A Frame-Theoretic Model of Bayesian Category Learning

Samuel D. Taylor and Peter R. Sutton

Abstract Bayesian models of category learning typically assume that the most probable categories are those that group input stimuli together around a maximally optimal number of shared features. One potential weakness of such feature list approaches, however, is that it is unclear how to weight observed features to be more or less diagnostic for a given category. In this theoretically oriented paper, we develop a frame-theoretic model of Bayesian category learning that weights the diagnosticity of observed attribute values in terms of their position within the structure of a frame (formalised as distance from the frame's central node). We argue that there are good grounds to further develop and empirically test frame-based learning models, because they have theoretical advantages over unweighted feature list models, and because frame structures provide a principled means of assigning weights to attribute values without appealing to supervised training data.

Keywords Category learning · Bayesian categorisation · Frames · Weighted Naïve Bayesian model · Frame-theoretic constraints

1 Introduction

Bayesian models of categorisation typically assume that there is both an input to categorisation—the stimulus to be categorised—and an output from categorisation—the (cognitive) behaviour of the categoriser (Kruschke 2008). But in order to count as cognitively adequate, the model must also represent the cognitive processes that mediate between input and output, and take these *representations* to be informative

S. D. Taylor (✉)

Institute of Philosophy, Heinrich-Heine-Universität, 40204 Düsseldorf, Germany

e-mail: sam.taylor@hhu.de

P. R. Sutton

Institute of Linguistics and Information Science, Heinrich-Heine-Universität,

40204 Düsseldorf, Germany

e-mail: peter.sutton@hhu.de

© The Author(s) 2021

S. Löbner et al. (eds.), *Concepts, Frames and Cascades in Semantics,*

Cognition and Ontology, Language, Cognition, and Mind 7,

https://doi.org/10.1007/978-3-030-50200-3_15

about the hypothesis space over which Bayesian inference operates. There are a number of possible candidates that could be sourced from cognitive scientific theories—e.g. prototypes, bundles of exemplars, or theory-like structures (Carey 1985; Lakoff 1987; McClelland and Rumelhart 1981; Nosofsky 1988; Rehder 2003). However, it has become standard practice to assume that Bayesian models operate over representations of unstructured lists of features; e.g. feature list representations (Anderson 1991; Sanborn 2006; Goodman et al. 2008; Shafto et al. 2011).

In this paper, we introduce and motivate frames as a candidate for the representations that mediate between (sensory) input and behavioural output, and as the representational format over which Bayesian inference operates in a Bayesian model of category learning. In other words, we introduce frame-theoretic representations (attribute-value structures) as the representational format of the data observed and operated on by the model. Our argument is that the resulting frame-theoretic model of Bayesian category learning is a theoretical improvement on feature list models, because our model can make fine-grained discrimination between competing categories without basing the weighting of attribute values on supervised training data. This is the case because frames—as the representational format of the input to our model—are not mere unordered lists of features, but, rather, are recursive attribute-value structures organised around a central node. For example, instead of three features such as **fur**, **black**, and **soft**, frames represent how these features are related by defining each feature as the value of some attribute i.e., that **fur** has (at least) two attributes COLOUR and TEXTURE, and that the values of these attributes are **black** and **soft**, respectively. As such, frames can be interpreted as assigning attribute values more or less weight depending on properties defined in terms of the structure of frames themselves. As a rough heuristic, our model proposes to weight attribute values as more or less diagnostic depending on whether or not they appear more centrally within a frame. In other words, our model takes a feature's 'path distance' from the central node to determine the diagnosticity of that feature for a given category.

As an example, suppose that the **fur**, **black**, and **soft** values appeared in a frame for a cat. Since, **black** and **soft** are values of attributes of **fur**, and **fur** is the value of an attribute of **cat**, a parameter based on distance from the central node would rank **black** and **soft** lower than **fur**. By incorporating this diagnosticity weighting in our model, we develop a frame-theoretic model of Bayesian category learning that introduces constraints on the most probable categories in terms of the diagnosticity of the observed features of entities being categorised.

The structure of this paper is as follows. In Sect. 2, we consider weighted Bayesian models of categorisation and argue that there is space to introduce a model that weights the relative diagnosticity of observed features that is not based on labelled training data. Then, in Sect. 3, we introduce a frame-theoretic representation of observed data and categories (e.g. the input and output of a categorisation model), in which frames are recursive attribute-value structures (Barsalou 1992; Barsalou and Hale 1993; Löbner 2014; Petersen 2015; Ziem 2014). Building upon this claim, we argue that the informational-structure of frames can be used to introduce a constraint on the relative diagnosticity of information encoded within a category and/or set of categories, where diagnosticity can be defined partly by properties of frame

structure (distance from the central node). Finally, we outline how feature list models of Bayesian category learning can be extended to operate over frames. On our frame-theoretic approach, the information-structural constraints of the model's frame representational-input influences the conditional probability of possible sets of categories by weighting the diagnosticity of the features of entities being categorised. We consider possible challenges to our model and possible future developments, before concluding that our model is better suited to describe and explain the unsupervised process of categorisation than comparable feature list based alternatives.

2 Weighted Bayesian Models of Categorisation

Categorisation is the cognitive process of representing given (natural) domains according to relevant features or properties. These features can be distinguished by our sense modalities—e.g. when we categorise objects in terms of their shape, size, or smell. But these features can also be distinguished by their informational content—e.g. when we can categorise foods in terms of their social role or nutritional content, or when we can categorise animals in terms of their ecological niches or taxonomic group (Shafto et al. 2011). In Bayesian models, categorisation occurs as the result of the model probabilistically grouping together sets of objects with shared features (e.g. **yellow**, **curved**). For instance, in the domain of, say, fruits, **yellow** and **curved** objects will have a relatively higher probability of being categorised together than all **yellow** objects, since other yellow fruits differ widely in their other properties (shape, size etc.), meaning that a clustering of all yellow fruits would yield a category with a below-optimal similarity of features. In this way, Bayesian models of categorisation explain how objects or sets of objects come to be categorised as one type or another (Anderson 1991; Tenenbaum 1999; Fei-Fei and Perona 2005; Wu et al. 2014 amongst many others).

