Modelling compounding across languages with analogy and composition

Aotao Xu[1,2] **(a26xu@cs.toronto.edu)**
Charles Kemp[3] **(c.kemp@unimelb.edu.au)**
Lea Frermann[2] **(lea.frermann@unimelb.edu.au)**
Yang Xu[1,4] **(yangxu@cs.toronto.edu)**

[1]Department of Computer Science, University of Toronto
[2]School of Computing and Information Systems, University of Melbourne
[3]School of Psychological Sciences, University of Melbourne
[4]Cognitive Science Program, University of Toronto

Abstract

Compounding is a common word formation process in many languages around the world. Previous semantic analyses of compounding suggest that analogy and composition are crucial cognitive processes that underlie the formation of new compounds, but these processes are typically considered separately. Here, we formulate a computational model of compounding that integrates both analogy and composition. Compared to simpler baselines, we show that the model combining both processes achieves the best performance in predicting the constituents of attested compounds in English, Chinese, and German. Our work extends previous semantic-based accounts of compounding via a computational approach that can be evaluated using large-scale crosslinguistic data.

Keywords: compounding; composition; analogy; lexical semantics; computational modelling

Introduction

Compounding refers to the process of creating compound words and is the most productive source of new lexical items in many languages around the world (Brinton & Traugott, 2005; Bauer, 2011; Schlücker, 2019). Although there are different definitions of compounds in the literature, one minimal definition holds that they are lexical items that express specific concepts by combining two or more existing words. For example, the concept defined by "a portable wireless telephone" is expressed by the compound *cellphone* in English and the compound *shǒu-jī* (lit. hand-machine) in Chinese, which both combine two separate words. Here, we aim to understand and model the cognitive processes that underlie compounding across languages.

Existing accounts of compounding consider both the form and meaning of compounds (e.g., Štekauer & Lieber, 2005). Here, we are primarily interested in the role of compounds in the expression of concepts, so we focus on reviewing previous semantic analyses of compounding. The traditional approach views compound interpretation as a process of identifying the relation between its constituents and aims to describe regularities in these relations (e.g., Levi, 1978; Lieber, 1983; Jackendoff, 2010; Levin, Glass, & Jurafsky, 2019). In turn, regularities in these relations constrain the possible compound expressions of intended concepts. For example, Levi (1978) argues that head-modifier relations in noun-noun compounds tend to belong to one of nine meta-relations, while other intended relations tend to be explicitly encoded in the compound (e.g., *eating* should not be removed from *fish-eating dinosaurs*). In a recent study, Levin et al. (2019) show that English noun compounds denoting artifacts tend to invoke head-modifier relations highlighting events of their creation or use, whereas noun compounds denoting natural kinds tend to invoke relations highlighting their so-called essence (e.g., perceptual properties). The examples above show that traditional semantic analyses of compounding differ in scope and granularity. However, since they focus on the semantic composition of compound constituents, these accounts tend to ignore non-compositional processes in compounding.

An alternate approach examines the semantics of compounding by considering the general cognitive process of analogy (e.g., Ryder, 1994; Booij, 2010; Klégr & Čermák, 2010). In this view, one or more existing compounds that share the same constituent can form a productive pattern that can be modified to create new compounds; these new compounds are regarded as analogical compounds. Crucially, the shared constituent in a productive pattern tends to develop bound meanings that shift away from its original lexicalized meaning, which are then inherited by new compounds based on the pattern (Booij, 2010). For example, *blue-collar* and *white-collar* form a productive pattern that led to the creation of the new compound *green-collar* (Mattiello & Dressler, 2018). Here, the pattern *X + collar* signifies membership in a certain profession, which is distinct from compositional formations like *dog collar*. Compounding can create word combinations that share no constituent with any existing compound and thus do not involve analogy, but the non-compositional nature of analogical compounds suggests that analogy may nonetheless complement compositional accounts of compounding.

In this study, we formulate a computational account of compounding that combines both composition and analogy. Specifically, we hypothesize that speakers create new compounds by inferring how word combinations may be interpreted by listeners. Listener interpretation may involve the composition of constituents, or in the case where one constituent overlaps with existing compounds that form a productive pattern, it may involve composing the bound meaning of the pattern with the other constituent. The speaker finally chooses a compound with an interpretation that is very similar to the intended concept. We illustrate this idea in Figure 1.

