5

Valuations and learnability

Contents

5.1 The likeliness scale 115
5.2 Naive inference (likeliness update) 118
5.3 Learning 122

In this chapter we describe a rational, but low resolution, model of probability. We do this for two reasons: first, to show how a naive theory, using only discrete categories, can still explain how people think about uncertainty, and second, as a model for fitting discrete theories of valuation (which arise in many other contexts from moral judgments to household finance) into the overall `4lang` framework. In 5.1 we introduce *likeliness*, which we take to be a valuation of propositions on a discrete (seven-point) scale. In 5.2 we turn to the inference mechanism supported by the naive theory, akin to Jeffreys-style probability updates, and argue that valuations are, for the most part, computed rather than learned. After these preparations, in 5.3 we address what we take to be the central concern for any cognitively inspired theory, learnability. We divide the problem in three parts, learning of (hyper)nodes, learning of edges, and learning of valuations. We argue for a system powered by embodied cognition, with all three parts operating simultaneously.

5.1 The likeliness scale

Historically, the theory of probability emerged from the efforts of Pascal and Fermat in the 1650s to solve problems posed by a gambler, Chevalier de Méré (Rényi, 1972; Devlin, 2008), and reached its current form in Kolmogorov, 1933. Remarkably, not even highly experienced gamblers can extract high precision probability estimates from observed data: one of de Méré's questions concerned comparing the probabilities of getting at least one 6 in four rolls of one die ($p = 0.5177$) and getting at least one double-6 in 24 throws of a pair of dice ($p = 0.4914$). Four decades later, Samuel Pepys is asking Newton to discern the difference between at least two 6s when 12 dice are rolled ($p = 0.6187$) and at least 3 6s when 18 dice are rolled ($p = 0.5973$).

A. Kornai, *Vector Semantics*, Cognitive Technologies, https://doi.org/10.1007/978-981-19-5607-2_5

Here we make this phenomenon, the very limited ability of people to deal with probabilities, the focal point of our inquiry. These limitations, we will argue, go beyond the well understood limits of numerosity (Dehaene, 1997), and touch upon areas such as cognitive limits of deduction (Kracht, 2011a) and default inheritance (Etherington, 1987). We introduce *likeliness*, which we take to be a valuation of propositions on a discrete (seven-point) scale. We defer the issue of "computing" with these values, the inference mechanism supported by the naive theory, to 5.2. For the case at hand (low resolution probabilities, this will be much like Jeffreys-style probability updates, but the same mechanism is available for other, non-probabilistic updates as well.

We use the term 'likeliness' for a valuation on a 7-point scale 0,...,6 which only roughly corresponds to a discretized notion of probability (we avoid the more natural-sounding 'likelihood' as this already has a well-established technical sense). 0 is assigned to *impossible* events, $l(e) = 0$, and 6 to *necessary* ones. Note that in this regard l corresponds better to everyday usage in that zero probability events ($p(e) = 0$) do occur, and $p(e) = 1$ guarantees only that the event e has measure zero exceptions of occurring. $l(e) = 2$ means *unlikely*: an example would be traffic accidents. $l(e) = 1$ means *conceivable*, events that are unlikely in the extreme, but not forbidden by physical law. An example would be being struck by a meteorite.

There is a duality between x and $6 - x$ as in Łukasiewicz L_7, so $l(e) = 4$ is assigned to *likely* events such as traveling without an accident and $l(e) = 5$ to *typical* or *expected* ones. Almost all lexical knowledge falls in this last category: chairs are by definition furniture that support a seated person, and if a particular instance collapses under ordinary weight we say it failed (whereas we don't conclude that my car failed when I get in a traffic accident – alternative hypotheses such as driver error are readily entertained). Events that are neither likely nor unlikely are assigned the value 3.