An important question for Bayesian models of categorisation, however, is how models should represent input feature spaces, and, furthermore, how the representation of feature spaces influences the process of Bayesian categorisation. On many approaches to Bayesian category learning, feature inputs are represented as unordered lists of features (Anderson 1991; Sanborn 2006; Goodman et al. 2008; Shafto et al. 2011). And, on this approach, Bayesian categorisation proceeds by making the most probable categories those categories that group input stimuli together around a maximally optimal number of shared features. But, unless weights are added to lists of features in some principled way, this approach can be criticised for failing to provide an account of the relative importance of the features around which categorisation occurs. For example, on this approach the features of **colour**, **shape**, **texture**, **genus**, and **region of first domestication** all count as equally relevant for the differentiation of, say, bananas and oranges. And this seems counter-intuitive, because the representation of certain features—say, **colour** and **shape** in the case of bananas and oranges—appears to be more important for categorisation and so should have a bearing on what is taken to be the maximally optimal grouping of shared features.

In order to resolve the problem of uniformly diagnostic features, weights have been added to Bayesian models of categorisation, which make different features more or less diagnostic for specific categories. Such weighted models, however, face the challenge of finding a principled way to assign weights to individual features. For example, Hall (2007) makes use of a “decision tree-based filter method for setting [feature] weights,” where feature weights are estimated by constructing an unpruned decision tree and looking at the depth at which features are tested in the tree (Hall 2007, p. 121). Similarly, Wu et al. (2014) assign weight values to features by allowing the model to construct an unpruned decision tree that can be used to estimate each feature’s dependence on other features (Wu et al. 2014, pp. 1675–1676). These example models—and many others like them—have contributed to a growing literature that aims to improve the performance of naive Bayesian models while retaining their simplicity and computational efficiency. Notably, however, models which assign weights to features do so on the basis of, for example, frequency of features for categories, where categories are established via supervised learning.

It follows that the weighting schemas implemented by frequency-based approaches are derived from periods of supervised learning; that is, they are schemas that are dependent upon the input of supervised training data (Wu et al. 2014, p. 1676). In principle, there is nothing wrong with the application of such supervised training-based weighting schemas. However, the simplicity and tractability of models based on naive Bayesian assumptions is attractive (Pham 2009), especially if such models can be used in unsupervised learning tasks. This is the challenge that we take up in this paper. We develop a model that maintains the independence assumptions of naive Bayes, whilst assigning weights to features without appealing to weighting schemas derived from a period of supervised learning. The price to pay for this is that one must enrich the data that is input into the model. We do this by taking the input data to be in representational format of frames and not of feature lists. Our justification for this move is set out in Sect. 3, where we argue that there is support for the view that human cognition is structured around richer structures than lists of features and, therefore, that the data made available to learning models ought to be enriched. Furthermore, we argue that the hierarchical structure of frames allows models to assign weights to attribute values in frames.

In the remainder of this paper, we develop a Bayesian frame-based model of category learning. Our model will assign weights to features in virtue of the information structure of the feature spaces observed by the model.¹ In doing so, we drop the assumption that the input feature spaces over which Bayesian models operate are themselves flat and uniformly diagnostic for all categorisation tasks. Our claim is that the relative diagnosticity of features for categories can be captured by enriching the representational format of the data observed by the model. Such an enrichment, we claim, makes explicit how the probability of a system of categories can be cal-

¹Many Bayesian models category learning already presuppose that observed features have an informational structure that makes them more or less diagnostic for a given category, because they introduce certain features—e.g. colour—without making explicit that other features must also be observed; e.g. they introduce the feature **colour** without making explicit that the feature **shape** must also be observed.

culated not only from features (the values of attributes in our terms), but also from the structure of the data itself (such as the path distance that attribute is from the central node). The end result, therefore, is that certain, observed features—e.g. the features **colour** and **shape** in the group of observed features **colour**, **shape**, **texture**, **genus**, and **region of first domestication**—will have more of an influence on the probability of categorising the observed data as one category or another—e.g. as banana or orange.

To be clear, we accept that the evaluation of our model will ultimately be empirical, whereby the model is compared to actual human performance in the course of experimental testing. However, the contribution of this paper is the theoretical development of a model that shows promise as an improvement on current models of Bayesian category learning, since it derives relative feature diagnosticity in an unsupervised manner.

3 Frames

According to Barsalou (1992), frame representations capture the general format of cognition. As attribute-value structures, frames represent both the “general properties or dimensions by which the respective concept is described (e.g., COLOR, SPOKESPERSON, HABITAT ...)” and the *values* that each property or dimension takes in any given instantiation “(e.g. [COLOR: **red**], [SPOKESPERSON: **Ellen Smith**], [HABITAT: **jungle**] ...)” (Petersen 2015, p. 151). Thus, “a frame is a representation of a concept for a category which is recursively composed out of attributes of the object to be represented, and the values of these attributes” (Löbner 2014, p. 11). For Barsalou, an attribute is “a concept that describes an aspect of at least some category members”; and values are “subordinate concepts of an attribute” (Barsalou 1992, pp. 30–31). And, thus, a picture emerges of frames as representations of categories that encode, at the attribute level, general properties, dimensions, or aspects of the category in question; and, at the value level, the values taken by specific instantiations of the category in question.

Frames, then, are constituted of attribute-value pairings, where for “every attribute there is the range of values which it can possibly adopt” and “The range of possible values for a given attribute constitutes a space of alternatives” (Löbner 2014, p. 11). For example, an attribute such as COLOUR maps entities to colour values (e.g., [COLOUR: **red**]), and an attribute such as SHAPE maps entities to geometrical values (e.g., [SHAPE: **round**]).² Frames can themselves be represented by directed-graphs, whereby labelled nodes specify instantiated regions of the value space and arcs

² There is an open question about how value spaces are learned by individual subjects. We shall not answer this question here, although we find it plausible that individual subjects have access to value spaces as the result of “hyperpriors” determined by the subject’s biological phylogeny, biological and social ontogeny, and sociocultural embedding (cf. Clark 2015; Newsome 2012).