Our work builds on a line of computational work on vector-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

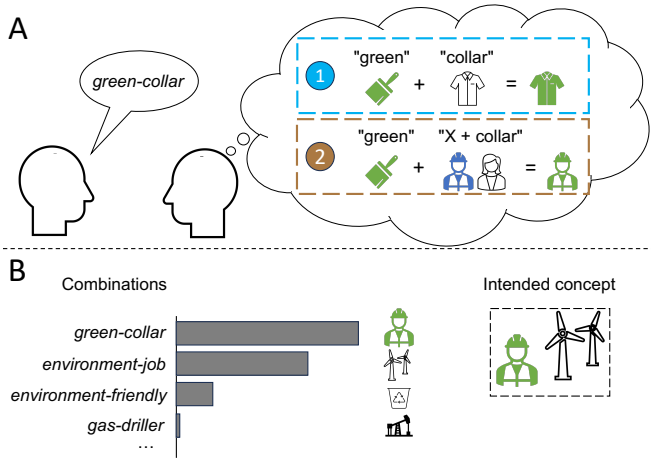

Figure 1: Analogy and composition in compounding. Panel (A) shows a listener that tries to interpret the compound *green-collar*. The listener chooses between 1) a compositional strategy and 2) an analogical strategy, in which the meaning of the pattern is based on *blue-collar* and *white-collar*. Panel (B) illustrates predictions on possible compound expressions for the concept "belonging to a green industry," which is illustrated on the right. Grey bars on the left illustrate the probability of choosing a compound, and the adjacent icons illustrate the corresponding listener interpretations.

based semantic composition (e.g., Mitchell & Lapata, 2010; Dima, de Kok, Witte, & Hinrichs, 2019). In this approach, word meanings are represented by real-valued vectors in semantic space, and semantic composition is modeled by composition functions that take word vectors as input and produce a new vector as output. In addition to using them to infer the meaning of novel combinations, previous work has applied composition functions to derive measures to predict human behavioural data on the processing of both existing and novel combinations (Vecchi, Marelli, Zamparelli, & Baroni, 2017; Günther & Marelli, 2016; Marelli, Gagné, & Spalding, 2017; Hsieh, Marelli, & Rastle, 2025). Here we extend these studies by using vector-based composition to reconstruct attested combinations given their meanings.

Our work is related to but distinctive from existing computational models of analogy in morphology (e.g., Krott, Baayen, & Schreuder, 2001; Plag, Kawaletz, Arndt-Lappe, & Lieber, 2023). With regards to compounding, analogy-based models have been applied to predicting the use of linking morphemes (Krott et al., 2001; Krott, Schreuder, Baayen, & Dressler, 2007) and the position of word stress (Arndt-Lappe, 2011). For a new compound, these models typically compute the probability of an outcome (e.g., stress position) based on the similarity between the new item and a group of existing items in the lexicon that are associated with the outcome. In the current study, we propose a model that computes the probability of a possible combination for an intended concept by using the concept's similarity with the meanings of existing compounds that belong to a productive pattern.

In the following, we first formulate our account in computational terms. We then compare our model to simpler models by using them to predict the constituents of attested compounds in English, Chinese and German. Finally, we discuss the implications of our work and future directions.

Computational Formulation

Our account assumes the following setting. Let $\mathcal{L}$ be the lexicon which contains the set of existing words, and let D be the set of existing compounds. Let C be a semantic space of real vectors, such that each word w corresponds to the semantic vector $c_w \in C$.

We define a compound family as a set containing existing compounds with the same head or modifier. We assume each family is partitioned into subfamilies, and each subfamily corresponds to a distinct productive pattern.[1] In this study, we focus on analogy with respect to compounds sharing the same head word (e.g., *X + collar*), and we leave modifier-based analogy for future work.

Probabilistic Model of Compounding

Given an intended concept $c \in C$ that is not expressed by compounds in D, we wish to predict an expression that consists of a head word $h \in \mathcal{L}$ and a modifier word $m \in \mathcal{L}$. We formalize this computational problem in probabilistic terms by applying Bayes rule:

$$p(m,h|c,D) \propto p(c|m,h,D)p(m|D)p(h|D). \quad (1)$$

There are three terms on the righthand side of Equation 1. The rightmost terms $p(m|D)$ and $p(h|D)$ represent the prior probability of selecting a specific modifier and head without considering the intended concept. The third term $p(c|m,h,D)$ is the likelihood that captures the similarity between c and listener interpretations of an observed combination. Importantly, all terms are dependent on statistical tendencies among existing compounds in the set D.

Likelihood. To formulate the likelihood, we first formalize listener interpretations via composition and analogy. Since we assumed meanings are represented in semantic space, a natural way to model composition is to use vector-based composition functions (e.g., Mitchell & Lapata, 2010). Let $f(\cdot,\cdot)$ be a general composition function that takes a pair of head and modifier and outputs a vector in C that represents the compositional interpretation. Similarly, let $g_i(\cdot|h)$ be a function that composes a modifier and existing compounds in the i-th subfamily associated with head h, such that it outputs a vector that represents an analogy-based interpretation. These functions correspond to processes illustrated in Figure 1A.

