Rating-based property axes: Agentivity and split intransitivity

Eva Neu
UMass Amherst
eneu@umass.edu

Brian Dillon
UMass Amherst
bwdillon@umass.edu

Katrin Erk
UMass Amherst
kerk@umass.edu

Abstract

Many gradable properties have been found to be encoded as axes in embedding space. Most commonly, property axes are computed using seed words, but recent work has noted limitations to seed-based axes. Here, we present a novel methodology for computing property axes that is based on human ratings and does not require seeds. We apply this methodology to a particular problem at the syntax-semantics interface: which semantic properties of intransitive verbs affect their likelihood to occur in one of two syntactic structures, unergative and unaccusative. Comparing property axes that encode different semantic dimensions of the concept of agentivity, we find that properties like movement and being alive are a better predictor of the syntactic behavior of intransitives than goal-directedness or intentionality. We discuss the potential of rating-based axes for future work in semantics and at the syntax-semantics interface.

1 Introduction

1.1 Properties and property axes

In linguistics and many other areas of research, it is often necessary to obtain a measure of a gradable property for a set of words. For instance, one might want to know how dangerous different kinds of animals are considered to be, what degree of wealth is associated with different kinds of sports, or how attractive people find different cities. Traditionally, such measures have been obtained by collecting ratings for these properties from human participants.

Recent work on word embeddings has shown that gradable properties are also encoded by *property axes* in embedding space (e.g., Grand et al. 2022; Kozlowski et al. 2019; Garí Soler and Apidianaki 2020; Lucy et al. 2022). The projection of a word embedding on such an axis encodes to what extent the word is associated with the positive or the negative end of this semantic dimension. The most widely used method for obtaining axes for gradable properties is through manually defined seed words. For example, for *danger*, the two poles of the property could be described with {*dangerous*, *unsafe*} and {*safe*, *harmless*}, respectively. An axis for *danger* can then be obtained as the mean over all difference vectors $\vec{safe} - \vec{dangerous}$, $\vec{safe} - \vec{unsafe}$, $\vec{harmless} - \vec{dangerous}$, $\vec{harmless} - \vec{unsafe}$. How close the projection of a word embedding sits towards the *dangerous*/*unsafe* end of the axis then serves as a measure of how dangerous the entity denoted by the word can be considered to be.

However, while simple and easy to apply, seed-based property axes face problems (Erk and Apidianaki, 2024; Lucy et al., 2022; Antoniak and Mimno, 2021). Their performance in capturing human judgments depends critically on which seeds are chosen. However, there is no agreed-upon method for finding seed words, and in practice, researchers adopt a wide range of strategies for settling on a set of seeds (e.g., selecting them by hand, crowdsourcing, etc.). The repercussions of these different methodological choices are unclear. It is rare for different sets of seed words to be compared to each other in order to determine which seeds are more or less appropriate for the property in question. Furthermore, the performance of seeds can be affected by word frequencies (Ethayarajh et al., 2019), and they can reflect the researchers' underlying social and cultural biases.

Against this background, this study develops a new methodology for computing property axes. These axes are based on human judgments but can be used to compute measures for words for which no human ratings have been obtained. We apply this methodology to a particular problem in linguistic theory, the semantic correlates of split intransitivity, which we introduce next.

1.2 Split intransitivity

It has long been known that the syntactic structures in which a verb can appear depend on its lexical semantics (e.g., Dowty 1991; Jackendoff 1983; Levin 1993; Levin and Rappaport Hovav 1995; Van Valin Jr 1990). For instance, verbs denoting a transfer of possession such as *give* typically appear in a ditransitive syntax, while verbs denoting a causal relation such as *kill* tend to surface in a transitive syntax.

One area in which the effect of lexical semantics on syntax has been discussed most extensively is split intransitivity. Intransitive verbs are commonly assumed to allow for two different underlying syntactic structures, unergative and unaccusative, which can be distinguish from each other based on a wide variety of syntactic diagnostics (Burzio, 1981, 1986; Perlmutter, 1978). For instance, *freeze* but not *play* can surface in prenominal participle constructions (1):

(1) a. the frozen lake
b. *the played child

Moreover, *freeze* but not *play* can take a secondary predicate describing the result of the event (2):

(2) a. The lake froze solid.
b. *The child played tired. (i.e., became tired as the result of playing)

These contrasts can be captured by positing two different syntactic structures for intransitives, unergative and unaccusative. Unaccusative structures allow for prenominal participle constructions and resultative predicates; unergative structures do not. In (1) and (2), *freeze* behaves as an unaccusative and *play* as an unergative.