Fig. 1 Lolly frame
(Petersen 2015)

specify attribute designations of regions in the value space (see Fig. 1).³ Importantly, however, frames cannot be reduced to simple lists of features, because:

[...] it is not possible to simply replace the nodes in the frame definition by their labels, since two distinct nodes of a graph can be labeled with the same type. E.g., we could modify the lolly-frame in [Fig. 1] so that the stick and the body of the described lollies were produced in two distinct factories, where one is located in Belgium and one in Canada. (Petersen 2015, pp. 49–50)

Two questions arise, the answers to which are important for justifying our model:

- (i) Why should we assume that the frames are the representations that mediate between (sensory) input and categorisation of that input (as opposed to feature lists)?;
- (ii) What benefits do frames have as such input over feature lists?

Our simple answer to (i) is that the construction of feature lists implicitly assumes a richer relation between features, which is made explicit when we construct frames. Take the frame in Fig. 1. As a feature list, one could represent part of this information with the following features **has a stick**, **has a body**, **body is red**, **stick is green**. For the latter two in particular, the alternative would be to list two incongruent colour features **red** and **green** (resulting in potential contradiction). Yet, given that features must be more fully specified in this way, such lists of features simultaneously assume an attribute-value structure and make the structure invisible to any model that attempts to form categories on the basis of those features. (Bear in mind, that for a categorisation model, the features **has a stick**, **has a body**, **body is red**, **stick is green** may as well be represented as f_1, f_2, f_3, f_4 , since the fact that two features share ‘stick’ and two features share ‘body’ as part of their labels is not something that a model based on feature lists can access.) Therefore, there is a very real sense in which providing feature lists as data input sells itself short by both implicitly assuming a

³Frames can also be represented as attribute-value (AV) matrices (cf. Carpenter 1992; Petersen 2015).

richer structure to the data, but also not allowing any learning model to access that structure.

With respect to (ii), our claim is that the reason why frames are useful and relevant to categorisation is that they can be used to constrain information. In the first place, frames provide constraints on the range of values at any given node, because “information represented in a frame does not depend on the concrete set of nodes. It depends rather on how the nodes are connected by directed arcs and how the nodes and arcs are labelled” (Petersen 2015, p. 49). In other words, if we assume that frames are the category representations that mediate between (sensory) input and behavioural output, then it follows that categories must have a structure that relates the general properties, dimensions, or aspects of a category to the possible values that such general properties, dimensions, or aspects can take. For example, if the value of COLOUR is given as square—e.g. [COLOUR: **square**]⁴—then it is clear that the established ‘category’ is, in fact, no category at all (**square** is not a possible COLOUR value). Thus, it follows that even where a notional ‘category’ contains attribute-value pairs, it may still follow that the ‘category’ in question is impermissible because some of the attributes are assigned infelicitous values.

A second way in which frames constrain information derives from the fact that they are recursive (the value of one attribute can itself have attributes). The central node (graphically, the double-ringed node) indicates what the frame represents (i.e., lollies in the case of Fig. 1). Attribute-value pairs ‘closer’ to the central node encode relatively important, but general, information about the represented object. And attribute-value pairs ‘further’ from the central node encode relatively less important, but more specific, information about the represented object (because they are, e.g., values of attributes of values of attributes of the central node). For example, in Fig. 1 the ‘closer’ attribute-value pairs specify what physical structure and component parts the lolly in question has; and the ‘further’ attributes specify the colour and producer of these components. It follows, therefore, that those attribute-value pairs that are closer to the central node are more likely to be diagnostic of the category into which the object represented should be sorted. Thus, we can conclude that, at least as a rough heuristic, frames with more uniform ‘closer’ attribute-value pairs will represent more likely categories than frames with less uniform ‘closer’ attribute-value pairs (even if the latter has more uniform ‘distant’ attribute-value pairs), because the former categories will be more effective in organising (sensory) input according to more ‘central’ properties.⁴ For example, looking again at the lolly frame in Fig. 1, a category containing only red things that may or may not have bodies and sticks will be a less probable category than one which contains objects of different colours that all have bodies and sticks.

In an important paper, Shafto et al. (2011, p. 5) observe that standard approaches to modelling category learning appeal to a ‘single system model’ of categorisation

⁴The question of what attribute-value pairs are the most diagnostic for any given (sensory) input or object is an empirical question which we would like to pursue further. Such empirical research is usually undertaken by considering typicality judgements or typicality rankings (Djalal et al. 2016; Rips 1989).

(although the aim of their paper is to develop and motivate a more sophisticated *cross categorisation* model). They define a single system model of categorisation as a model that “embodies two intuitions about category structure in the world: the world tends to be clumpy, with objects clustered into a relatively small number of categories, and objects in the same category tend to have similar features.” So a single system model “assumes as input a matrix of objects and features, D , where entry $D_{o,f}$ contains the value of feature f for object o ” (Shafto et al. 2011). For the single system model, therefore, “there are an unknown number of categories that underlie the [input],” but the objects that are categorised within the same category “tend to have the same value for a given feature” (Shafto et al. 2011). As a result, the ultimate goal of the model is to infer—by means of establishing groupings within D according to shared features—likely set of categories, $w \in W$, where the process of categorisation occurs as the result of a trade-off between two *goals* or *constraints*: “minimizing the number of [categories] posited and maximizing the relative similarity of objects within [each category]” (Shafto et al. 2011).

Such models, and the model we develop here, make independence assumptions regarding feature spaces (value spaces for attributes, in our terms). For example, that the colour of the body of a lolly is independent from the manufacturer of the body. Single system models of categorisation proceed by partitioning the hypothesis space—e.g. the objects in the input matrix, D —according to more or less probable sets of categories, w . Finally, the posterior probability of hypotheses given the data ($p(w|D)$) is calculated, where this posterior probability is influenced by the extent to which objects grouped into categories share features (are homogeneous) (Shafto et al. 2011, p. 6).

