymmetric similarity measures we use assume sparse vectors (or the "inclusion" criterion would be meaningless), so we project back the outcome of SVD and NMF to sparse 10K-dimensional but low-rank spaces. In total, we explore 20 distinct semantic spaces.

We also collect co-occurrence vectors for the phrases needed to estimate the composition method parameters (see Section 3.1 above). We use DISSECT's "peripheral space" option to project the phrase raw count vectors into the various spaces without affecting their structure.

Due to memory constraints, we restrict evaluation in the full-rank spaces to the *wadd* and *mult* models.

4 Experiments

Given the methods described above, the main question we want to answer is: Which combination of compositional model and asymmetric similarity measure yields a better fit for the data in the NBP dataset?

We start however with a sanity check on the ability of the measures to capture the *direction* of the instance-class membership relation. Even a measure that is good at capturing degrees of membership/typicality won't be of much practical use

[12] http://wacky.sslmit.unibo.it
[13] http://en.wikipedia.org
[14] http://www.natcorp.ox.ac.uk

clarkede	weedsprec	balapinc	cosweeds	invcl
Low-rank spaces				
10	8	11	8	7
Full-rank spaces				
2	4	4	4	2

Table 3: Number of spaces (over totals of 16 low-rank and 4 full-rank spaces) in which each measure was able to predict class membership direction significantly above chance.

if it is not able to tell us which item in a pair is the instance and which is the class.

Detecting membership direction As described in Section 2 above, NBP also contains single-word $h \rightarrow c$ pairs (*parrot*→ *pet*). We extracted the subset of those that all judges considered to be in the category membership relation, and we checked them manually to make sure that the direction was one-way only. This resulted in a set of 639 pairs where the membership relation holds unidirectionally. We tested all combination of semantic spaces (Section 3.3) and asymmetric similarity measures (Section 3.2) on the task of assigning a higher score to the pairs in the $h \rightarrow c$ (vs. $c \rightarrow h$) direction (e.g., ($score$(*parrot* $\rightarrow$ *pet*) $>$ $score$(*pet* $\rightarrow$ *parrot*)). Table 3 reports, for each measure, the number of spaces in which the measure was able to predict membership direction significantly better than chance (binomial test, $p < 0.05$). We report results on full- and low-rank (SVD, NMF) spaces separately since, as discussed above, for most composition models we can only use the latter. We observe that all measures are able to significantly detect directionality in at least some spaces. For all the analyses below, we exclude from further testing the space-measure combinations that failed to pass this sanity check, since they are clearly failing to capture properties pertaining to the instance-class relation (if a combination is not able to tell that it is a *parrot* that is a *pet*, and not *vice versa*, there is no point in asking if the same combination is able to model how typical a *dead parrot* is as a *pet*).

Modeling typicality ratings of $mh \rightarrow c$ pairs Next, for each of the remaining spaces, we first performed composition as described in Section 3.1 above to build the representations for the nominal phrases in the NBP dataset, and then computed asymmetric similarity scores for pairs made of a phrase and the corresponding potential class.

We computed the correlations between mean human membership or typicality ratings and the scores produced with each combination of composition model, similarity measure and space. The resulting performance profiles for membership and typicality are very highly correlated ($r = .99$), and we thus report only the latter. We leave it to further work to devise measures that are more specifically tuned to capture membership or typicality.

Table 4 reports the top correlation coefficients between typicality judgments and scores of each $mh \rightarrow c$ pair (*dead parrot*→ *pet*) across spaces, organized by measures and composition methods. The best correlation is achieved with the weedsprec measure using the mult composition model in a full-rank space (precisely that of context window size 2 and ppmi weighting). Recall that mult returns the component-wise product of the vectors it combines. Thus, modification under mult is carried out by picking only those features of the head that are also present in the modifier, and enhancing them by a factor given by the modifier's feature value. The weedsprec measure is then given by the weighted proportion of active features in mh that are also active in c. Therefore, the more the modifier shares features with the parent category, the higher weedsprec will be. This might explain why weedsprec is a good fit for the mult model in measuring degrees of category typicality.