We formulate the likelihood in Equation 1 by using the similarity between c and listener interpretations of m and h.

[1] For example, the subfamily containing *software* and *freeware* is separate from the subfamily containing *tableware* and *kitchenware*.

Because a listener may choose between the compositional interpretation and an analogy-based interpretation, if available, we define the likelihood as an average of similarities:

$$p(c|m,h,D) \propto q_0 sim(c, f(m,h)) + \sum_i q_i sim(c, g_i(m|h)) \quad (2)$$

$$sim(c_1, c_2) \propto \exp(s \cdot cos(c_1, c_2)) \quad (3)$$

where $0 \leq q_i \leq 1$ is the probability of choosing a specific interpretation, $s > 0$ is a sensitivity parameter, and $cos(\cdot,\cdot)$ is cosine similarity. We choose to use cosine because it is standard in previous applications of composition functions (e.g., Mitchell & Lapata, 2010).[2]

We define the probability of invoking a specific interpretation by using the prevalence of each pattern among lexicalized compounds. Let N_h be the size of the family associated with h, and let $N_{h,i}$ be the size of its i-th subfamily. We define q_j by using family and subfamily sizes:

$$q_0 = \frac{\theta}{N_h + \theta},\ q_{i>0} = \frac{N_{h,i}}{N_h + \theta} \quad (4)$$

where $\theta > 0$ is a pseudocount that determines the frequency of invoking composition. Equation 4 is inspired by previous work showing that the ease of understanding combination w is sensitive to the frequency of head-modifier relations associated with the constituents of w (Gagné & Shoben, 1997; Maguire, Maguire, & Cater, 2010).

Priors. In principle, any open-class word in the lexicon can be chosen as the head or modifier, but the distribution of compound family sizes also tends to be skewed. Previous work in psycholinguistics suggests that morphological family size facilitates processing (e.g., De Jong, Feldman, Schreuder, Pastizzo, & Baayen, 2002; Martín, Bertram, Häikiö, Schreuder, & Baayen, 2004), and thus we hypothesize that words corresponding to larger families tend to be preferred by speakers:

$$p(h|D) \propto N_h + \alpha,\ p(m|D) \propto N'_m + \beta \quad (5)$$

where $\alpha, \beta > 0$ are pseudocounts of compound families, and N'_m is the number of compounds in D with modifier m. Smaller pseudocounts increase the preferences for reusing a word from a large family, and vice versa.

Functions of Compound Interpretation

We now specify the composition function $f(\cdot,\cdot)$ and the function $g_i(\cdot|h)$ that captures analogical reasoning with respect to the i-th pattern associated with h. Given a modifier m and a head word h, we define their semantic composition using the simple additive function (Mitchell & Lapata, 2008):

$$f(m,h) = c_m + c_h \quad (6)$$

Because Equation 6 composes the lexicalized meanings of m and h, the resulting vector represents a compositional interpretation (e.g., *dog* + *collar*).

If the head h is associated with a compound family in D, the interpretation may also be obtained from analogy. Let $r_{h,i}$ be the bound meaning of the i-th pattern associated with head h (e.g., the job-related meaning of *X* + *collar*). We define the interpretation via analogy with respect to this pattern by composing its bound meaning with the modifier word:

$$g_i(m|h) = c_m + r_{h,i} \quad (7)$$

In contrast to Equation 6, here the interpretation is based on a bound meaning at the pattern level.

The bound meaning $r_{h,i}$ should capture how existing compounds that form the corresponding pattern tend to relate to their modifiers.[3] Here, we represent relations in semantic space via vector differences (e.g., Rumelhart & Abrahamson, 1973; Vylomova, Rimell, Cohn, & Baldwin, 2016). Let $S_{h,i}$ be the i-th subfamily associated with head h, and m_w be the modifier of compound w. We define $r_{h,i}$ via the central tendency of these relations:

$$r_{h,i} = \frac{\mu_{h,i}}{||\mu_{h,i}||_2},\ \mu_{h,i} = \sum_{w \in S_{h,i}} c_w - \gamma c_{m_w} \quad (8)$$

where $\gamma \geq 0$ is an additional free parameter that governs the importance of the modifier in contributing to relational meaning. Equation 8 normalizes $r_{h,i}$ to ensure its magnitude is on par with c_m in Equation 7.

Note that in the case where a pattern is based on a single existing compound w, Equation 8 reduces to a single vector difference. If we plug this difference into Equation 7 and disregard the coefficients on each semantic vector, we obtain the parallelogram-model solution to a proportional analogy involving four concepts in semantic space (e.g., after observing *software*, the listener may infer its meaning by solving "hard" : "hardware" :: "soft" : ?).