Replacing feature lists with frames amounts to making the input matrix D richer. When the input matrix specifies frames and not merely feature lists, the structure of frames can be used to define parameters for a categorisation model. Here, we investigate the possibility of exploiting the fact that frames are hierarchical. Graphically, each node can be measured in terms of path distance from the central node. Added to the fact that attributes are functional, this allows us to define, as a rough heuristic, the relative diagnostic strength of an attribute value from that value’s distance from the central node. Hence, by including in D weighted values, where weights are derived from frame structure, Bayesian inference operates over a richer information set.

Consider the simple feature list matrix for four witnessed objects a, b, c, d and four features **fur**, **feathers**, **brown**, **black** in Table 1. If we assume that, even as feature lists, these features can be grouped into classes, which we label colour and layer, the joint probability distribution for the data can be given as shown in Table 2.

The possible groupings of objects into categories for this sample already numbers 15. Four such are given in (1) with the additional information of how these groupings relate to the features of objects.

$$\begin{aligned}
 w_1 &= \begin{cases} \mathbf{fu} \wedge \mathbf{br} = \{a\} \\ \mathbf{fu} \wedge \mathbf{bl} = \{b\} \\ \mathbf{fe} \wedge \mathbf{br} = \{c\} \\ \mathbf{fe} \wedge \mathbf{bl} = \{d\} \end{cases} & w_2 &= \begin{cases} \mathbf{fu} &= \{a, b\} \\ \mathbf{fe} \wedge \mathbf{br} &= \{c\} \\ \mathbf{fe} \wedge \mathbf{bl} &= \{d\} \end{cases} \\
 w_8 &= \begin{cases} \mathbf{fu} = \{a, b\} \\ \mathbf{fe} = \{c, d\} \end{cases} & w_{15} &= \{ \mathbf{fu} \vee \mathbf{fe} \vee \mathbf{br} \vee \mathbf{bl} = \{a, b, c, d\} \end{aligned} \tag{1}$$

However, the number of possible sets of categories increases exponentially with the number of objects. This presents a categorisation challenge. Given a huge number of hypotheses for categorising a set of objects, the options must be whittled down. Bayesian approaches to categorisation can do this by calculating the maximum probability for some set of categories w_i , given the data D , namely: $\text{MAX}_{w_i \in W} [p(w_i|D)]$ (such that these probabilities can be updated in the light of new data). (Other alternatives include Markov Chain Monte Carlo Variational Bayesian methods.) For example, Shafto et al. (2011), following Anderson (1991), argue that this probability depends on the prior probability of assigning objects to categories (in a set of categories w) and the probability of the data given a set of categories.

We adopt Shafto et al.'s (2011) use of two parameters and the way in which they contribute to calculating $p(w|D, \alpha, \delta)$ ⁵:

$$p(w|D, \alpha, \delta) \propto p(w|\alpha) \times p(D|w, \delta) \tag{2}$$

In (2), $p(w|\alpha)$ contains the parameter α which sets the extent to which the number categories should be minimised. $p(D|w, \delta)$ contains the parameter δ which sets the extent to which features of objects within categories should be similar (i.e., that members of categories should have the same feature/attribute values).

Table 1 Distribution of skin covering and colour features (simulated)

	a	b	c	d
fur	1	1	0	0
feathers	0	0	1	1
brown	1	0	1	0
black	0	1	0	1

Table 2 Joint probability distribution: $f_{L,C}(l, c)$

		LAYER (L)	
		fur	feathers
COLOUR (C)	brown	0.25	0.25
	black	0.25	0.25

⁵Our model differs from theirs, however. See Appendix 1.

As a simple example of how these parameters work, take the data in Table 1. If the α parameter is set to maximally minimise the number of categories, then maximising $p(w|\alpha)$ would select w_{15} in (1); namely, a singleton set of one category that includes all objects so far observed. If, however, the parameter δ is set to maximise feature harmony within categories, then maximising $p(D|w, \delta)$ would select w_1 in (1); namely, a set of categories that contains as many categories as there are ways to distinguish objects by their features.

Such feature list models have been implemented for categorisation tasks (Chater and Oaksford 2008; Shafto et al. 2011). However, notice that for some data sets, although we would intuitively categorise some entities together, unweighted feature lists provide an insufficient amount of information to distinguish between competing hypotheses. Take, once more, the data in Table 1. No matter how one sets parameters such as α and δ in a feature list based Bayesian categorisation model, the probability value for w_8 in (1) could not differ from the value for w_9 in (3):

$$w_9 = \begin{cases} \mathbf{br} = \{a, c\} \\ \mathbf{bl} = \{b, d\} \end{cases} \quad (3)$$

The reason for this is because even if we grant that a model can be set up to see brown versus black and feathers versus fur as two distinct comparison classes, the flat nature of feature lists does not allow for (observed) relations between features to be expressed, which, were they articulated, could be used to inform judgements regarding probable sets of categories. In other words, as has been recognised, feature lists must, at the very least, be weighted in some principled way. The problem is that, in an unsupervised learning task, it is difficult to justify the selection of one feature over another.

Given frames as input data, however, such weightings can be defined by parameterising the structure of frames themselves. In other words, with frames, a categorisation model can be defined that can distinguish cases such as w_8 and w_9 . This is made possible because frames introduce a hierarchy between feature values in virtue of the fact that some values are values of attributes of other values. For the case in hand, for example, **black** and **brown** could be observed to be values of a COLOUR attribute, such that COLOUR is an attribute of the values **fe** and/or **fu**.⁶ That is to say the data in Table 1 could license the attribute-value structure shown in Fig. 2.

⁶In this paper, we are making the assumption that fur/feather-based categories are preferable. We take this to be reasonable on common-sense grounds. However, we also accept that there may be cross-cultural variation in the kinds of feature-based categories preferred. For example, it may be the case that individuals in certain cultures—e.g. Yucatec-speaking cultures—prefer material-based categories, while individuals in other cultures—e.g. English-speaking cultures—prefer shape-based categories (Lucy and Gaskins 2001). The kinds of cross-cultural differences that may be apparent in categorisation tasks cannot be dealt with adequately in this paper due to lack of space. Still, it is worth noting that our model—like any other Bayesian model of category learning—could be supplemented with further constraints to account for such differences in categorisation tasks. Such supplementation would first have to be justified in the light of ongoing debates about the relation between language, culture, and thought (cf. McWhorter 2014; Lucy 1992a, b).

