Towards a Better Rational Speech Act Framework for Context-Aware Modeling of Metaphor Understanding

Gaia Carenini [1] **Luca Bischetti** [2] **Walter Schaeken** [3] **Valentina Bambini** [2]

Abstract

Modeling language is a fundamental step for understanding human communication and improving human-computer interaction. The Rational Speech Act (RSA) model provides a flexible framework to pursue this objective by catching pragmatic reasoning. However, state-of-the-art models still have limitations in dealing with context. We present a new RSA framework for metaphor understanding that accounts explicitly for the role of context by emphasizing the mutual shared information between the speaker and the listener in the estimation of the communicative goal. The model is tested extensively against 24 metaphors (with either intrinsic or emergent properties) and its predictions are compared to human data.

1. Introduction

The development of formal models of natural language is an emerging promising field of research. The introduction of the Rational Speech Act (RSA) framework (Frank & Goodman, 2012) provides a flexible and intuitive way to model - in a probabilistic way (Goodman & Frank, 2016) - pragmatic reasoning and cooperative communication between interlocutors, thus formalizing the general principles of verbal interactions (Grice, 1989; Sperber & Wilson, 1995). RSA allows us to explore many pragmatic phenomena, e.g., scalar implicatures and irony (Goodman & Stuhlmüller, 2013; Kao & Goodman, 2015), including metaphor (Kao et al., 2014; Mayn & Demberg, 2022). In particular, Kao and colleagues (2014) developed an RSA model in which the speaker and listener recursively reason about each other's knowledge to reach the intended meaning. Mayn & Demberg (2022) further introduced a graded approach accounting for the typicality of metaphor vehicles. As a whole, these models capture relevant facets of metaphor understanding with different degrees of success (rs between models' outputs and human ratings .70 and .54, respectively), yet they present several weaknesses. First, they work with nominal predicative metaphors ("X is Y"), with X (topic, that is, the subject of the metaphor) being male proper names and Y (vehicle, that is, the term used metaphorically to predicate about the topic) being animals (e.g., *John is a fox*), whereas metaphor vehicles might also include non-animate entities referring to a broader set of topics (Carston, 2016). Secondly, they compared models' performance against human judgements, and not interpretations of metaphors. Thirdly, RSA models on metaphor often do not account for the role of context, whereas metaphor interpretations are inferentially derived using contextual cues (Wilson & Carston, 2007). Fourthly, previous RSA models did not consider metaphors predicating emergent properties, that is, features that are not standardly associated with individual constituents in isolation (Wilson & Carston, 2007). Lastly, on the technical side, the available models on metaphor present several parameters obtained through data interpolation (e.g., the parameter about the expected communicative goal).

In this work, we address these weaknesses by presenting a new model that: a) can be applied to nominal predicative metaphors of the form "X is Y", where X is a human category and Y is an animal or an object category, also with emergent properties (e.g., *Philosophers are airplanes*), and b) explicitly accounts for a (minimal) communicative context - consisting of the metaphor's topic and the background knowledge associated with it - to determine the parameter accounting for the communicative goal. We test extensively our model on a dataset of 24 metaphors.

2. Computational Model

In the Rational Speech Act model (Frank & Goodman, 2012), a listener and speaker engage in recursive reasoning to arrive at enriched meanings. The speaker selects an utterance based on its informativeness, considering a literal

[1]Department of Computer Science, ENS-PSL Research University, Paris, France [2]Laboratory of Neurolinguistics and Experimental Pragmatics (NEP), Department of Humanities and Life Sciences, University School for Advanced Studies IUSS, Pavia, Italy [3]Laboratory for Experimental Psychology, KU Leuven, Leuven, Belgium. Correspondence to: Gaia Carenini <gaia.careniniens.psl.eu>.

Proceedings of the First Workshop on Theory of Mind in Communicating Agents, Honolulu, Hawaii, USA. PMLR 202, 2023.

listener's perspective. The pragmatic listener then infers the meaning given the utterance using Bayes' rule, considering the speaker's reasoning. To accommodate non-literal interpretation, the model extends by incorporating various communicative goals (Kao et al., 2014). Introducing multiple goals enables the speaker to produce an utterance that satisfies their goal, even if it is literally false. The optimal utterance is informative and relevant in fulfilling the speaker's communicative goal. Since the listener may not know the precise goal, a joint inference that considers both the goal and intended meaning must be performed. This is achieved by leveraging the speaker and listener's shared knowledge, or common ground, to reason about the information gained from a literal interpretation.