Looking at composition methods, there is no evidence that the more complex, matrix-based fulladd and lexfunc approaches are performing any better than the simple multiplicative and additive methods. Indeed, mult shows the most consistent overall performance, confirming the conclusion of Blacoe and Lapata (2012) that, at the present time, when it comes to composition, "simpler is better". A related point emerges from the comparison of the low- and full-rank results for mult and wadd. The smoothing process due to dimensionality reduction is quite disruptive for the current asymmetric measures, that are based on feature inclusion. This is a further reason to stick to simpler composition methods, that can be applied directly in the full-rank spaces.

Regarding the measures themselves, we see that cosweeds, that balances weedsprec with the classic cosine score, is the most robust, returning good

	clarkede	weedsprec	balapinc	cosweeds	invcl
Low-rank spaces					
dil	9*	15*	16*	19*	8*
fulladd	17*	16*	12*	24*	−3
lexfunc	17*	12*	12*	27*	−2
mult	13*	19*	19*	29*	12*
wadd	14*	14*	16*	27*	−2
Full-rank spaces					
mult	9*	**39***	**33***	**36***	**15***
wadd	**30***	34*	31*	35*	14*

Table 4: Percentage Pearson r between asymmetric similarity measures and $mh \rightarrow c$ typicality ratings. *$p < 0.001$

	clarkede	weedsprec	balapinc	cosweeds	invcl
Low-rank spaces					
dil	5	−1	−1	−2	7*
fulladd	10*	7*	5+	7+	−2
lexfunc	**15***	9*	10*	18*	−2
mult	4+	14*	13*	15*	**9***
wadd	7+	7*	9*	12+	−2
Full-rank spaces					
mult	1	**25***	**21***	**24***	5+
wadd	11*	18*	13*	20*	2

Table 5: Percentage Pearson r between asymmetric similarity measures and $mh \rightarrow c$ typicality ratings where $h \rightarrow c$ scores have been partialed out. *$p < 0.001$, +$p < 0.05$

results across all composition methods. On the other hand, the related clarkede and invcl measures turn out to be quite brittle.

The highly significant correlations show that the measures do capture to some extent the patterns of variance in the data. However, when considering potential practical applications, even the highest reported correlation (.39) is certainly not impressive, indicating that there is plenty of room for further research into developing better composition methods and/or membership/typicality measures.

Focusing on the modifier effect for $mh \rightarrow c$ pairs The typicality judgment for *dead parrot* as a *pet* is influenced by two factors: how typical *parrots* are as *pets*, and how much more or less typical *dead parrots* are as *pets*, as opposed to *parrots* in general. A good model must be able to capture both factors (and this is what we tested above). However, we are also interested in assessing to what extent the models are capturing the modification effectproper, as opposed to the overall degree of typicality of the h concept as member of the c category. To focus on the modification factor, we partialed out the $h \rightarrow c$ (*parrot→pet*) ratings from the $mh \rightarrow c$ (*dead parrot→pet*) ratings and from the corresponding model scores (that is, we correlated the residuals of $mh \rightarrow c$ ratings and model-produced scores after regressing the $h \rightarrow c$ ratings on both). The results are shown in Table 5. Correlations are lower overall, but the general picture from the previous analysis still holds, confirming that the computational models are (also) capturing modifier effects. Interestingly, wadd, dil and fulladd generally undergo larger performance drops than mult and lexfunc. Evidently, models like the latter, in which the modifier selects the relevant features from the head, are better suited to explain modification than the former, in which the modifier features are just added to those of the head by means of a linear combination.

Modeling typicality ratings of $mh \rightarrow h$ pairs We repeated the first analysis for pairs of the type $mh \rightarrow h$ (*dead parrot→ parrot*). The results, shown in Table 6, are lower than in the previous analysis. This is probably due to the fact that, as discussed in Section 2, when the very same concept is used as phrase head and category, judgments are subject to a strong ceiling effect, and none of our measures is designed to flatten out above a certain threshold. Indeed, if we measure the skewness of the typicality ratings,[15] we obtain that, while for $h \rightarrow c$ and $mh \rightarrow c$ the skewness is of -1.9 and -1.5, respectively, for $mh \rightarrow h$ it gets to -3.9.