Fig. 2 Attribute-value structure for data in Table 1

Table 3 Distribution of fur layer and colour features relative to distance (simulated)

	a	b	c	d
⟨fur, 1⟩	1	1	0	0
⟨feathers, 1⟩	0	0	1	1
⟨brown, 2⟩	1	0	1	0
⟨black, 2⟩	0	1	0	1

Our proposal is that, in general, the importance of the similarity of feature values of objects within categories is proportional to how ‘close’ these feature values are to the central node measured by (minimum) path distance. The intuitive idea is that properties of objects within the same category tend to be similar, at least in terms of type, when these properties are more diagnostic of the category in question (see Sect. 3). Take the frame from Petersen (2015) in Fig. 1. The type of value for the BODY and STICK attributes will be very similar across different lollies. Indeed, if something had, e.g., lolly properties but no stick, one might judge it to be a sweet, not a lolly. However, the shape, colour, and producer for each lolly component may vary to a greater extent without giving one cause to judge, e.g., that two differently coloured objects belong to different categories *qua lolly* or *not a lolly*.

Using unweighted feature lists alone, one cannot formally capture that similarity between values is more important for more central nodes. With frames we can. Given that we will not here be exploiting further properties of frames, data sets can be minimally changed to include a distance measure. For the frame in Fig. 2, for example, V_1 measures a distance of 1 from the central node. V_2 measures a distance of 2. (For more complex frames, this means that there may be multiple values that measure the same distance.⁷) This requires a fairly minimal adjustment in how data sets are represented. The data in Table 1, for example, will be represented as in Table 3. The adjustment made is that we now represent features as pairs $\langle f, d \rangle$ where f is a feature (e.g. **brown** or **feathers**) and d is a measure of distance such that $d \in \mathbb{N}$. This change is not trivial. Enriching the data set could be seen as some kind of cheat, i.e., by providing more information that guides the process of forming categories. However, as we argued in Sect. 3, such structure is often implicit in feature lists, even if it invisible to the learning model. In our model, we make this implicit information available.⁸

⁷We assume, in cases where a node is connected to the central node along multiple paths, that this is calculated as the minimum distance.

⁸It should be stressed that we lose a lot of information by compressing frames in this way. However, we do this for simplicity and do not rule out that retaining more information in frames may be required in future developments of this model.

A full specification of our model is given in Appendix 1. In brief, we calculate the value for $p(w|\alpha)$ from the sum of the entropy of the set of categories in w with respect to the assignment of objects to categories in w , weighted by α . In other words, in terms of the average amount of information required to determine which object a category is in, given a set of categories. A w with only one category will minimise entropy (no information is required to know which category an object is in because all objects are in one category). This translates into a high value for $p(w|\alpha)$. Depending on the value of α , a w with many categories will have comparably higher entropy (especially if the categories are evenly distributed/of similar size). This translates into a comparably lower value for $p(w|\alpha)$. Values of $p(D|w, \delta)$ are calculated from the δ -weighted entropy of each category with respect to the features of objects within that category. If all objects within each category have the same features, then entropy will be minimised (one would need no information to know which features an object has given the category it is in). This translates into a high value for $p(D|w, \delta)$. If objects in the same category differ with respect to their attribute values, then, depending on the setting for δ , this probability will be lower.

The difference between our model and one based on feature lists, therefore, is that unsupervised feature list models do not have a principled way to weight similarity with respect to some features more heavily than similarity with respect to others. For feature list models, given the data set in Fig. 1 and w_8 and w_9 from (1) and (3), for example, $p(w_8|D, \alpha, \delta) = p(w_9|D, \alpha, \delta)$ for all settings of α and δ . However, our frame-based model can discriminate between these two sets of categories. Objects in categories in w_8 have the same attribute values at distance 1 from the central node (viz. **fe** and **fu**), but different attribute values at distance 2 from the central node (viz. **br** and **bl**). In contrast, objects in categories in w_9 have different attribute values at distance 1 from the central node (viz. **fe** and **fu**), and the same attribute values at distance 2 from the central node (viz. **br** and **bl**). (See Appendix 1 for details.)⁹

3.1 Challenges and Future Developments

Refining the model to discriminate between subkinds/superkinds. This kind of model opens up an intriguing avenue for further research: we could define levels of granularity for categorisation by manipulating the function which underpins δ . For example, relatively coarse-grained categorisation would prioritise similarity of object features only for nodes that are small distances from the central node. This

⁹We do not claim that there is no other way to do this. For example, possible sets of categories, formed from unweighted feature list input, could be ranked according to other principles such as *simplicity* in which sets of categories are preferred if they minimise similarities within categories and maximise differences between categories (Chater 1999; Pothos and Chater 2002). Indeed, it is an open and interesting question whether our model ends up approximating the results of a simplicity-driven strategy, or, if not, whether both a frame based input and a simplicity-driven categorisation strategy could be combined in some way. We leave the comparison between our model and others for future work.

Fig. 3 (Partial) Frame for BIRD

Fig. 4 Partial frame for SHOE

might, for example, cluster birds together and mammals together. If, however, δ is set to push towards similarity of values in ‘further out’ nodes, then distinctions between categories would be more fine grained. This could, for example, allow for the BIRD category to be partitioned into species of birds. The reason for this is that there is a general tendency for birds to be similar with respect to values closer to the central node (e.g. **feathers**, **wings**, **beak** etc.), but dissimilar with respect to less central values. For example, beaks, wings, and feathers may differ with respect to shape, size, and colour. The basic idea is shown in Fig. 3. If values at distance 1 from the central node are enforced to be similar ($V_{1.1}$, $V_{1.2}$, and $V_{1.3}$), but values at distance 2 can differ ($V_{2.1}$ – $V_{2.5}$), then we would expect birds to be categorised together. However, if the setting for δ was such that values at distance 1 and at distance 2 were enforced to be (more-or-less) similar, we would get a categorisation of, say, different bird species.

An interesting avenue for further research is whether or not our model, which is a single system model in the sense of Shafto et al. (2011), could be used as a cross categorisation model by manipulating the function that underpins the δ parameter.