Clearly, using exactly 7 degrees is somewhat arbitrary, but it is evident that using only 3 (say impossible, unknown, possible) would be a gross oversimplification of how people deal with probability, and using a very fine scale would create illusory precision that goes beyond people's actual abilities. With 7, we stick to a relatively small but descriptive enough scale. Even if one could argue that, say on cognitive grounds, 5 or 9 degrees would be better, the overall methodology would be the exact same, and everything below could be easily modified and worked out with that scale. Altogether, our choice of having a 7-degree scale is more of an illustration than a commitment, albeit one well supported by practical experience with semantic differentials (Osgood, Suci, and Tannenbaum, 1957).

The commonsensical valuation, which is our object of study here, differs from probabilities in several respects. The most important from our perspective is *lack of additivity*. At this point, it is worth emphasizing that the theory of likeliness valuation is not intended as a replacement of the standard (Kolmogorov) notion of probability, which we take to be the correct theory of the phenomena studied under this heading, but rather as an explanatory theory of how the *naive* worldview accounts for these phenomena. The fact that as a computational device the standard theory is superior to the naive theory is

no more a reason to abandon study of the naive theory than the superiority of eukaryotes is reason to abandon study of prokaryotes.

By lack of additivity we don't just mean lack of σ-additivity, but something that is already visible on finite sums. Consider the Law of Total Probability, that $p(A)$ can be computed as $\sum_n p(A|B_i)p(B_i)$ where the B_i provide a (typically finite) partition of the event space. The equivalent formulation with likeliness normed to 1 would be

$$l(A) = \bigoplus_i l(B_i) \otimes l(B_i \to A) \tag{5.1}$$

Here we retain the assumption that likeliness is a valuation in a semiring where addition $\oplus$ and multiplication $\otimes$ are defined, but instead of conditional probability we will speak about relevant implication $\to$ having a valuation of its own. The semiring of greatest interest is the one familiar from n-valued logic, where $\otimes$ is min, and $\oplus$ is max. In this simplified model, we allow two types of propositions only: standalone sentences A and sentences in the form of an implication $A \to B$ (see 5.2).

To put lack of additivity in sharp relief, consider the following commonsensical example: *all men are mortal*. If we take A to be eventual death, we have $l(A) = 6$. If we ask people to elicit causes of death B_i, they will produce a handful of causes such as cancer or heart attack that they consider likely ($l = 4$); some like accidents or tropical diseases they consider neither very likely nor very unlikely ($l = 3$); some like autoimmune diseases or freezing to death they consider less likely ($l = 2$); and some they consider conceivable but extremely unlikely such as murder/suicide or terrorism ($l = 1$). Needless to say, such valuations are not precisely uniform across people, but they do have high intrasubjective consistency (as measured e.g. by κ statistics). Since $l(B_i \to A)$ is by definition 6, we are left with an enumeration of causes:

$$l(A) = \bigoplus l(B_i) = \oplus_{i=0}^{6} \oplus_{l(B_j)=i} i \tag{5.2}$$

The problem here is that no amount of heaping on more of less likely causes will increase the $\oplus$ above the valuation of its highest term. The phenomenon is already perceptible at the low end: if we collect all conceivable causes of death from lightning strike to shark attack, we have 'death by (barely) conceivable causes' which itself is unlikely, not just conceivable.

In actual mortality tables, this phenomenon is reflected in the proliferation of categories like 'unknown', 'unspecified', and 'other', which take up the slack. Depending on the depth of tabulation, the catchall category typically takes up between .5% and 5% of the total data, which corresponds well to the lack of sensitivity below 1% observed in the de Méré and Pepys examples we started with.

Another obvious difference between the standard and the naive theory is the way extremely low or extremely high probability events are treated. When we want to draw the line between impossible and conceivable events, we don't rely on a single numerical cutoff. But if we take the proverbial 'one in a billion chance' as marking, in some fuzzy sense, the impossible/conceivable boundary, and use log odds scale, as argued by

Jaynes, 2003, the next natural order of magnitude (Gordon and Hobbs, 2017) brings us to $p = 0.0014$, which we can take to mark the conceivable/unlikely boundary, and the one beyond that to $p = 0.1118$, which marks the unlikely/neutral boundary.