The unergative/unaccusative split is not a neat divide between two distinct classes of verbs. Rather, verbs have a gradient preference towards one structure or the other, with some leaning strongly towards an unergative structure, others towards an unaccusative structure, and others being more variable in their syntactic behavior (Sorace, 2000, 2011, 2004). Crucially, this tendency in a verb's syntactic behavior has been linked to its semantics: verbs that denote a more *agentive* activity appear to be more likely to be used as unergatives. For instance, *play* describes a more agentive activity than (intransitive) *freeze*, thus favouring an unergative structure.

Theoretically, such a correlation between an agentive meaning and an unergative syntax has been motivated based on *thematic roles*. The Unaccusativity Hypothesis developed in the generative tradition holds that in an unergative syntax, the sole argument is merged in the same position as the subject argument of transitives, whereas in an unaccusative syntax, it starts out in the same position as the object argument of transitives (Burzio 1981, 1986; Perlmutter 1978, but see Van Valin Jr 1990). This analysis suggests that the arguments of unergatives and unaccusatives should receive the same thematic role as the subject and object argument of transitives, respectively, with the argument of unergatives being assigned the thematic role of an Agent, and the argument of unaccusatives the thematic role of a Patient. Thus, a verb should be the more likely to adopt an unergative syntax the more agentive its meaning is.

Empirically, however, the link between agentivity and unergativity remains contested. It has been documented most extensively in a series of papers by Sorace (2000, 2011, 2004), who grouped verbs into seven different classes based on their degree of agentivity – besides another semantic property, telicity – and argued that more agentive verbs are more likely to behave as unergatives (see also Acartürk and Zeyrek 2010; Allman 2017; Baker 2019; Huang 2018 for similar, smaller-scale studies). However, a major limitation to Sorace's work is that her classification of verbs into semantic groups was only based on intuition. Testing if – and if so, how strongly – agentivity predicts the syntactic behavior of intransitives requires an actual quantitative measure of agentive semantics rooted in empirical data.

In recent work, Kim et al. (2024) aimed to fill this gap by conducting a systematic comparison of various semantic measures to determine their predictive value for the syntactic behavior of English intransitives. These measures include Sorace's seven verb classes as well as a similar verb classification developed by Levin (1993). Another set of predictors was derived from the GloVe embeddings of the verbs by means of a Principal Component Analysis. In addition, Kim et al. collected semantic ratings for each verb on a scale from 0 to 6 for two sets of features. One consisted of 6 event-related features traditionally considered relevant for the unergative/unaccusative distinction (Agentivity, Telicity, Caused, Transitivity, Dynam-

icity, Requires energy input). The other contained a set of 66 properties developed by Binder et al. (2016) such as Color, Pain, Duration and Angry. Binder and colleagues argue that these 66 experiential features each have a distinct neurological representation and can thus jointly capture how concepts are represented in the brain.

To evaluate the performance of these predictors, Kim et al. collected ratings for 138 verbs on how acceptable they are in reduced relative clauses, an unaccusativity diagnostic (see (1)). Ratings are expected to be higher for more strongly unaccusative verbs. Each verb was presented in the context of three different phrases and rated on a 1–5 Likert scale. The event-related features predicted the syntactic data better than Levin's and Sorace's categorizations, but only Caused emerged as a significant predictor after correcting for multiple comparisons. The model combining experiential and event-related features predicted the syntactic ratings best; Caused and Agentivity were significant predictors among several others. The GloVe-based model came in second. Kim et al. concluded that the unergative/unaccusative distinction is best described as being rooted in graded, embodied features of sensory experience, as formalized in the set of properties developed by Binder et al. (2016). This picture suggests that the effect of agentivity on the syntax of intransitives might have been overestimated.