Distance may be insufficient as a measure. Our model has limitations as a result of our simplistic adoption of distance from the central node as the basis for justifying the weighting of certain attribute values over others, namely, for some cases, such as a coarse measure is unlikely to get the right results. For example, take a frame for shoes such that one wishes to discriminate high-heeled shoes from loafers. In such a case the height of the heel is surely a highly diagnostic factor. However, as indicated

by Fig. 4, other, far less relevant factors, such as the colour of the heel will appear at the same distance from the central node. Developments of our account will therefore have to investigate if there are other features of frames that can be parameterized in a categorisation model to capture such cases. For example, an extra feature of frames that we have not discussed here are constraints between values. For example, finding out the height of a shoe's heel may be highly informative as to other attribute values (such as the shape of the upper, the (un)likelihood of shoelaces etc.). One possible extension would therefore be to enrich the model with a parameter based upon numbers of constraints a node has linking it to other nodes. (The colour of a heel will be less likely to constrain other values than the height of the heel.)¹⁰

Necessity of empirical verification of the model. We submit that our frame-theoretic model of Bayesian category learning is an important theoretical development in one crucial respect: the model incorporates weights on the relative diagnosticity of attribute-value pairs without having to index such weightings to properties discerned from a period of supervised learning. In other words, our model provides an unsupervised way of introducing weights on the relative diagnosticity of attribute-value pairs, such that one need not train the model on a data set already imbued with category distinctions. However, we also accept that, in this paper, we have only been able to make explicit a *theoretical* difference between our model and comparable alternatives. It follows that our model—if it is to be taken as an accurate representation of human performance in categorisation tasks—must be empirically tested. That is, experimental methods must be employed to compare the categorisation performance of our model with the categorisation performance of other available models. In this way, our model must be comparatively evaluated according to how well it accounts for a given set of data relating to human performance, so that it can be empirically demonstrated that our model better explains human performance than its rivals. We therefore plan to test our model empirically in future research.

4 Conclusion

Although a number of representational formats have been exploited to account for the input to Bayesian categorisation models, it remains unclear which is best suited to modelling human categorisation. On the received view, Bayesian inference is taken to operate over input in the form of object-feature list matrices. Although such models have made progress, we have argued here that they only have sufficient discriminatory power because they tend to implement weighting schemas based on supervised learning (weights are derived from exemplars of categories provided in a period of supervised (or semi-supervised) learning).

Our central contribution has been to introduce and exploit frames as the representational format of the input to Bayesian models of category learning. Frames have a richer informational-structure than do feature lists, and so can be used to determine

¹⁰Such an enrichment would amount to dropping many of our independence assumptions, however.

the weighted diagnosticity of the information encoded within a category. As a result, the frame-based model we developed can discriminate between competing sets of categories without having to define weights based on samples of data labelled with categories. In other words, we have given a theoretical basis for a Bayesian categorisation model that, in principle, can approximate weighted naive Bayesian models without a period of supervised learning or weakening the independence assumptions of such models. This follows because the structure frames inherently have (and feature lists lack) can be used to define such weights directly from training data that is not tagged with categories to be learned.

Our adoption of frames as representations of data input and category output extends and consolidates the enlightened Bayesian paradigm, which looks to developments in the cognitive sciences to inform Bayesian modelling techniques (Chater et al. 2011; Jones and Love 2011). As postulates of cognitive scientific theories, frames are already a well-established representational architecture (among many others, see Barsalou 1992; Löbner 2014; Ziem 2014). However, until now, the theoretical benefits of frames had not been made explicit within the context of Bayesian models of category learning. By arguing that frames allow for the development of a more intuitively discriminatory model of category learning based on enriched input, we hope to have shown one way that an account of categorisation based upon the mathematical ideals of Bayesianism can still be subject to principled representational constraints. Although we accept that more work is needed to spell-out the evolutionary and practical relationship between Bayesian inference and (mental) representations in the broader domain of cognitive development, we think that our frame-theoretic approach to Bayesian category learning serves as a welcome further step on the path to developing a mechanistically-grounded and formally rigorous picture of cognition.

Acknowledgements We would like to thank the participants of *Cognitive Structures: Linguistic, Philosophical and Psychological Perspectives* (2016) for their constructive comments and critique. We would also like to extend our thanks to the two anonymous reviewers for their helpful recommendations and advice about how the paper could be improved. This work was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) as part of the Collaborative Research Centre 991: The Structure of Representations in Language, Cognition, and Science, projects C09 and D02. Thanks also to the members of C09 and D02 for their support.

Appendix 1: A Frame-Based Bayesian Categorisation Model

Our model is based, like other single system models, on the calculation of $p(w|D, \alpha, \delta)$ from the joint probability distribution over w , D , α , and δ (elements of the model). We use the same formula (reprinted here with an M label on p to indicate the probability function based on this joint distribution):

$$p_M(w_i|D, \alpha, \delta) \propto p_M(w_i|\alpha) \times p_M(D|w_i, \delta) \quad (4)$$

Table 4 Definitions for elements of the frame based Bayesian categorisation model

$O = \{o_1, ..., o_n\}$	A set of observed objects
$C = \{c \mid c \neq \emptyset, c \subseteq O\}$	A set of categories
$W = \{w \mid w \neq \emptyset, w \subseteq C\}$	A set of sets of categories
$F = \{\mathbf{f} \mid \mathbf{f} \subseteq O\}$	A set of attribute values (i.e. a set of predicates of objects)
$\iota = f : F \rightarrow \mathbb{N}_{>0}$	A function from attribute values to their distance from the central node
$A = \{(\mathbf{f}, n) \mid \mathbf{f} \in F, n = \iota(\mathbf{f})\}$	A set feature-distance pairs (i.e. a set of distance indexed features)
$D = \{(\langle o, X \rangle \mid o \in O, X \subseteq A)\}$	The data: a set of tuples such that for each object, there is a set of feature-distance pairs
$\alpha \in [0, 1]$	The small categories preference parameter
$\delta = f : \mathbb{N}_{>0} \rightarrow \mathbb{R}$	The similar features preference parameter. A function from distance measures to real numbers

We maintain the small categories preference parameter α , but the similar features preference δ , on our model, sets the preference for how strongly distance from the central node affects the overall similarity score for a set of categories. Definitions of elements of the model are given in Table 4. Categories are sets of objects and category schemas are sets of categories. The data input for the model consists of frames, here simplified to objects paired with attribute values and a distance of this value from the central node. Distance from the central node forms the basis for the weighting of attribute values determined by δ .