In this reckoning everything between $p = 0.1118$ and $p = 0.8882$ is considered $l = 3$, neither particularly likely nor particularly unlikely. Likely events are between $p = 0.8882$ and $p = 0.9986$, while typical events are above that limit though still with a one in a billion chance of failure. As at the low end, the naive theory lacks the resolution to distinguish such failure rates from necessity (total absence of failure).

We should emphasize here that it is the overall logic of the scheme that we are vested in, not the particular numbers. For example, if we assume an initial threshold of one in a million instead of one in a billion, the limits will be at 0.0125 and 0.2008 (and by symmetry at 0.7992 and 0.9875), but the major characteristics of the system, such as the 'neither likely nor unlikely' category takes up the bulk of the cases, or that $l = 2$ cases are noticeable, whereas $l = 1$ cases are barely detectable, remain unchanged. Further, we should emphasize that such limits, however we set them, are not intended as a crisp characterization of human classification ability, the decision boundaries are fuzzy. Returning to lack of additivity, there may well be several likely causes of death beyond cancer and heart attack, but no closed list of such is sufficient for accounting for the fact that eventual death is typical (as assumed by Christian doctrine that posits Jesus as an exception), let alone necessary, as assumed by the irreligious. For this, we need a slack variable that lifts the $\oplus$ of the likely $l = 4$ causes to $l = 5$ or $l = 6$, which we find in B_n 'death by other causes'. We note that historically old age was seen as a legitimate cause of death, and only very recently (since the 1980s) do US coroner's reports and obituaries find it necessary to list the failure of a specific organ or subsystem as the cause of death, and a catchall category, *geriatric malady*, remains available in many countries.

Finally, in contradistinction to the standard theory, $\oplus$ can extend only to a handful of terms, especially as the terms are implicitly assumed independent. By the above reckoning, it takes less than 80 unlikely causes to make one neutral, and less than 8 neutral to make a likely one. The geometry of the likeliness space is *tropical* (Maclagan and Sturmfels, 2015), with the naive theory approximating the log odds (max) semiring.

5.2 Naive inference (likeliness update)

We have two types of propositions: stand alone sentences A and sentences in the form of an implication $A \to B$. A context is a (finite) collection of propositions, which can be represented by a directed graph: nodes of the graph denote propositions A and edges of the graph denote implications $A \to B$. The likeliness function is an evaluation acting on the graph: both vertices and edges can have numeric values between 0 and 6, 0 representing impossibility, 6 representing necessity.

Values $l(A \to B)$ belong to the inner model (adult competence, see 5.3) therefore they are hardly subject to change. Take the following example as an illustration. Snow-

bird is a ski resort in Utah. Say, for a typical European, Snowbird is related to traveling, skiing, and snowing with the likeliness

$$\begin{aligned} l(\text{Snowbird} \to \text{traveling}) &= 5 \\ l(\text{Snowbird} \to \text{skiing}) &= 5 \\ l(\text{Snowbird} \to \text{snowing}) &= 5 \end{aligned}$$

Such likelinesses express *typicality* of these relations. Skiing is related to some extent, say, to ski-accident, and ski-accident to death. Take the example below (for the sake of example we differentiate between ski-accidents and accidents; the latter excludes accidents occurring while skiing).

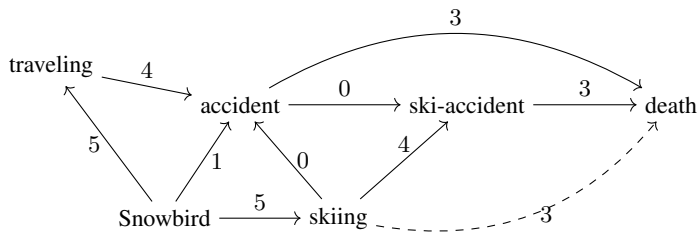

Fig. 5.1: Skiing, accidents, Snowbird

In a typical scenario one does not have any likeliness of the implication Snowbird→death inside the inner model. However, naive inference works: Snowbird *typically* implies skiing; skiing is *likely* to imply ski-accident; finally, it is neither likely nor unlikely that ski-accident results in death. Therefore, one may say [visiting] Snowbird is neither likely nor unlikely to result in death, i.e.

$$l(\text{Snowbird} \to \text{death}) = 3$$

In a similar manner, one could obtain the likeliness $l(\text{skiing}\to\text{death}) = 3$ by saying that skiing is *likely* to ensure a ski-accident, while it is neither likely nor unlikely that ski-accident results in death.