In any case, the results confirm the brittleness of the clarkede and invcl measures. The linguistically motivated lexfunc model emerges here as a competitive alternative to the simpler models. Still, the best results are obtained with mult and cosweeds (on the full-rank, context window size 20, ppmi weighted space). Notably, weedsprec applied to a pair of the type $mh \rightarrow h$, where the phrase is constructed using the mult model, results in a constant value of 1, whatever the modifier and the head noun is. This is due to the fact that the features of a phrase composed using mult are a subset of the features of the head,[16] and in this case the head is the same as the category. Therefore, by definition, weedsprec yields a score of 1 for every pair, the variance is null and hence the correlation is unde-

[15] A skewness factor of 0 means that the distribution is balanced around the mean, while the more negative the coefficient is, the more the left tail is longer and the distribution is concentrated to the right (toward high typicality values in our case).

[16] In set notation: $F_u \cap F_v = F_u$ since $F_u \subseteq F_v$

	clarkede	weedsprec	balapinc	cosweeds	invcl
Low-rank spaces					
dil	2	−1	−2	−3	4
fulladd	5+	5+	2	1	−1
lexfunc	**14***	8*	14*	17*	−1
mult	3	-	13*	15*	**5+**
wadd	6+	8*	7+	6	−3
Full-rank spaces					
mult	−2	-	**18***	**19***	−2
wadd	7*	**13***	7*	12*	−2

Table 6: Percentage Pearson r between asymmetric similarity measures and $mh \rightarrow h$ typicality ratings. $^*p < 0.001$, $^+p < 0.05$

fined. As a consequence, in this case cosweeds, which is the geometric mean between weedsprec and cosine, reduces to cosine similarity! The latter might be effective in capturing the degree of similarity between the phrase and its potential category but, as a symmetric measure, it cannot, alone, provide a full account of category typicality effects.

5 Conclusion

We introduced the challenge of quantifying the impact of modification on the meaning of noun phrases to the computational linguistics community. We presented a new dataset that collects membership and typicality ratings for modifier-head phrases with respect to the category represented by the head as well as a broader category. Since accounting for modifier distortion requires semantic representations of phrases and modeling graded judgments, we consider this an ideal testbed for compositional distributional semantics.

In the interaction between compositional models and directional similarity measures, we have observed that simpler models yield better results. Specifically, mult and wadd are economical composition models than can be applied on full-rank spaces, which in turn work best with our similarity measures.

Psychologists studying modification effects in concept combination have proposed models that are usually quite complex, relying on hand-crafted feature definitions and making very strong assumptions about the combination process (see for example Cohen and Murphy (1984), Smith et al. (1988)). Some of these assumptions have led other researchers to argue that prototypes do not compose at all (Connolly et al., 2007). In contrast, the approach we borrow from distributional semantics, while only mildly successful for now, has the advantage of being very simple both in its construction and application, and in the assumptions that it makes.

Also notable is that we are putting under the same umbrella tasks that have been traditionally tackled separately. For example, among the effects present in the dataset, we can find both word sense disambiguation (see discussion at the end of Section 2) and what Murphy (2002) calls "knowledge effects" (e.g., a *plane* makes a very good *machine*, but a *paper plane* doesn't). Moreover, these effects can also interact (people know that a *human egg* is actually a single, small cell, and hence not even cannibals would consider it satisfactory food). We can thus explore the empirical question of whether all these related phenomena can be tackled together, with a single model accounting for all of them.

In conclusion, the challenge that we introduced brings together concept combination and non-subsective modification phenomena studied in psychology and theoretical linguistics, and tries to handle them with the standard machinery of computational linguistics. This challenge has proved quite difficult for current tools, but this is exactly what we expected in the first place. Our goal, from the outset, was to create a task that could help us delimiting the boundaries of computational methods for characterizing human concepts, while delimiting, at the same time, the notion of human concepts itself.

Acknowledgments

We acknowledge ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the EMNLP GEMS Workshop*, pages 1–10, Edinburgh, UK.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, pages 1183–1193, Boston, MA.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-Chieh Shan. 2012. Entailment above

the word level in distributional semantics. In *Proceedings of EACL*, pages 23–32, Avignon, France.