Materials and Methods

In the current study, we evaluate a version of our model that assumes a single productive pattern per compound family.[4] To do so, we first compiled datasets of compound words drawn from English, Chinese, and German, which are commonly studied but show crosslinguistic differences in morphology (e.g., J. Xu & Li, 2014; Günther, Smolka, & Marelli, 2019). We then split each dataset into training and test sets, which simulate the lexicon, the set of existing compounds, and the concepts to be expressed via compounding.

Datasets. We compiled large datasets of English, Chinese, and German compounds. The English dataset consists of closed English compounds from the Large Database of English Compounds (Gagné, Spalding, & Schmidtke, 2019) and entries that are labeled as etymologically obtained from compounding in a scrape of Wiktionary (Wu & Yarowsky,

[2] We assume every c is on the unit sphere, which makes Equation 2 a mixture of von-Mises-Fisher distributions. In this case, the omitted normalizing constant is only a function of s.

[3] For example, *X* + *ware* may represent a computer program characterized by the modifier.

[4] All code used in analyses is available at `https://github.com/johnaot/Compounds-compose-analogy`

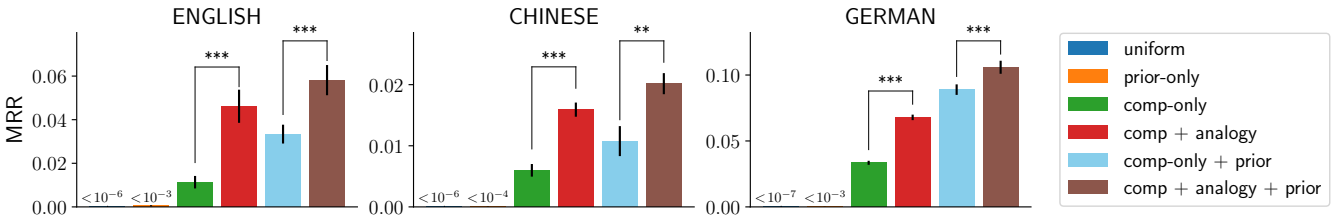

Figure 2: Comparing the full model against simpler variants across languages. Each bar shows the MRR averaged over five folds, and error bars indicate standard deviation across the folds. Stars indicate the p-values obtained from comparing models that involve both processes to composition-only models ("**" indicates $p < .01$ and "***" indicates $p < .001$).

2020); these compounds were intersected with Princeton WordNet (Fellbaum, 1998) after filtering out combinations involving stopwords and chemical names in WordNet. The Chinese dataset was obtained by using 2-character compounds in the Chinese Open WordNet (Wang & Bond, 2013), and words that are chemical names according to WordNet were removed. The German dataset was obtained from a large compilation of German compounds (Günther, Marelli, & Bölte, 2020), CELEX (Baayen, Piepenbrock, & Gulikers, 1995), and the same scrape of Wiktionary. We removed combinations involving stopwords or translations of English chemical names that were obtained from the Google Translate API. All compounds in the datasets contain exactly two constituents; we ignored linking elements in German and English compounds and we removed compounds containing other compounds.

In all languages, we implemented the semantic space using word embeddings trained on Wikipedia (Mikolov, Chen, Corrado, & Dean, 2013; Li et al., 2018; Yamada et al., 2020); all vectors were normalized. After ensuring the compounds and constituents map to vectors, we obtained 5,093 English compounds, 11,420 Chinese compounds, and 42,715 German compounds. Note that this implies the remaining Chinese character constituents also function as independent words.

Methods. For each language, we set the lexicon $\mathcal{L}$ to be the set of all constituent words in the compound dataset. The lexicon contained 2,646 words, 3,535 characters, and 9,487 words for English, Chinese, and German, respectively. We then randomly sampled at most 10,000 compounds from the dataset and applied 5-fold cross-validation to the subset. In all languages, we assumed all compounds are right-headed, except for Chinese verbs which we assumed are left-headed.[5]

We used the training set to estimate model parameters. For each training set, we randomly set aside 3/4 of compounds to initialize the set D. We used the remaining 1/4 as a development set to estimate the parameters via a limited grid search, searching over $\{10^n : n = -2,-1,..,2\}$ for α and β, $\{x/(1-x) : x = 0.5, 0.4, 0.3, 0.2\}$ for γ, $\{1, 10, 100\}$ for θ, and $\{0.01^{-1}, 0.02^{-1}, ..., 1\}$ for s. For every combination of α, β, γ and θ, we first set s by maximizing the likelihood $\prod_{D'} p(c|m,h,D)$, where D' contains compounds in the development set, and then we ranked all possible combinations given the meaning of each attested compound in D' according to the score given by Equation 1. We chose the set of values that maximizes the mean reciprocal rank (MRR) obtained from the ranks of attested combinations.