In our model, we consider nominal predicative metaphors of the form "X is Y", where X is a human category and Y is an animal or an object category. We restrict the possible features of X under consideration to a vector of size n : $f = [f_1, \ldots, f_n]$, where f_i is a number in the interval $[0, 1]$. The literal listener L_0 will interpret the utterance "X is Y" as meaning that X is literally a member of the category Y and has corresponding features. Formally, if u is the uttered category:

$$L_0(c, f|u) = \begin{cases} P(f|c), & \text{if } c = u \\ 0 & \text{otherwise} \end{cases}$$

where $P(f|c)$ is the prior probability that a member of category c has feature vector f.

Our assumption is that the speaker aims to convey the significance of a specific feature. Consequently, the speaker's goal can be seen as a mapping from the complete feature space to the subset that holds relevance for them. Formally, the goal to communicate about feature $i \in \{1, \ldots, n\}$ is the function $g_i(f) = f_i$. Adhering to the Rational Speech Act model, the speaker's utility is defined as the negative surprisal of the true state, given an utterance, under the listener's distribution. In this case, our attention is solely on the surprisal along the goal dimension. To achieve this, we project along the goal dimension as in (Kao et al., 2014), leading to the following utility function for the pragmatic speaker S_1:

$$U(u|g, f) = \log \sum_{c,f'} \delta_{g(f)=g(f')} L_0(c, f'|u) \qquad (1)$$

Based on this utility function, the pragmatic speaker employs a softmax decision rule, which approximates the behavior of a rational planner, to select an utterance:

$$S_1(u|g, f) \sim e^{\lambda U(u|g,f)}$$

where λ is an optimality parameter.

The listener L_1 uses Bayesian inference to infer the intended meaning based on prior knowledge and their understanding of the speaker. To determine the speaker's intended meaning, L_1 considers all possible speaker goals and integrates over them. In particular, the pragmatic listener L_1 is defined as:

$$L_1(c, f|u) \sim P(c)P(f|c) \sum_g \mathcal{R}(g|t) S_1(u|g, f)$$

where c, f, u, and t denote respectively a category, a feature vector, an utterance, and the topic of the metaphor under discussion and $\mathcal{R}$ is a function expressing the relevance of the goal g to the topic t.

Other parameters involved in the modeling are: a) the prior probability that the entity discussed belongs to category c, $P(c)$, estimated by fitting the model to the data; b) the prior probability that a member of category c has feature vector f, $P(f|c)$, estimated in Experiment 2; and c) the parameter accounting for minimal context, $\mathcal{R}(g|t)$, estimated from Experiment 2 (and not implicit, as in previous studies). This results in a direct dependence of the communicative goal estimate on the topic, and such relationship is expressed through the typicality of features obtained also for the selected topic from Experiment 1. Moreover, we accounted for the metaphor typicality by encoding the feature vector f with real values, instead of integers (see Mayn & Demberg, 2022). This contributes to get a more fine-grained quantification.

The key difference with previous models lies precisely in the introduction of the term $\mathcal{R}(g|t)$, which biases the communicative goal estimate according to the background knowledge on the topic.

3. Behavioral Experiments

We conducted 3 experiments, mostly following Kao et al. (2014), on native speakers of Italian. For all the experiments, we started from a set of 24 metaphors, extracted from the set of metaphors used in (Canal et al., 2022). This set included metaphors with intrinsic properties (i.e., attested in norms of semantic features of vehicles, based on (McRae et al., 2005)) or emergent (i.e., absent in norms of semantic features of vehicles). Experiment 1 was carried out to elicit topics and vehicles' features, accounting for the minimal context; Experiment 2 to compute the conditional probabilities for the categories of interest; and Experiment 3 to measure people's interpretations of metaphors.

Experiment 1. *Participants:* 58 adults (Age, mean=24.98; M=36%, F=59%, other=5%). *Materials:* 48 nouns (topics and the vehicles of the 24 metaphors, see Table 1, Appendix A). *Methods:* Elicitation of 3-to-5 relevant features for each of the proposed nouns. Extraction of features present in at least 10% of answers, after aggregating synonyms. *Results:* From 4615 observations, 59 features for topics and vehicles were extracted (see Table 2, Appendix A).