However, we see a caveat to this finding. Agentivity is a broad and multifaceted concept, having been associated with a wide range of properties like intentionality, animacy, causal power, volition, sentience, and others (Dowty, 1991). The question that Kim et al. asked speakers to determine the degree of agentivity of a verb was: *To what extent does this verb describe something that is actively or intentionally done?* However, this question might have operationalized agentivity too narrowly. In other words, the question is not only whether agentivity matters for the syntactic behavior of intransitives, but also *what kind* of agentivity matters. Moreover, we note that rating tasks are not as straightforwardly applied to verbs as they are to nouns, as it can be difficult to formulate rating questions that target the semantic properties of verbs in a way that is natural for human participants. Here, we aim to address both of these concerns.

1.3 This study

In this study, we develop a novel methodology to set up property axes based on human ratings. We compute axes in embedding space that optimize the fit to human ratings and then project other word embeddings onto them, thus simulating ratings for unseen items. Crucially, our axes achieve a good fit to human ratings even without seed words.

We apply this methodology to the problem of split intransitivity by using a set of human ratings from VanArsdall and Blunt (2022). They collected ratings for 1,200 concrete nouns on eight different dimensions of animacy such as similarity to a person, ability to reproduce and goal-directedness. For each dimension, we computed a property axis that optimizes the fit to VanArsdall and Blunt's noun ratings and projected intransitive verbs onto it to derive a semantic measure of the verbs on this property. Rating-based property axes thus allow us to explore different ways of conceptualizing agentivity, while also circumventing the methodological challenges for collecting human ratings for verbs.

To evaluate the performance of these property axes, we tested how well the semantic measures derived from them predict the syntactic measure of unergativity/unaccusativity used by Kim et al. (2024). We find that a broad concept of animacy, being associated with properties like movement and being alive, is a stronger predictor of the unergative/unaccusativity split than a narrow concept of goal-directedness. This also accounts for the fact that Kim et al.'s measure of agentivity, which they operationalized in terms of intentionality, did not perform particularly strongly in their study in predicting the syntactic data.

Our ambition in the following is not to solve the general question of which semantic properties affect the syntactic behavior of intransitives. Rather, the more modest goal of this study is to focus on one semantic property – agentivity – that has emerged from the previous literature as the most likely contender to affect to syntactic behavior of intransitives and determine what kind of agentivity is the strongest predictor for the unergative/unaccusative distinction. We hope that our findings stimulate further research on the semantic correlates of split intransitivity, comparing agentivity to a wider range of semantic properties.

The major contribution of this study, however, lies in developing a novel methodology for detecting semantic properties in embedding

space. Rating-based property axes constitute a low-resource strategy for high-quality simulation of human ratings, making them a fruitful method for a wide range of research projects.

2 Modeling

2.1 Methods

Many gradable properties seem to be linearly encoded in embedding space, a fact that has been used for analyses in linguistics, cognition, and social sciences (Bolukbasi et al., 2016; Kozlowski et al., 2019; Garí Soler and Apidianaki, 2020; Grand et al., 2022). It has even been proposed that many high-level concepts are encoded linearly (Park et al., 2024), although this does not hold for all concepts (Engels et al., 2025). Traditionally, axes in embedding space that encode gradable properties have been computed using seed words (e.g., Grand et al. 2022; Kozlowski et al. 2019; Garí Soler and Apidianaki 2020; Lucy et al. 2022). However, Erk and Apidianaki (2024); Lucy et al. (2022); Antoniak and Mimno (2021) highlight the limitations of this method, as discussed earlier in Section 1.1.

To improve the accuracy of seed-based axes, Erk and Apidianaki (2024) (below: EA) introduced a method for computing property axes that interpolates seed words with training data in the form of human ratings. To compute an axis $\vec{f}$ for a property f of concepts w, they used a loss function that penalizes the predicted scalar projection of $\vec{w}$ onto $\vec{f}$ for deviation from the gold rating y_w for w. This was combined with a second loss that enforces closeness to a seed-based axis.