[5]Ceccagno and Basciano (2007) estimates that about 75% of Chinese compound verbs are left or two-headed. These words are usually combinations of V + V or V + N.

Language	Model type	Spearman ρ	p-value
English	comp-only	−0.276	$< .001$
	comp + analogy	−0.491	$< .001$
Chinese	comp-only	−0.058	$< .001$
	comp + analogy	−0.421	$< .001$
German	comp-only	−0.240	$< .001$
	comp + analogy	−0.424	$< .001$

Table 1: Correlation between predicted rank and family size.

We used the test set to evaluate the model by using Equation 1 to score each attested combination given its meaning. The full model was compared with simpler baselines: 1) a model that only involves composition, 2) a model that sets the priors to uniform, 3) a model that combines the previous two settings, 4) a prior-only model, and 5) a uniform distribution. The parameters of all baseline models were estimated in the same way as the full model based on the training set.

Results

We first evaluate the model by reconstructing an attested combination given its meaning. We then perform fine-grained analyses of bound meanings at the pattern level and across compound types.

Model evaluation. Figure 2 summarizes model performance through MRR. We observe that the full model obtained the best performance in all three languages. We also observe that the priors tend to improve model performance across languages and that the model with full likelihood but no prior is superior to composition-only models in English and Chinese. Similar observations can be made if we used top-k accuracy to evaluate the models.

Figure 3 shows the ranks predicted by the two models with priors and the (head-based) family size of every attested compound. We observe that the full model tends to dominate the simpler baseline across most family sizes. We also observe

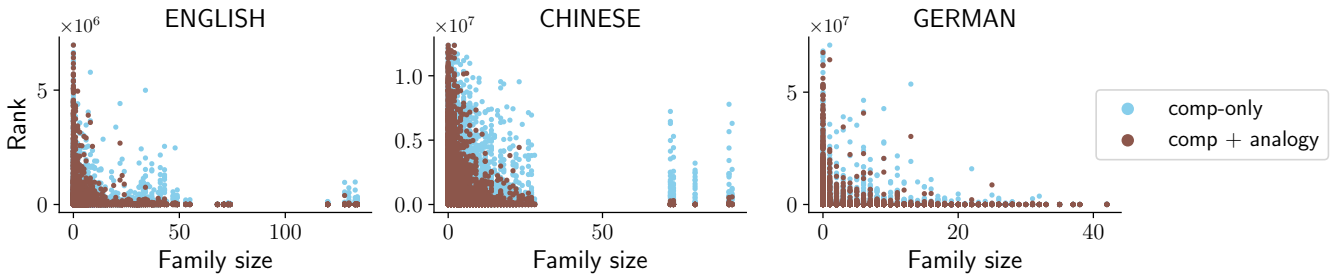

Figure 3: Comparing the full model and the composition-only model (including priors) in terms of attested-compound rank across family sizes. Predicted ranks lower in magnitude indicate better performance.

Head	Lexicalized meaning NNs	Bound meaning NNs
front	rear, frontmost	downtown, condo
head	coach, helm	dopy, cruelness

Table 2: Samples of nearest neighbours (NNs) of the lexicalized and bound meanings of certain head words.

Language	Spearman ρ	p-value	N
English	−0.487	< .001	4,187
Chinese	−0.316	< .001	8,625
German	−0.413	< .001	7,856

Table 3: Correlation between cosine similarity and model predicted rank.

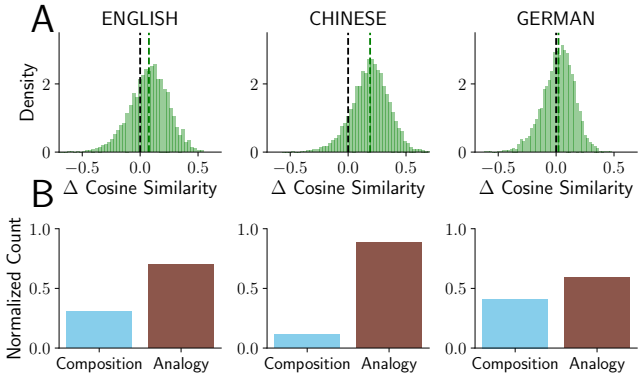

Figure 4: Semantic similarity between head and compound meanings. Panel (A) shows the distribution of differences in similarity between intended compound meanings and different meanings of the head; green lines show mean differences. Panel (B) shows the proportion of compounds that are more similar to lexicalized meanings and the proportion of compounds that are more similar to bound meanings.

that predicted rank tends to improve with family size for both models in the cases of English and German, but this only applies to the full model in the case of Chinese. This is confirmed by the correlations between predicted rank and family size in Table 1, and it may be in part explained by the previous observation that including priors which encode family size information improves performance.