Experiment 2. *Participants:* 37 adults (Age, mean=25.43; M=24%, F=76%). *Materials:* 48 nouns (topics and the vehicles of the 24 metaphors, see Table 1, Appendix A) and the list of 59 features from Experiment 1. *Method:* Evaluation of typicality (1-7 Likert scale) of the 48 nouns along the 59 features from Experiment 1. *Results:* Value of the estimators for $P(f|c)$ and $\mathcal{R}(g|t)$.

Experiment 3. *Participants:* 40 adults (Age, mean=27.08; M=23%, F=77%). *Materials:* 24 metaphors (see Table 1). *Methods:* Single-forced-choice metaphor interpretation task, asking participants to select one among the 59 features presented. *Results:* Percentages of the most probable interpretation for each metaphor.

4. Model Evaluation

The parameters of the models, namely the multiplicative coefficient γ, an estimator for the likelihood of the literal interpretation (therefore proportional to $P(c)$), and the optimality parameter, were set at fixed values ($\gamma = 0.1, \lambda = 0.5$). This choice is dictated by a data-based estimate of the probability of the literal interpretation of the metaphor and the quantification of the speech actor's rationality.

To potentially improve the numerical results, an unconstrained optimization using the BFGS method was performed, which allowed for obtaining an optimal interpolation of the data points. It should be noted that this procedure may create values that do not generalize well, that is, not suited for "X is Y" metaphors not included in our dataset. Nevertheless, our improvements to the model are able to mitigate possible bias and can be applied to other sets of predicative nominal metaphors.

By employing these parameters, our model effectively produced a distribution of potential interpretations that exhibits a significant association with the one derived from the behavioral experiments. We evaluated our model with two global indicators of (dis)similarity between human data and model-based distributions, namely Pearson's correlation coefficient and Jensen-Shannon Divergence (JSD). The mean Pearson correlation coefficient was determined to be 0.45, with a standard deviation of 0.16. This coefficient indicates a moderate positive correlation between the distributions generated by the model and the interpretations obtained from the behavioral experiments. Moreover, the standard deviation highlights notable variation in the correlation across different interpretations, which we will further discuss in relation to the specific type of metaphor considered. Furthermore, the average JSD was computed as 0.36, with a standard deviation of 0.07. The JSD serves as a measure of dissimilarity between the model's distribution and the observed interpretations. In our case, this index suggests a strong similarity between them (see Figure 1). We then

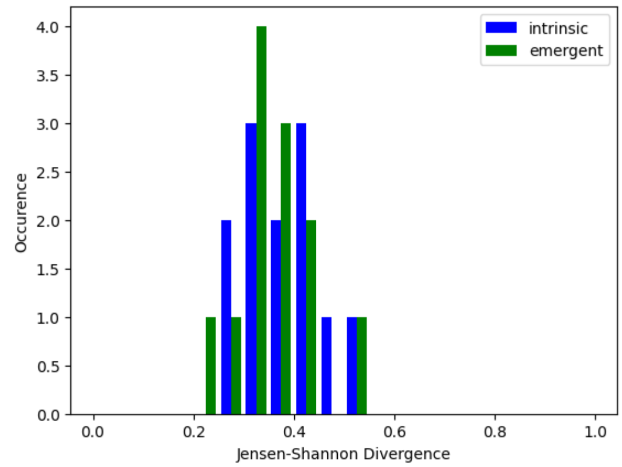

Figure 1. The gradient of information, in terms of Jensen Shannon divergence among the distribution generated by our RSA model and the one associated to participants answers in Experiment 3.

also considered metaphorical interpretations, namely the most probable interpretations as provided by in Experiment 3, against model's predictions. We examined the presence of shared features among the top k likely interpretations, as this may mirror ease of interpretation. Specifically, we analyzed the number of common features for increasing values of k (see Figure 2, 3) . When $k = 3$, only three metaphors exhibit no shared interpretations between the two distributions, while for as many as 9 metaphors, at least two out of the top three interpretations are shared between the distributions.

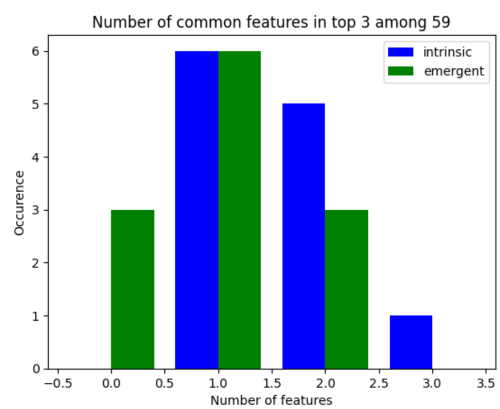

Figure 2. The number of common features among the 3 most likely interpretation according to the distribution generated by our RSA model and the one associated to participants answers in Experiment 3.