The EA model optimizes pointwise fit to human ratings. But the typical use of property axes is to predict *rankings* of concepts along the property. We introduce a new model that, like the EA model, fits a property axis to human ratings, but uses a ranking loss to more directly approximate the characteristic of interest and as a result does not require seeds to fit the data well. We use a margin ranking loss (Nayyeri et al., 2019): for any pair (a, b) of concepts where the gold rating of b is higher than that of a, it encourages the predicted value for b to be higher than a's by at least a margin of d:

$$J_r = \sum_{(a,b)\in P, \hat{y}_b > \hat{y}_a |} max(0, d - y_b + y_a)$$

where P is a set of training item pairs, an N-size set sampled from all training item pairs with at least a difference of d in their gold ratings; for a concept pair (a, b), $\hat{y}_a, \hat{y}_b$ are gold ratings, and y_a, y_b are predictions. N and d are the parameters of the model.

model	POC	XPOC
seed	.629	.631
pointwise	.701	.779
ranking	.703	.797

Table 1: Comparing our ranking-based fitted axes to EA (pointwise) and seed-based axes on the data of Grand et al. (2022). Pointwise uses interpolated losses based on human ratings and seed axes. All values are averaged over folds and conditions.

2.2 Evaluation

We evaluated the ranking loss model on the large collection of concepts and human property ratings introduced by Grand et al. (2022) and compared it to the original seed based axes of Grand et al. (2022) and the EA model. Grand et al. (2022) measured performance through correlation (pearson *r*) as well as *pairwise order consistency* (POC), the percentage of test pairs ordered correctly by the model. However, we cannot use correlation because the data sets become too small once part of the data is used for training. We measured POC as well as *extended* POC (XPOC), the percentage of test pairs and train/test pairs ordered correctly by the model (Erk and Apidianaki, 2024) – this checks whether test items are ranked correctly with respect to training or test items. Results from a 5-fold crossvalidation on the Grand et al. data are shown in Table 1. *Ranking* is our new model. In the evaluation, as throughout in this paper, we used GLoVE embeddings with 300 dimensions pre-trained on Wikipedia and Gigaword. [1]

We see that the ranking loss model, like the EA model, clearly improves fit over the seed-based axes, and achieves even better performance than the EA model. The EA model interpolates human ratings with seeds, and in fact flounders when no seed axis is given. The ranking model, in contrast, manages to fit the data very well even without the help of a seed axis.

To extend our evaluation of the ranking loss model, we evaluated on the VanArsdall and Blunt

[1] Hyperparameters of the EA model are as given in that paper; hyperparameters for our ranking model were optimized on the development portion of the Grand et al. (2022) data defined by Erk and Apidianaki (2024), $d = 0.2 \cdot$ sdev, $N = 300$.

Living	Thought	Reproduction	Person	Goals	Move
waitress	expert	physician	referee	president	twister
waiter	engineer	mother	boyfriend	leader	squirrel
son	human	mom	teenager	governor	waitress
officer	biologist	dog	salesman	physician	ambulance
kitten	astronaut	human	person	doctor	cougar
frog	woman	walrus	officer	inventor	tornado
elephant	surgeon	toad	man	astronaut	bunny
cousin	leader	cheerleader	human	attorney	wasp
citizen	captain	bee	detective	scientist	politician
chipmunk	professor	man	uncle	detective	runner

Table 2: Highest-scoring nouns per animacy dimension in the VanArsdall and Blunt (2022) dataset.

(2022) dataset which we draw on later to predict the syntactic behavior of intransitives. VanArsdall and Blunt collected ratings for 1,200 concrete nouns on 6 different animacy dimensions: general living/non-living scale, ability to think, ability to reproduce, similarity to a person, goal-directedness and movement likelihood. Table 2 summarizes the highest-scoring nouns for each animacy dimension in their study. Using factor analysis, these axes were then further clustered into two coarser dimensions, mental animacy and physical animacy. Note that while VanArsdall and Blunt refer to the semantic properties measured as variants of *animacy*, they are broad enough to also be describable as *agentivity*. In the following, we largely use the two terms interchangeably.

This dataset differs from Grand et al. (2022) in two ways. First, it contains much more data for each individual property. This allows us to compute the correlation between gold ratings and predictions because the test folds of the cross-validation are larger. Second, the VanArsdall and Blunt dataset does not come with seeds. Hence, we compare the ranking loss model to the EA pointwise loss model without seeds. Hyperparameter values for the EA and ranking loss models were not fit anew; we reused the values fit to the Grand et al. data by EA.