Marco Baroni. 2013. Composition in distributional semantics. *Language and Linguistics Compass*, 7(10):511–522.

William Blacoe and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of EMNLP*, pages 546–556, Jeju Island, Korea.

Gemma Boleda, Eva Maria Vecchi, Miquel Cornudella, and Louise McNally. 2012. First order vs. higher order modification in distributional semantics. In *Proceedings of EMNLP*, pages 1223–1233, Jeju Island, Korea.

Gemma Boleda, Marco Baroni, Louise McNally, and Nghia The Pham. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of IWCS*, pages 35–46, Potsdam, Germany.

Graham Chapman. 1989. *The complete Monty Python's flying circus : all the words*. Pantheon Books, New York.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36:345–384.

Benjamin Cohen and Gregory L Murphy. 1984. Models of concepts. *Cognitive Science*, 8(1):27–58.

Louise Connell and Michael Ramscar. 2001. Using distributional measures to model typicality in categorization. In *Proceedings of CogSci*, pages 226–231, Edinburgh, UK.

Andrew Connolly, Jerry Fodor, Lila Gleitman, and Henry Gleitman. 2007. Why stereotypes don't even make good defaults. *Cognition*, 103(1):1–22.

Ann Copestake and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15–67.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: rationale, evaluation and approaches. *Natural Language Engineering*, 15:459–476.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. General estimation and evaluation of compositional distributional semantic models. In *Proceedings of ACL Workshop on Continuous Vector Space Models and their Compositionality*, pages 50–58, Sofia, Bulgaria.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences*. Ph.D dissertation, Stuttgart University.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Gene Golub and Charles Van Loan. 1996. *Matrix Computations (3rd ed.)*. JHU Press, Baltimore, MD.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of GEMS*, pages 33–37, Uppsala, Sweden.

James Hampton. 1991. The combination of prototype concepts. In Paula Schwanenflugel, editor, *The psychology of word meanings*, pages 91–116. Erlbaum, Hillsdale, NJ.

Charles Kalish. 1995. Essentialism and graded membership in animal and artifact categories. *Memory and Cognition*, 23(3):335–353.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31:91–113.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Daniel Lee and Sebastian Seung. 2000. Algorithms for Non-negative Matrix Factorization. In *Proceedings of NIPS*, pages 556–562.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of *SEM*, pages 75–79, Montreal, Canada.

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

Louise McNally. 2013. Modification. In Maria Aloni and Paul Dekker, editors, *Cambridge Handbook of Semantics*. Cambridge University Press, Cambridge, UK. In press.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Gregory Murphy. 2002. *The Big Book of Concepts*. MIT Press, Cambridge, MA.

Massimo Poesio, Simone Ponzetto, and Yannick Versley. 2010. Computational models of anaphora resolution: A survey. `http://clic.cimec.unitn.it/massimo/Publications/lilt.pdf`.

Edward Smith and Daniel Osherson. 1984. Conceptual combination with prototype concepts. *Cognitive Science*, 8(4):337–361.

Edward E Smith, Daniel N Osherson, Lance J Rips, and Margaret Keane. 1988. Combining prototypes: A selective modification model. *Cognitive Science*, 12(4):485–527.

Robert Speer and Catherine Havasi. 2013. ConceptNet 5: A large semantic network for relational knowledge. In Iryna Gurevych and Jungi Kim, editors, *The People's Web Meets NLP*, pages 161–176. Springer, Berlin.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of COLING*, pages 1015–1021, Geneva, Switzerland.

Edward Wisniewski. 1997. When concepts combine. *Psychonomic Bulletin & Review*, 4(2):167–183.

Fabio Zanzotto, Ioannis Korkontzelos, Francesca Falucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of COLING*, pages 1263–1271, Beijing, China.

German Kruszewski, Marco Baroni: Dead parrots make bad pets: Exploring modifier effects in noun phrases. TomÃ¡Å¡ Neugebauer: Topic-to-question generation. Yllias Chali, Sadid A. Hasan: Towards Topic-to-Question Generation.