The 3 most likely interpretations provided by the model and obtained in the experiments are summarized in Table 3

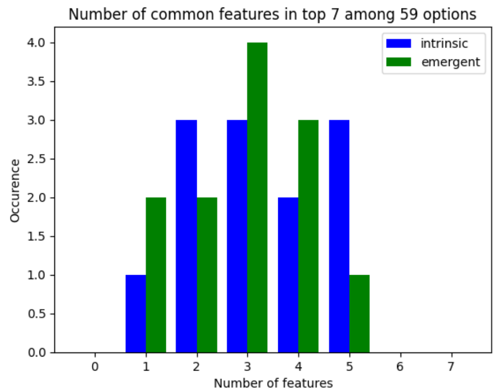

Figure 3. The number of common features among the 7 most likely interpretation according to the distribution generated by our RSA model and the one associated to participants answers in Experiment 3.

(Appendix A).

Metaphors with Emergent vs Intrinsic Properties The model showed different levels of performance in the case of metaphor with emergent and intrinsic properties. In particular, for metaphors with intrinsic properties, the distribution generated by the model always provides at least one valid interpretation (among the 3 most likely). This is not the case for metaphors with emergent properties: to obtain at least one common interpretation between human and model interpretations, the 7 most likely interpretations were to be considered (see Figure 3). In other words, metaphors with intrinsic properties, compared to those with emergent properties, seem to be more sensitive to topic properties, as reflected by the better performance and the more fine-tuned interpretative process.

Comparability with Kao et al. (2014) & Mayn & Demberg (2022) Despite the apparent similarity with our model, the models presented by (Kao & Goodman, 2015) and (Mayn & Demberg, 2022) can be considered a subset of our approach. As such, a direct comparison does not apply. The lack of a parameter for typicality, as well as the constant role of $\mathcal{R}(g|t)$ based on interpolation, point to a reduced generalization performance on different metaphors. More specifically, interpolation inherently leads to models that become closely tailored to the dataset considered. In the context of our work, the primary objective was to explicitly express the parameters rather than relying solely on interpolation. By explicitly defining the parameters, we were able to capture the nuances and intricacies of the metaphor interpretation more accurately. Such a novel framework is possibly able to model metaphor beyond the training-testing set, thus being more ecologically valid or opening to multilingual description of metaphor understanding in RSA terms. To clarify the potential of our model, ablation studies might be useful and recommended for future studies

5. Discussion

Our model seems to be able to capture the implicated meaning of a metaphor with high accuracy – with a gradient of precision, as found for metaphors with emergent properties – overall suggesting that the inclusion of context, defined as the topic and its associated knowledge, is key to understand metaphorical meanings. Compared to previous RSA models of metaphor, where only the vehicle explicitly drives the interpretation, our model innovately considers the role of speaker's intentions and speaker's rationality, and also ”speaks” about the topic and its characteristics, being able to interpret a wider range of metaphoric expressions, including the ones with emergent properties. In doing so, our model has a two-fold value. First, it provides a more explicit formulation of the inferential machinery supporting metaphor understanding – closer to theoretical proposals (Wilson & Carston, 2007) and known mechanisms of human interpretation of metaphors (see (Bambini et al., 2016)). Second, it provides an explicit framework of analysis that can be used as a tool for NLP and potentially as a mean to better understand “blackbox” Large Language Models (Hu et al., 2022) and how they deal with complex communicative tasks such as metaphor understanding.

References

Bambini, V., Bertini, C., Schaeken, W., Stella, A., and Di Russo, F. Disentangling metaphor from context: an ERP study. *Frontiers in Psychology*, 7:559, 2016. doi: 10.3389/fpsyg.2016.00559.

Canal, P., Bischetti, L., Bertini, C., Ricci, I., Lecce, S., and Bambini, V. N400 differences between physical and mental metaphors: The role of Theories of Mind. *Brain and Cognition*, 161:105879, 2022. doi: 10.1016/j.bandc.2022.105879.

Carston, R. Relevance theory and metaphor. In Semino, E. and Demjén, Z. (eds.), *The Routledge Handbook of Metaphor and Language*, pp. 42–55. Routledge, 2016. doi: 10.4324/9781315672953.ch3.