Table 3 summarizes the results of a 5-fold cross-validation on the VanArsdall and Blunt (2022) data. We omit XPOC, which patterns with POC, and instead report correlation. We find that the ranking loss model again achieves a solid fit to the data, while the pointwise loss model performs poorly in the absence of seeds.

Tables 4 and 5 show the performance of the ranking loss model and the pointwise loss model for each individual animacy dimensions. Again, the ranking loss model is superior: with the exception of Thought, it scores over .7 on pearson *r* for all dimensions, whereas the pointwise loss model only performs well for Thought, Mental and Physical, with values close to or below 0 in the other categories.

feature	POC	pearson *r*
pointwise	.516	0.05
ranking	.783	.789

Table 3: Crossvalidating the fitted axes on the VanArsdall and Blunt (2022) data, comparing pointwise and ranking loss. Values are averaged over folds and conditions. Both sets of axes are fitted without seeds.

feature	POC	pearson *r*
Living	.791	.788
Thought	.505	.024
Repr.	.789	.800
Person	.779	.801
Goals	.796	.800
Move	.764	.740
Mental	.811	.808
Physical	.794	.808

Table 4: Crossvalidating axes fitted with ranking loss (no seeds) on the VanArsdall and Blunt (2022) data, results for individual dimensions. Values are averaged over folds and conditions.

We further tested the performance of the ranking loss axes by taking the 60,000 most frequent words in COCA, extracting the verbs from that list, and among them, computing the words with the highest values on each property axis for the six fine-grained animacy dimensions. The results are summarized

feature	POC	pearson *r*
Living	.503	.002
Thought	.780	.809
Repr.	.543	.128
Person	.520	.077
Goals	.533	.098
Move	.493	-.016
Mental	.811	.808
Physical	.794	.808

Table 5: Crossvalidating axes fitted with pointwise loss (no seeds) on the VanArsdall and Blunt (2022) data, results for individual dimensions. Values are averaged over folds and conditions.

in Table 6.

For all axes, the highest-scoring words are indeed associated with animacy/agentivity, but we also see that the axes single out different shades of this concept. For example, *infected* is only in the top 10 verbs for the dimensions Living and Reproduction, being associated with biological life rather than human action more specifically. Similarly, *preys* only scores highly for Reproduction. *Graduate* occurs exclusively in the top 10 verbs for Thought, Person and Goals; *interviewed* for Thought and Person.

Lastly, we compute the corelation between the eight animacy axes and Kim et al.'s agentivity ratings, for which subjects were asked how intentionally an action was performed. We find a significant correlation for Goals (.286, *p*-val. .0), Thought (.191, *p*-val. .001), Person (.191, *p*-val. .001), and coarse-grained Mental Animacy (.254, *p*-val. .0), but not for the Physical Animacy features. This is in line with the fact that Kim et al.'s rating task targeted a particular dimension of agentivity, matching most closely VanArsdall and Blunt's category of goal-directedness. Property axes allow us to test a wider spectrum of semantic dimensions with respect to how well they predict the syntactic behavior of intransitives.

3 Predicting the syntactic data

We evaluated how well the eight animacy axes predict the unergative/unaccusative split using the experimental data collected by Kim et al. (2024). Recall that Kim et al. tested the acceptability of 138 intransitive verbs in reduced relative clauses, rated on a 5-point Likert scale. Each verb was shown in the context of three different phrases (e.g., *the frozen ground*, *the frozen lake*, *the frozen popsicles*). Since reduced relatives are an unaccusativity diagnostic, and since we expect animacy/agentivity to be correlated with unergativity, higher animacy scores should correspond to lower acceptability in reduced relatives.