In virtue of the examples above we give a formal model. Let assume we have a finite directed graph $G = (V, E)$ and an evaluation $l : E \to \{0, \dots, 6\}$. We would like to evaluate edges of the complete graph on V that are not in E. Pick two vertices $a, b \in V$, $a \neq b$ and suppose $(a, b) \notin E$. Let $p = (v_1, \dots, v_n)$ be a path in G from $a = v_1$ to $b = v_n$. We write

$$l(p) = \min\left\{l(v_i \to v_{i+1}) : \ i = 1 \dots n-1\right\} \tag{5.3}$$

The value $l(p)$ expresses how likely the inference $a \to b$ is in case we are relying on the chain of already evaluated implications belonging to the path p. Then the value $l(a \to b)$ is obtained by

$$l(a \to b) = \min\{l(p) : \ p \text{ is a path in } G \text{ from } a \text{ to } b\} \tag{5.4}$$

In the example above vertices of the graph did not have likelinesses. Suppose we get new information about John: he is *likely* to be in Snowbird, i.e. $l(\text{Snowbird}) = 4$. What consequences can we draw? Being a typical European, if John is in Snowbird, then he must be traveling and it is really typical that people travel to Snowbird to ski. The information that $l(\text{Snowbird}) = 4$ propagates via the edges of the graph: the likeliness of those propositions that are related to Snowbird (that is, they are connected by an edge in the graph to Snowbird) will be updated given new information: $l(\text{traveling})$ and $l(\text{skiing})$ become 4. In the formal model, given the value $l(a)$ and a path $p = (v_1, \ldots, v_n)$ from $a = v_1$ to $b = v_n$, using the definition of $l(p)$ in equation (5.3) we can update the likeliness of b writing

$$l(b) = \max\{l(a), l(p) : \ p \text{ is a path in } G \text{ from } a \text{ to } b\} \tag{5.5}$$

This process of updating iterates: neighbors of just updated vertices get updates in the next round, etc. Supposing the graph is connected, all vertices are assigned with likeliness:

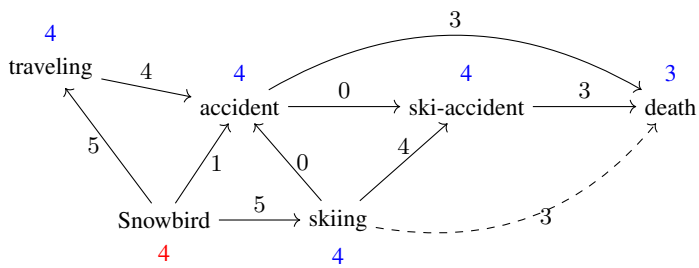

Fig. 5.2: John in Snowbird

Let us now suppose that we learn that John died abroad. The first column of Table 1 describes the default likelinesses we assign to various causes of death, with subsequent columns showing the updates based on whether we learn ($l = 6$) that the death took place in Reykjavík, Istanbul, or on a tourist trip, destination unspecified. Some rows are easy to explain: for example death at home in bed is considered likely, but if we know that John was on a tourist trip the implication is that he is not at home, and the likeliness is demoted to 1. Not 0, because there are extremely unlikely but not inconceivable scenarios whereby he fell in love with the place, bought a home, and resettled there, cf. Jaynes, 2003 5.2.2. This is a scenario that is, perhaps, worth considering if we know only

that John went to Reykyavík or Istanbul and tourism was merely an inferred, rather than explicitly stated, goal of the trip, but if we *know* it was a tourist trip and nothing more (last column) this is logically incompatible with being at home.