Bound meanings at the pattern level. Table 2 shows a sample of the neighbours of the lexicalized meanings and bound meanings of the same head words, which were retrieved from the vocabulary of our embeddings. We can see that *X + front* clearly developed a property-related meaning (as in *beachfront*) distinct from the original location-related meaning, and *X + head* obtained a meaning that is more related to personal character (as in *metalhead*). These English examples illustrate the semantic shift of pattern-level meanings from the lexicalized meanings of head words.

We hypothesized that compounds tend to inherit the bound meanings of productive patterns to a degree greater than their degree of inheritance from the lexicalized meanings of head words. We quantified the degree to which a specific meaning is inherited by measuring the cosine similarity between the former and the intended meaning of a compound. Figure 4A shows the distribution of differences given by subtracting bound-meaning similarities from lexicalized-meaning similarities. We observe that bound meanings are on average much more similar to intended compound meanings in English ($t(4,186) = 29.57, p < 0.001$), Chinese ($t(8,624) = 109.40, p < 0.001$), and German ($t(7,855) = 13.01, p < 0.001$). A similar trend can be observed in Figure 4B. These results may in part explain the better performance of the full model since higher similarity between bound and intended meanings correlates with better predicted rank (see Table 3).

In Figure 4, we can also observe that differences between lexicalized and bound meanings in terms of their similarity with intended meanings tend to be smaller for German compounds relative to English compounds. This is consistent with an earlier empirical finding that German complex words tend to be more transparent than English ones (Günther et al., 2019) and earlier observations by linguists suggesting that the German lexicon is more motivated than the English lexicon (de Saussure, 1916; Ullmann, 1953).

Literal and non-literal head words. Recall that an advantage of analogy over composition is that it is capable of addressing certain non-compositional compounds (e.g., *green-collar*). We examined one type of non-compositional compound known as exocentric compounds, in which the

Model type	Compound type	MRR	N
comp-only	endocentric	0.0482	1,962
	exocentric	0.0257	2,818
comp + analogy	endocentric	0.0660	1,962
	exocentric	0.0565	2,818

Table 4: Comparing model performance across English endocentric and exocentric noun compounds.

Target	F. size	Rank	Top model predictions
taxicab	0	23	taxi-car, taxi-bus
mastermind	0	13k	evil-leader, shadow-leader
beachfront	3	1	beach-front, beach-land
metalhead	43	4	metal-boy, metal-man

Table 5: Examples of predictions from the full model. The second and third columns show the family size and predicted rank of the attested compound (target), respectively.

head is not a hypernym of the compound (e.g., *seahorse* is not a *horse*). Specifically, we hypothesized that the full model is able to better account for exocentric compounds than the best-performing composition-only baseline. Following A. Xu, Kemp, Frermann, and Xu (2024), we used WordNet (Fellbaum, 1998) to categorize English noun compounds into exocentric and non-exocentric (or endocentric) compounds. We compare model performance across compound types in Table 4. We observe that both models tend to perform better for endocentric compounds. Although we observe that the ranking of models remains the same across compound types, we also observe that the performance gap across types appears to be smaller for the full model.

We demonstrate the relative advantage of the full model with English examples in Table 5. Similar to a composition-only model, the full model is able to generate compositional expressions, which accounts well for compositional compounds like *taxicab* but not for non-compositional and non-analogical compounds like *mastermind*. However, the full model gains an advantage when an analogy-based interpretation of the attested combination is highly similar to its intended meaning. This can be demonstrated by the examples of *beachfront* and *metalhead*: using the bound meanings of *X + front* and *X + head*, the model assigned similar ranks to non-compositional attested combinations and hypothetical combinations that have similar interpretations but are compositional (e.g., *beach-land*, *metal-man*).

Discussion

We have presented a computational account of compounding that integrates the cognitive processes of composition and analogy. Using datasets of compounds from English, Chinese, and German, we tested computational models that reconstruct these compounds given their intended meanings. We showed that our full model, which integrates both processes, performs better than simpler baselines across all three languages.

Our work connects two separate lines of research on compounding by integrating the underlying cognitive processes. Specifically, traditional accounts focus on regularities among compositional relations between the constituents of attested compounds (e.g., Levi, 1978; Jackendoff, 2010), which correspond to the composition process, while a separate line of work appeals to the general cognitive process of analogy (e.g., Booij, 2010; Klégr & Čermák, 2010). In our account, a speaker creates a new compound by choosing a word combination that has an interpretation similar to their intended concept. Since compound interpretation may take place via the composition of the constituents or via analogical reasoning with respect to existing compounds, both processes are naturally integrated into our account.