We tested this prediction by fitting mixed-effects Bayesian ordinal regression models with default priors, using the `brms` library in R. All models were computed with a cumulative probit link function and fitted with 2000 iterations (1000 warm-up, 1000 samples taken). R-hat was 1.00 throughout; no divergences were observed during sampling. All gradable features were z-scored. Unlike Kim et al. (2024), we did not average the syntactic ratings over phrases and subjects, which would not allow us to account for inter-subject variation and would generally simplify the data, potentially obscuring important effects. Instead, all our models included by-subject intercepts. The formula for the models is given in (3), with different models differing in

Living	Thought	Reproduction	Person	Goals	Move
infected	acquitted	infected	soldier	killed	killed
emigrated	abducted	preys	acquitted	abducted	soldier
acquitted	killed	emigrated	interviewed	graduate	trained
kill	soldier	kill	graduate	trained	kill
tormented	marshal	acquitted	suspect	soldier	encounters
rat	trained	killed	committed	committed	marshal
killed	interviewed	rat	killed	acquitted	kills
kills	freed	abducted	trained	killing	brave
marshal	graduate	typecast	abducted	confirmed	hates
trained	committed	kills	identified	identified	emigrates

Table 6: Verbs with the highest values on the six fine-grained animacy dimensions fitted to the VanArsdall and Blunt (2022) dataset with ranking loss. Items are extracted from the list of the 60,000 most frequent words in COCA.

the value of animacy_score.

(3) answer ∼ animacy_score + (1 | subject)

Table 7 summarizes the fixed effect regression coefficients for each of the eight regression models. As expected, higher animacy scores across all dimensions result in lower acceptability of reduced relative clauses, an unaccusativity diagnostic. All 95% confidence intervals exclude 0.

predictor	est.	est. error	l-95%	u-95%
Living	-0.24	0.01	-0.26	-0.21
Thought	-0.23	0.01	-0.25	-0.20
Repr.	-0.20	0.01	-0.22	-0.18
Person	-0.22	0.01	-0.24	-0.20
Goals	-0.14	0.01	-0.16	-0.12
Move	-0.35	0.01	-0.37	-0.32
Mental	-0.20	0.01	-0.23	-0.18
Physical	-0.25	0.01	-0.27	-0.22

Table 7: Estimates for fixed-effect coefficients with estimate error and upper and lower bound 95% confidence interval.

We then compared the goodness of fit of the different models with a leave-one-out (LOO) analysis using the `loo` library in R. LOO provides a measure of predictive accuracy by training a model on all data points except one and then testing how well the model predicts the held-out data point. Each model is compared to a null model that only includes by-subject intercepts (4):

(4) answer ∼ (1 | subject)

This allows us to evaluate how well the two semantic measures predict the syntactic data relative to each other.

Table 8 summarizes the result of the LOO analysis in terms of expected log predictive density (ELPD). ELPD is a measure of the log probability that the model attributes to all the held-out data points; a higher value (closer to zero) corresponds to better predictive performance.

For the six fine-grained axes, we see that the strongest improvement is achieved by Move (444.6), followed by Living (214.4). Goals occupies the very bottom of the list (74.3). This indicates that a broader notion of animacy associated with properties like movement and being alive is a better predictor for the syntactic behavior of intransitives than a narrow notion of goal-directedness or intentionality.

models	ELPD diff	SD
Move vs. null model	444.6	28.5
Living vs. null model	214.4	2.6
Thought vs. null model	195.5	19.2
Person vs. null model	184.3	18.8
Reproduction vs. null model	149.2	17.2
Goals vs. null model	74.3	12.1
Mental vs. null model	156.3	17.4
Physical vs. null model	23.5	21.2

Table 8: LOO analysis with different animacy axes.

However, for the coarse-grained axes, Physical Animacy performs better than Mental Animacy. At first, this finding seems to point in the opposite direction than the results from the fine-grained axes. We argue that this apparent contradiction is related to some inherent limitations to the Physical dimension itself. Recall that VanArsdall and Blunt (2022) computed the two coarse-grained animacy axes via factor analysis. While taken together, these two axes explain a total of 86.07% of the variance in the data, 79.05% were accounted for by Mental, and only 7.02% by Physical. Moreover, the factor loading of Move – the strongest predictor among the narrow dimensions – on Physical is fairly low at .56, contrasting with .96 for Living and .93 for Reproduction. Overall, Physical does not predict much of the variance in the dataset in general, and it is only moderately related to the dimension of Move. We conclude that the results from the fine-grained animacy dimensions are more informative for our purposes.