Frank, M. C. and Goodman, N. D. Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084):998–998, 2012. doi: 10.1126/science.1218633.

Goodman, N. D. and Frank, M. C. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, 20(11):818–829, 2016. doi: 10.1016/j.tics.2016.08.005.

Goodman, N. D. and Stuhlmüller, A. Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, 5(1):173–184, 2013. ISSN 17568757. doi: 10.1111/tops.12007.

Grice, P. *Studies in the Way of Words*. Harvard University Press, 1989.

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., and Gibson, E. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv*, 2022. doi: 10.48550/arXiv.2212.06801.

Kao, J. T. and Goodman, N. D. Let's talk (ironically) about the weather: Modeling verbal irony. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015*, pp. 1051–1056, 2015.

Kao, J. T., Bergen, L., and Goodman, N. D. Formalizing the Pragmatics of Metaphor Understanding. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society, CogSci 2014*, pp. 719–724, 2014.

Mayn, A. and Demberg, V. Pragmatics of Metaphor Revisited: Formalizing the Role of Typicality and Alternative Utterances in Metaphor Understanding. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society: Cognitive Diversity, CogSci 2022*, pp. 3154–3160, 2022.

McRae, K., Cree, G. S., Seidenberg, M. S., and Mcnorgan, C. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559, 2005. doi: 10.3758/BF03192726.

Sperber, D. and Wilson, D. *Relevance: Communication and Cognition (2nd edition)*. Blackwell, Oxford, 1995.

Wilson, D. and Carston, R. Metaphor and the 'Emergent Property' Problem: A Relevance-Theoretic Approach. *Baltic International Yearbook of Cognition, Logic and Communication*, 3(1):1–40, 2007. doi: 10.4148/biyclc.v3i0.23.

A. Supplementary Materials

Table 1. The table includes the set of metaphors used in the study, in Italian, with the English translation in italics (first column). For each metaphor, we isolated the topic and the vehicle (here provided in English only, second and third columns). The fourth column described the type of metaphor, namely with intrinsic or emergent properties.

Metaphor	Topic	Vehicle	Properties
I ballerini sono cigni. *Dancers are swans.*	Dancers	Swans	Intrinsic
Gli anziani sono lumache. *The elderly are snails.*	Elderly	Snails	Intrinsic
I ciclisti sono razzi. *Cyclists are rockets.*	Cyclists	Rockets	Intrinsic
I muratori sono rocce. *Masons are rocks.*	Masons	Rocks	Intrinsic
I corridori sono lepri. *Runners are hares.*	Runners	Hares	Intrinsic
I rugbisti sono tori. *Rugby players are bulls.*	Rugby players	Bulls	Intrinsic
I cantanti sono usignoli. *Singers are nightingales.*	Singers	Nightingales	Intrinsic
I papà sono ombrelli. *Dads are ombrellas.*	Dads	Ombrellas	Intrinsic
I genitori sono scudi. *Parents are shields.*	Parents	Shields	Intrinsic
I giocatori sono elefanti. *Players are elephants*.	Players	Elephants	Intrinsic
Le indossatrici sono bambole. *Models are dolls*.	Models	Dolls	Intrinsic
Gli scalatori sono scoiattoli. *Climbers are squirrels*.	Climbers	Squirrels	Intrinsic
I credenti sono greggi. *Believers are flocks*.	Believers	Flocks	Emergent
I buttafuori sono armadi. *Bouncers are closets*.	Bouncers	Closets	Emergent
I fanciulli sono agnelli. *Children are lambs.*	Children	Lambs	Emergent
I capi ufficio sono iene. *Office managers are hyenas*.	Office managers	Hyenas	Emergent
I giornalisti sono avvoltoi. *Journalists are vultures*.	Journalists	Vultures	Emergent
I maestri sono libri. *Teachers are books*.	Teachers	Books	Emergent
Le mogli sono martelli. *Wives are hammers*.	Wives	Hammers	Emergent
I filosofi sono aeroplani. *Philosophers are airplanes*.	Philosophers	Airplanes	Emergent
Le nuore sono trapani. *Daughters-in-law are drills*.	Daughters-in-law	Drills	Emergent
Gli operai sono formiche. *Workers are ants*.	Workers	Ants	Emergent
I cuochi sono mongolfiere. *Cooks are hot-air balloons*.	Cooks	Air balloons	Emergent
Gli impiegati sono zerbini. *Office workers are doormats*.	Office workers	Doormats	Emergent

Table 2. List of the features obtained in Experiment 1.