Our account emphasizes the similarity between listener interpretations and intended meanings, and thus is closely related to functionalist accounts of compounding which argue that new compounds are shaped by a pressure for informativeness (Downing, 1977; A. Xu et al., 2024). However, these accounts typically assume that listeners interpret compounds via some type of semantic composition, and thus they have limited applicability to the creation of non-compositional compounds. Our findings suggest that functionalist accounts may provide a better explanation for certain non-compositional compounds by considering analogical processes in compound interpretation.

The full model examined in this paper can be extended in various ways. For example, there are various composition functions that are more involved than our simple additive function (e.g., Mitchell & Lapata, 2010; Marelli et al., 2017), which may be able to better capture the regularities of compositional compound interpretation. Moreover, we made the simplifying assumption that every compound family corresponds to a single productive pattern. In contrast, existing semantic analyses of compounding that emphasize analogy tend to consider patterns that are subsets of a compound family or local analogies based on a single existing compound (e.g., Booij, 2010; Mattiello & Dressler, 2018). One way to model productive patterns in a more fine-grained manner is to use clustering to identify meaningful compound subfamilies, which can be straightforwardly integrated into our computational model.

Since we formulated our account in computational terms, we were able to validate the role of composition and analogy in compounding by reconstructing attested compounds. In contrast, theoretical accounts of morphology have often associated analogy with unpredictability (Arndt-Lappe, 2015). Although recent computational models have in part addressed this concern (e.g., Krott et al., 2001; Plag et al., 2023), to our knowledge we have presented the first computational account that specifies how analogy can predict the head and modifier in new compounds. In the future, our approach may be extended to analyze a more typologically diverse set of languages and datasets of recent neologisms.

Acknowledgments

We thank anonymous reviewers for constructive comments on an earlier version of the paper. AX is funded in part by the U of T–UoM IRTG program, and CK is supported by ARC FT190100200. YX is supported by an Ontario Early Researcher Award #ER19-15-050 and NSERC Discovery Grant RGPIN-2018-05872.

References

Arndt-Lappe, S. (2011). Towards an exemplar-based model of stress in English noun–noun compounds. *Journal of Linguistics*, *47*(3), 549–585.

Arndt-Lappe, S. (2015). Word-formation and analogy. In C. Iacobini, P. Müller, I. Ohnheiser, S. Olsen, & F. Rainer (Eds.), *Word-formation : An international handbook of the languages of Europe* (Vol. 2, pp. 822–841). De Gruyter.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database.* University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Bauer, L. (2011). Typology of compounds. In R. Lieber & P. Štekauer (Eds.), *The Oxford handbook of compounding* (pp. 343–356). Oxford University Press.

Booij, G. (2010). Compound construction: Schemas or analogy? a construction morphology perspective. *Cross-disciplinary issues in compounding*, 93–108.

Brinton, L. J., & Traugott, E. C. (2005). *Lexicalization and language change*. Cambridge University Press.

Ceccagno, A., & Basciano, B. (2007). Compound headedness in Chinese: An analysis of neologisms. *Morphology*, *17*, 207–231.

De Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M., & Baayen, R. H. (2002). The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects. *Brain and Language*, *81*(1-3), 555–567.

de Saussure, F. (1916). *Course in general linguistics* (R. Harris, Trans.). London, UK: Bloomsbury Academic. (Translated edition published 2013)

Dima, C., de Kok, D., Witte, N., & Hinrichs, E. (2019). No word is an Island—A transformation weighting model for semantic composition. *Transactions of the Association for Computational Linguistics*, *7*, 437–451. Retrieved from https://aclanthology.org/Q19-1025/ doi: 10.1162/tacl_a_00275

Downing, P. (1977). On the creation and use of English compound nouns. *Language*, 810–842.

Fellbaum, C. (1998). *Wordnet: An electronic lexical database*. MIT press.

Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier–noun combinations. *Journal of experimental psychology: Learning, memory, and cognition*, *23*(1), 71.

Gagné, C. L., Spalding, T. L., & Schmidtke, D. (2019). LADEC: the large database of English compounds. *Behavior research methods*, *51*(5), 2152–2179.

Günther, F., & Marelli, M. (2016). Understanding karma police: The perceived plausibility of noun compounds as predicted by distributional models of semantic representation. *PloS one*, *11*(10), e0163200.

Günther, F., Marelli, M., & Bölte, J. (2020). Semantic transparency effects in German compounds: A large dataset and multiple-task investigation. *Behavior Research Methods*, *52*(3), 1208–1224.

Günther, F., Smolka, E., & Marelli, M. (2019). 'understanding' differs between English and German: Capturing systematic language differences of complex words. *Cortex*, *116*, 168–175.