4 Discussion

This study has developed a novel method for computing property axes in embedding space based on human ratings using a ranking loss function. We have evaluated the ranking loss axes on two large-scale datasets, Grand et al. (2022) and VanArsdall and Blunt (2022), in both cases finding a solid fit to the data. For both datasets, the ranking loss axes clearly outperformed rating-based property axes computed with pointwise loss (Erk and Apidianaki, 2024), which required seed words to match human ratings well.

We have then applied this methodology to the problem of determining the semantic correlates of the unergative/unaccusative contrast. We have focused on the concept of agentivity/animacy as the most likely contender for a semantic property that

affects split intransitivity, and investigated which dimension of agentivity/animacy best predicts a measure of the syntactic behavior of intransitive verbs, namely, their acceptability in reduced relative clauses. We found that a lower-level concept of animacy, associated with properties such as movement and being alive, emerged as a stronger predictor than a narrow notion of goal-directedness. The vastly different performance of the different property axes highlights the importance of carefully operationalizing agentivity/animacy by taking into account the various semantic dimensions associated with it, rather than equating it with a single, more specific concept like intentionality.

Our findings for split intransitivity come with a number of limitations. First, in terms of the syntactic data used here, we have tested the performance of our property axes for a single unaccusativity diagnostic, reduced relative clauses. Adding further unaccusativity diagnostics such as resultative secondary predicates or agent nominalizations would give us a more complete picture.

Second, this study has represented each verb with a single lemma, its unmarked form. However, it would be more accurate encode the verb by abstracting over its different morphological forms (e.g., *kill*, *kills*, *killed* and *killing*). In future work, we aim to represent verbs as sets or clusters of embeddings, each encoding a different morphological form, rather than as a single embedding. This is especially imperative for morphologically richer languages in which verbs can take on dozens of different forms.

Third, we have focused here exclusively on the concept of agentivity/animacy. However, the syntactic behavior of unergatives might be equally, or perhaps even more strongly, affected by other semantic properties, with especially telicity being another promising candidate (e.g., Sorace 2000, 2011, 2004). We hope that by exploring the different dimensions of agentivity, this paper has established a more solid ground for future studies to further compare agentivity to other properties. This might also take the form of exploring the embedding space more directly: rather than identifying an axis for a particular property that is hypothesized to predict the syntactic data and then test its performance, we might also identify an axis that performs well in predicting the syntactic data and then attempt to determine the semantic property associated with it.

Fourth, a confound comes from the fact that many verbs in our sample of intransitives – e.g., *break* – can also be used as transitives, as in *The glass broke/I broke the glass*. For predicting the syntactic behavior of the intransitive, we are interested in the degree of animacy associated with *the glass*. However, projecting the embedding for *break* onto an animacy axis does not allow us to specifically target properties associated with one argument rather than another, thus potentially distorting the semantic measures derived from it. A related issue is that some English verbs double as nouns: in Table 6, which reported the highest-scoring verbs for each animacy dimensions, several items arguably made it to the top of the list not because of their verbal, but because of their nominal use (e.g., *soldier*, *marshal*, *rat*). In response to this predicament, for the future we plan to extend our work to token-level embeddings, which encode verb meaning in context. This would allow us to specifically extract intransitive verbs and test their behavior with respect to property axes.

Lastly, another direction for future research lies in broadening the scope of our work beyond English. Investigating to what extent split intransitivity is sensitive to the same semantic properties cross-linguistically would also tie in with research on the cognitive status of thematic roles such as Agent and Patient (see Rissman and Majid 2019 for an overview), which the unergative/unaccusative split has been argued to be based on.

Beyond split intransitivity, we believe that rating-based property axes hold much promise for research in semantics and at the syntax-semantics interface. Compared to seed-based axes, they do not require selecting a set of seed words manually, which is prone to various biases and confounds, but allow us to directly recover axes in embedding space that map onto human judgments. Rating-based axes thus constitute a low-resource strategy for high-quality simulation of human ratings. A methodological question to address in the future is how much data in the form of human ratings is needed to compute axes that perform well. Moreover, the axes we have used here were both trained and cross-validated on nouns, but we then extrapolated ratings from them to verbs. Further research should address whether the application of property axes across categories poses any special challenges. Overall, we hope that our work contributes to a better understanding of how semantic properties are encoded in embedding space.