We assume, for simplicity, that for any set of categories, w , no object is in more than one category and every object is in a category. (Sets of categories completely partition the domain of objects.) In other words, as given in (5), for a set of objects, O , for each w , we have a distribution over the categories $c_i \in w$ (the probability function is accordingly labelled O, w , we suppress O in most of the following since we will not consider cases for multiple O sets).

$$\sum_{c_i \in w} p_{O,w}(c_i) = 1 \quad (5)$$

The prior probability of a category c relative to a set of categories w is calculated as the number of objects in the category divided by the number of objects so far observed:

$$\text{For each } c \in w, \quad p_{O,w}(c) = \frac{|c|_w}{|O|} \quad (6)$$

Other distributions occur at the level of nodes in frames. Each node has a set of possible values (e.g., **red**, **green** etc., for COLOUR, and **feathers**, **fur**, **scales** etc. for COVERING). We say more about such distributions in Appendix 1.2.

Table 5 The effect of α on calculating the prior $p(w|\alpha)$ for the data in Table 3 restricted to w_1 and w_{15}

	α			
	1	0.5	0.1	0
$p(w_1 \alpha)$	0.20	0.3	0.47	0.50
$p(w_{15} \alpha)$	0.80	0.6	0.53	0.50

1.1 The α Parameter

The intuitive idea behind the calculation of $p_M(w|\alpha)$ is that w should minimise entropy over the object space (minimise the average amount of information required to identify in which category in w an object belongs). This is given in (7). If alpha is set to 1, then the probability is proportional to the inverse log of the entropy of w . If $\alpha = 0$, then, assuming a base-2 logarithm, for all $w \in W$, $p_M(w|\alpha) \propto 2^0$ (i.e. $\propto 1$), thus all $w \in W$ would receive the same prior.¹¹ In other words, there would be no preferential effect of reducing (or increasing) the number of categories.

$$\text{For all } w \in W : p_M(w|\alpha) \propto 2^{\left(\alpha \times \sum_{c_i \in w} (p_w(c_i) \times \log_2(p_w(c_i)))\right)} \quad (7)$$

As an example of how α operates, consider four objects a, b, c, d and a space of two category sets w_1, w_{15} . If $w_1 = \{c_1 = a, c_2 = b, c_3 = c, c_4 = d\}$ and $w_{15} = \{c_5 = \{a, b, c, d\}\}$, then, for varying vales for α , we get the results in Table 5 (values given to 2 decimal places).

1.2 The δ Parameter

The intuitive idea behind the calculation of $p(D|w_i, \delta)$ is that, with respect to the values for an attribute, each category should minimise entropy (weighted by distance the attribute is from the central node). In other words, minimise the average amount of information it takes to decide which properties an object has if it is in a particular category.

Given that each $d \in D$ is a tuple of an object and a set of attribute value-distance pairs, calculating $p_M(D|w, \delta)$ turns on calculating, for each category c in w , the probability that the objects in c have some particular value for the relevant attribute. Let $|\mathbf{f}_j|_{c_k, w, D}$ be the number of times the attribute value $\mathbf{f}_j$ occurs as a value in category $c_k \in w$ for a data set D . Let $|c_k|_{w, D}$ be the number of objects in $c_k \in w$. $p_{w, D}(\mathbf{f}_j|c_k)$ is, then:

¹¹The actual probability is calculated by dividing by the sum of the values given in (7) over all $w \in W$.

$$p_{w,D}(\mathbf{f}_j|c_k) = \frac{|\mathbf{f}_j|_{c_k,w,D}}{|c_k|_{w,D}} \quad (8)$$

namely, for a set of categories w , the total number of times objects in $c_k \in w$ have value $\mathbf{f}_j$, divided by the total number of objects in c_k . This forms a distribution for any set of attribute values that are the mutually exclusive values of some attribute (e.g., a distribution over **feathers** and **fur**, and a distribution over **black** and **brown** in our toy example).

The entropy values for attribute value spaces, given a category, are weighted depending on the distance d the feature is from the central node. This weighting is set by δ , which is a function from d to a real number in the range $[0, 1]$. The weighted entropy value for a category is, then, the sum of the weighted sum of the surprisal values for each attribute value, given a category, also weighted by δ . The weighted entropy value for a set of categories w is the weighted average of the entropy values for each category in w (relative to $p_w(c)$). So, for all $w \in W$:

$$p_M(D|w, \delta) = 2^\wedge \left(\sum_{c_k \in w} p_w(c_k) \times \sum_{(\mathbf{f}_j, n_j) \in \pi - 2(D)} \left(p_{w,f,c}(\mathbf{f}_j|c_k) \times \log_2(p_{w,f,c}(\mathbf{f}_j|c_k)) \times \delta(n_j) \right) \right) \quad (9)$$

Intuitively, $p_M(D|w, \delta)$ is a measure on how well the data is predicted by each w (weighted by δ). This value will be 1 if every piece of data (an object and its attribute values and distances) falls into a totally homogenous category with respect to the objects it contains. This is because the average amount of information to determine the attribute values of members of each category is 0. As categories get more and more heterogeneous, the value of $p(D|w, \delta)$ will get lower. This is because the average amount of information need to determine the attribute values of members of each category is high.