Hsieh, C.-Y., Marelli, M., & Rastle, K. (2025). Compositional processing in the recognition of Chinese compounds: Behavioural and computational studies. *Psychonomic Bulletin & Review*, 1–12.

Jackendoff, R. (2010). The ecology of English noun-noun compounds. In *Meaning and the lexicon* (pp. 413–451). Oxford University Press Oxford.

Klégr, A., & Čermák, J. (2010). Neologisms of the 'on-the-pattern-of' type: Analogy as a word formation process. *The Prague school and theories of structure*, 229–241.

Krott, A., Baayen, R. H., & Schreuder, R. (2001). Analogy in morphology: modeling the choice of linking morphemes in Dutch. *Linguistics*, *39*(1), 51–93.

Krott, A., Schreuder, R., Baayen, R. H., & Dressler, W. U. (2007). Analogical effects on linking elements in German compound words. *Language and cognitive processes*, *22*(1), 25–57.

Levi, J. (1978). *The syntax and semantics of complex nominals*. New York: Academic Press.

Levin, B., Glass, L., & Jurafsky, D. (2019). Systematicity in the semantics of noun compounds: The role of artifacts vs. natural kinds. *Linguistics*, *57*(3), 429–471.

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 138–143). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/P18-2023

Lieber, R. (1983). Argument linking and compounds in English. *Linguistic inquiry*, *14*(2), 251–285.

Maguire, P., Maguire, R., & Cater, A. W. (2010). The influence of interactional semantic patterns on the interpretation of noun–noun compounds. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 288.

Marelli, M., Gagné, C. L., & Spalding, T. L. (2017). Compounding as abstract operation in semantic space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition*, *166*, 207–224.

Martín, F. M. d. P., Bertram, R., Häikiö, T., Schreuder, R., & Baayen, R. H. (2004). Morphological family size in a morphologically rich language: the case of finnish compared with dutch and hebrew. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1271.

Mattiello, E., & Dressler, W. U. (2018). The morphosemantic transparency/opacity of novel English analogical compounds and compound families. *Studia Anglica Posnaniensia*, *53*(1), 67–114.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mitchell, J., & Lapata, M. (2008, June). Vector-based models of semantic composition. In J. D. Moore, S. Teufel, J. Allan, & S. Furui (Eds.), *Proceedings of acl-08: Hlt* (pp. 236–244). Columbus, Ohio: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P08-1028

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, *34*(8), 1388–1429.

Plag, I., Kawaletz, L., Arndt-Lappe, S., & Lieber, R. (2023). Analogical modeling of derivational semantics: Two case studies. In S. Kotowski & I. Plag (Eds.), *The semantics of derivational morphology.* (pp. 103–141). De Gruyter.

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(1), 1–28.

Ryder, M. E. (1994). *Ordered chaos: The interpretation of English noun-noun compounds*. University of California Press.

Schlücker, B. (2019). *Complex lexical units: Compounds and multi-word expressions*. Berlin, Boston: De Gruyter.

Štekauer, P., & Lieber, R. (2005). *Handbook of word-formation* (Vol. 64). Springer Science & Business Media.

Ullmann, S. (1953). Descriptive semantics and linguistic typology. *Word*, *9*(3), 225–240.

Vecchi, E. M., Marelli, M., Zamparelli, R., & Baroni, M. (2017). Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive science*, *41*(1), 102–136.

Vylomova, E., Rimell, L., Cohn, T., & Baldwin, T. (2016, August). Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1671–1682). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P16-1158/ doi: 10.18653/v1/P16-1158

Wang, S., & Bond, F. (2013, October). Building the Chinese open Wordnet (COW): Starting from core synsets. In P. Bhattacharyya & K.-S. Choi (Eds.), *Proceedings of the 11th workshop on Asian language resources* (pp. 10–18). Nagoya, Japan: Asian Federation of Natural Language Processing. Retrieved from https://aclanthology.org/W13-4302

Wu, W., & Yarowsky, D. (2020, May). Computational etymology and word emergence. In N. Calzolari et al. (Eds.), *Proceedings of the 12th language resources and evaluation conference.* Marseille, France: European Language Resources Association. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.397

Xu, A., Kemp, C., Frermann, L., & Xu, Y. (2024). Word reuse and combination support efficient communication of emerging concepts. *Proceedings of the National Academy of Sciences*, *121*(46), e2406971121.

Xu, J., & Li, X. (2014). Structural and semantic non-correspondences between Chinese splittable compounds and their English translations: A Chinese-English parallel corpus-based study. *Corpus Linguistics and Linguistic Theory*, *10*(1), 79–101.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., & Matsumoto, Y. (2020). Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 23–30). Association for Computational Linguistics.