References

Cengiz Acartürk and Deniz Zeyrek. 2010. Unaccusative/unergative distinction in Turkish: A connectionist approach. In *Proceedings of the 8th Workshop on Asian Language Resources*, pages 111–119, Beijing, China.

JungAe Lee Allman. 2017. *Empirical examination of two diagnostics of Korean unaccusativity*. Ph.D. thesis, The University of Texas at Arlington, Arlington, TX.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online. Association for Computational Linguistics.

James Baker. 2019. Split intransitivity in English. *English Language and Linguistics*, 23:557–589.

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. 2016. Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33:130–174.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in neural information processing systems*, 29.

Luigi Burzio. 1981. *Intranitive Verbs and Italian Auxiliaries*. Ph.D. thesis, MIT, Cambridge, MA.

Luigi Burzio. 1986. *Italian Syntax: A Government and Binding Approach*. D. Reidel, Dordrecht.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67:547–619.

Joshua Engels, Isaac Liao, Eric J Michaud, Wes Gurnee, and Max Tegmark. 2025. Not all language model features are linear. In *Proceedings of ICLR*.

Katrin Erk and Marianna Apidianaki. 2024. Adjusting interpretable dimensions in embedding space with human judgments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2675–2686, Mexico City, Mexico. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705.

Aina Garí Soler and Marianna Apidianaki. 2020. BERT knows Punta Cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualised representations. In *Proceedings of EMNLP*, pages 7371–7385.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, , and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, 6:975–987.

Yujing Huang. 2018. *Linking form to meaning: Reevaluating the evidence for the unaccusative hypothesis*. Ph.D. thesis, Harvard University, Cambridge, MA.

Ray Jackendoff. 1983. *Semantics and Cognition*. MIT Press, Cambridge, MA.

Songhee Kim, Jeffrey R Binder, Colin Humphries, and Lisa L Conant. 2024. Decomposing unaccusativity: a statistical modelling approach. *Language, Cognition and Neuroscience*, 39:1189–1211.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84:905–949.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Beth Levin and Malka Rappaport Hovav. 1995. *Unaccusativity. At the Syntax-Lexical Semantics Interface*. MIT Press, Cambridge, MA.

Li Lucy, Divya Tadimeti, and David Bamman. 2022. Discovering differences in the representation of people using contextualized semantic axes. In *Proceedings of EMNLP*, pages 3477–3494.

Mojtaba Nayyeri, Xiaotian Zhou, Sahar Vahdati, Hamed Shariat Yazdi, and Jens Lehmann. 2019. Adaptive margin ranking loss for knowledge graph embeddings via a correntropy objective function.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of ICML*, pages 39643–39666.

David Perlmutter. 1978. Impersonal passives and the Unaccusative Hypothesis. *Papers from the Annual Meeting of the Berkeley Linguistic Society*, 4:157–189.

Lilia Rissman and Asifa Majid. 2019. Thematic roles: Core knowledge or linguistic construct? *Psychonomic Bulletin amp; Review*, 26:1850–1869.

Antonella Sorace. 2000. Gradients in auxiliary selection with intransitive verbs. *Language*, 76(4):859–890.

Antonella Sorace. 2004. Gradience at the lexicon-syntax interface: Evidence from auxiliary selection and implications for unaccusativity. In Artemis Alexiadou, Elena Anagnostopoulou, and Martin Everaert, editors, *The Unaccusativity Puzzle*, pages 243–268. Oxford UP, Oxford.

Antonella Sorace. 2011. Gradience in split intransitivity: the end of the Unaccusative Hypothesis? *Archivio Glottologico Italiano*, XCVI(1):67–86.

Robert D Van Valin Jr. 1990. Semantic parameters of split intransitivity. *Language*, 66:221–260.

Joshua E VanArsdall and Janell R Blunt. 2022. Analyzing the structure of animacy: Exploring relationships among six new animacy and 15 existing normative dimensions for 1,200 concrete nouns. *Memory & Cognition*, 50:997–1012.