For example, for the data in Table 3, so with four objects a, b, c, d , and also with the four category sets w_1, w_8, w_9, w_{15} , if $w_1 = \{c_1 = \{a\}, c_2 = \{b\}, c_3 = \{c\}, c_4 = \{d\}\}$, $w_8 = \{c_5 = \{a, b\}, c_6 = \{b, c\}\}$, $w_9 = \{c_7 = \{a, c\}, c_8 = \{b, d\}\}$, and $w_{15} = \{c_9 = \{a, b, c, d\}\}$ and attribute values are as displayed in Table 3, then we get the impact of altering the δ function as given in Table 6 (values given to 2 decimal places). Since w_1 contains only singleton categories, the probability of the data given w_1

Table 6 The effect of δ on calculating the likelihood $p(D|w, \delta)$ for the data in Table 3 for w_1, w_8, w_9 and w_{15} , for $\delta(n_j) = n_j^0$, $\delta(n_j) = n_j^{-1}$, and $\delta(n_j) = n_j^{-2}$

	$\delta(n_j)$		
	n_j^0	n_j^{-1}	n_j^{-2}
$p(D w_1, \delta)$	1	1	1
$p(D w_8, \delta)$	0.5	0.71	0.84
$p(D w_9, \delta)$	0.5	0.5	0.5
$p(D w_{15}, \delta)$	0.25	0.35	0.42

is 1 no matter how $\delta(n_j)$ is defined, since for all attribute values and all categories $p_{w_1, f, c}(\mathbf{f}_j | c)$ equals 1 or zero (so the weighted entropy value is 0 and $2^0 = 1$). The worst performing is w_{15} , since this contains only one category so heterogeneity for features is high (this is mitigated a little when $\delta(n_j)$ is defined to decrease the homogeneity requirement for attribute values with larger distances from the central node).

We now turn to the comparison between w_8 and w_9 (which is important for our toy example). In the case where $\delta(n_j) = n_j^0$ (i.e. where $\delta(n_j)$ is always equal to 1), there is no weighting towards the importance of similarity of values with respect to being close to the central node. This gives us the same result as would be given for a simple unweighted feature list. In other words, given some things that are furry and black, furry and brown, feathered and black, and feathered and brown, the model has no preference towards grouping furry things together and feathered things together over grouping black things together and brown things together. When $\delta(n_j) = n_j^{-1}$, the result is that entropy is weighted to be halved for values at a distance of two nodes away from the central node. When $\delta(n_j) = n_j^{-2}$, the result is that entropy is weighted to be quartered for values at a distance of two nodes away from the central node. This translates into an increasing preference for no entropy at the inner most nodes and an allowance of higher entropy at further out nodes.

References

- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological review*, 98(3), 409.
- Barsalou, L. W. (1992). Frames, concepts, and conceptual fields. In E. Kittay & A. Lehrer (Eds.), *Frames, fields, and contrasts: New essays in semantic and lexical organization* (pp. 21–74). Erlbaum.
- Barsalou, L. W., & Hale, C. (1993). Frames, concepts, and conceptual fields. In I. Van Mechelen, J. Hampton, R. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 97–144). Academic Press.
- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.
- Carpenter, B. (1992). *The logic of typed feature structures*. Cambridge: Cambridge University Press.
- Chater, N. (1999). The search for simplicity: A fundamental cognitive principle. *Quarterly Journal of Experimental Psychology: Section A*, 52(2), 273–302.
- Chater, N., & Oaksford, M. (2008). The probabilistic mind: Where next? In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 501–514). Oxford: Oxford University Press.
- Chater, N., Goodman, N., Griffiths, T. L., Kemp, C., Oaksford, M., & Tenenbaum, J. B. (2011). The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science. *Behavioral and Brain Sciences*, 34(4), 194–196.
- Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.
- Djalal, F. M., Ameer, E., & Storms, G. (2016). The typicality ranking task: A new method to derive typicality judgments from children. *PLoS ONE*, 6(11), 1–17.
- Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Society Conference, Computer Vision and Pattern Recognition, 2005, CVPR 2005* (Vol. 2, pp. 524–531).

- Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1), 108–154.
- Hall, M. (2007). A decision tree-based attribute weighting filter for naive Bayes. *Knowledge-Based Systems*, 20(2), 120–126.
- Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34(4), 169–188.
- Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning and Behavior*, 36(3), 210–226.
- Lakoff, G. (1987). Cognitive models and prototype theory. In U. Neisser (Ed.), *Concepts and conceptual development* (pp. 63–100). Cambridge: Cambridge University Press.
- Löbner, S. (2014). Evidence for frames from human language. In *Frames and concept types* (pp. 23–67). Springer International Publishing.
- Lucy, J. A. (1992a). *Language diversity and thought: A reformulation of the linguistic relativity hypothesis*. Cambridge: Cambridge University Press.
- Lucy, J. A. (1992b). *Grammatical categories and cognition: A case study of the linguistic relativity hypothesis*. Cambridge: Cambridge University Press.
- Lucy, J. A., & Gaskins, S. (2001). Grammatical categories and the development of classification preferences: A comparative approach. In *Language acquisition and conceptual development* (pp. 257–283).
- McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1 an account of basic findings. *Psychological Review*, 88, 375–407.
- McWhorter, J. H. (2014). *The language hoax: Why the world looks the same in any language*. Oxford University Press.
- Newsome, W. (2012). Complementing predictive coding. *Frontiers in Psychology*, 3, 554.
- Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 700–708.
- Petersen, W. (2015). Representation of concepts as frames. In T. Gärdenfors, D. Gerland, R. Osswald, & W. Petersen (Eds.), *Meaning, frames, and conceptual representation* (pp. 39–63). Düsseldorf: Düsseldorf University Press.
- Pham, D., & Ruz, G. (2009). Unsupervised training of Bayesian networks for data clustering. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* (Vol. 465, No. 2109, pp. 2927–2948).
- Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorisation. *Cognitive Science*, 26(3), 303–343.
- Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 1141–1159.
- Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). Cambridge: Cambridge University Press.
- Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 726–731).
- Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, 120(1), 1–25.
- Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems* (Vol. 11, pp. 59–65). Cambridge, MA: MIT Press.
- Wu, J., Pan, S., Cai, Z., Zhu, X., & Zhang, C. (2014). Dual instance and attribute weighting for naive Bayes classification. In *2014 International Joint Conference on Neural Networks (IJCNN)* (pp. 1675–1679).
- Ziem, A. (2014). FrameNet, Barsalou frames and the case of associative anaphora. In *Studies in language and cognition* (pp. 93–112).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.