English Term	Translation
Aggressività	Aggressiveness
Agilità	Agility
Altezza	Height
Amore	Love
Armonia	Harmony
Arte	Art
Atleticità	Athleticism
Autoritarietà	Authoritarianism
Bellezza	Beauty
Candore	Candor
Capienza	Capacity
Colore	Color
Competenza	Competence
Competitività	Competitiveness
Coraggio	Courage
Creatività	Creativity
Curiosità	Curiosity
Devozione	Devotion
Disponibilità	Availability
Dolcezza	Sweetness
Durezza	Hardness
Eleganza	Elegance
Fedeltà	Fidelity
Forza	Strength
Fragilità	Fragility
Gioventù	Youth
Grandezza	Size
Innocenza	Innocence
Intelligenza	Intelligence
Interesse	Interest
Invadenza	Intrusiveness
Laboriosità	Diligence
Leggerezza	Lightness
Lentezza	Slowness
Magrezza	Thinness
Metallicità	Metallic
Musicalità	Musicality
Numerosità	Numerosity
Perforanza	Penetrating power
Pericolosità	Danger
Pesantezza	Heaviness
Piccolezza	Smallness
Potenza	Power
Preoccupazione	Concern
Protezione	Protection
Resistenza	Resistance
Robustezza	Robustness
Rumorosità	Noisiness
Saggezza	Wisdom
Sporcizia	Dirtiness
Tenerezza	Tenderness
Utilità	Utility
Velocità	Speed
Viscidità	Stickiness
Volo	Flight
Voracità	Voracity
Opportunismo	Opportunism
Sottomissione	Submissiveness
Distrazione	Distraction

Table 3. The table includes the set of metaphors used in the experiment, in the English translation (first column). For each metaphor, we include the three most probable interpretations provided by humans (second column) and by the RSA model (third column).

Metaphor	Human interpretation	RSA model interpretation
Dancers are swans.	Elegance, Lightness, Beauty	Beauty, Elegance, Harmony
The elderly are snails.	Slowness, Tenderness, Fragility	Slowness, Stickiness, Smallness
Cyclists are rockets.	Speed, Athleticism, Opportunism	Speed, Noisiness, Flight
Masons are rocks.	Strength, Robustness, Hardness	Hardness, Robustness, Heaviness
Runners are hares.	Speed, Agility, Athleticism	Agility, Speed, Athleticism
Rugby players are bulls.	Power, Robustness, Strength	Strength, Aggressiveness, Power
Singers are nightingales.	Musicality, Harmony, Sweetness	Lightness, Smallness, Musicality
Dads are umbrellas.	Protection, Love, Concern	Utility, Protection, Color
Parents are shields.	Protection, Resistance, Robustness	Protection, Robustness, Utility
Players are elephants.	Heaviness, Robustness, Size	Size, Heaviness, Height
Models are dolls.	Beauty, Elegance, Submissiveness	Beauty, Youth, Tenderness
Climbers are squirrels.	Agility, Athleticism, Harmony	Speed, Agility, Smallness
Believers are flocks.	Submissiveness, Devotion, Numerosity	Numerosity, Innocence, Submissiveness
Bouncers are closets.	Robustness, Height, Size	Utility, Capacity, Height
Children are lambs.	Innocence, Tenderness, Purity	Innocence, Fragility, Tenderness
Office managers are hyenas.	Aggressiveness, Authoritarianism, Opportunism	Danger, Aggressiveness, Agility
Journalists are vultures.	Opportunism, Intrusiveness, Competitiveness	Flight, Aggressiveness, Danger
Teachers are books.	Wisdom, Competence, Intelligence	Wisdom, Art, Interest
Wives are hammers.	Heaviness, Intrusiveness, Loudness	Utility, Heaviness, Danger
Philosophers are airplanes.	Creativity, Flight, Wisdom	Flight, Height, Utility
Daughters-in-law are drills.	Intrusiveness, Heaviness, Penetrating power	Noisiness, Power, Utility
Workers are ants.	Diligence, Numerosity, Slowness	Diligence, Smallness, Numerosity
Cooks are hot-air balloons.	Creativity, Size, Robustness	Flight, Height, Size
Office workers are doormats.	Submissiveness, Opportunism, Smallness	Utility, Dirtiness, Color