Cause of death	Default	Reykjavík	Istanbul	trip
in hospital	4	4	5	4
by accident (non-ski)	4	4	4	5
at home in bed	4	1	1	0
in war	1	0	0	1
by homicide	1	1	1	1
by suicide	2	2	2	1
by forces of nature	1	4	1	2
by ski accident	1	2	1	1

Table 5.1: Likeliness of cause of death

The same logic is operative in the next row (war): since we know there is no war in Reykjavík or Istanbul the likeliness is demoted to 0, but for a generic trip it is not, since we do know that there are war zones on the globe and John may have visited one of these.

We obtain that death by ski accident is less likely in Reykjavík (2) than in Snowbird (3) not because skiing is inherently more safe in Iceland, but simply because one can travel to Reykjavík for many reasons, and the likeliness that one goes skiing there is 3, perhaps 4, whereas to ski in Snowbird is typical (5). In connection of Reykjavík we are much more likely to think of death by forces of nature, as there are many natural dangers nearby, from volcanoes to geysers and sneaker waves, indeed this class rises to the top category (4).

This line also illustrates the nonmonotonic nature of the calculus: in general we consider death by forces of nature conceivable but unlikely in the extreme (1), knowing that John went to a tourist trip increases this to 2, but further learning that he went to Istanbul, not particularly known as a natural danger zone, demotes this back to 1.

With this, our rational reconstruction of the naive theory of probability is complete. This theory is not as powerful computational device as the standard theory, and generally only leads to rough estimates of likeliness. However, it is better suited for studying human cognitive behavior, as it requires very little data, and extends to a broad range of cases where the statistical data undergirding the standard theory is unavailable.

Importantly, the idea of updates extends well beyond probability/likeliness. Equation 5.2 and the update formulas remain meaningful for all other kinds of valuations. Following Osgood, May, and Miron, 1975 in essence, but not in terminology, we take the first principal component of their analysis as our primary example. Unluckily for the present discussion, they called this factor EVALUATION, but here we use a more specific name, GOOD/BAD, since it is just one (though clearly the most important) of many val-

uations. We take the evolutionary roots of this valuation to be hardwired. In S19:3.4 we wrote:

> The pain/pleasure valuation is largely fixed. A human being may have the power to acquire new tastes, and make similar small modifications around the edges, but key values, such as the fact that harming or destroying sensors and effectors is painful, can not be changed.

In terms of the thought vector analysis we sketched in 2.3, a word like *good* is strongly anchored as the density center of the projection of pleasurable mind states on the linguistic subspace, and a word like *bad* is similarly anchored as the density center of painful mind states. These two may not be orthogonal (clearly a mind state, including external and proprioceptive sensor states, can be both pleasurable and painful at the same time), but their existence is sufficient for setting up an initial valuation (say, on a scale of -3 to $+3$) that applies to novel mind states as well. It requires further acquisition work to generalize this from current sensory state to anticipation of future events,
`bad` which is what our definition of *bad* as `cause_ hurt` assumes. This requires nothing far-fetched, given that primary linguistic data, parents' utterances of *bad*, will typically refer to events and behaviors which, upon continuation, would indeed lead to bodily harm. Other valuations, such as Osgood et al.'s POTENCY or ACTIVITY are also richly embedded in sensory data, making them quite learnable early on. As the likeliness case shows, we don't actually need for every word (fixed linguistic data) or set of propositions (transient linguistic data) to have a stored valuation: it is sufficient for there to be deductive methods for dynamically computing such valuations on stored data.

5.3 Learning

How the adult system of mental representation, called the 'inner model' above, is formed in regards to probabilities? Before we turn to the theory of learning, a word of caution is in order: it is not our goal to supplant the existing theories such as (Tomasello, 2003), especially not when it comes to descriptive detail of child language acquisition. Rather, our goal is to provide *explanatory adequacy* in the sense of Chomsky (1973) who states this quite clearly:

> [t]he fundamental empirical problem of linguistics is to explain how a person can acquire knowledge of language.

To paraphrase (Hertz, Krogh, and Palmer, 1991) whom we quoted in 2.3, we will speak of children, because much of the inspiration for learning comes from developmental psychology, *not* because we are concerned with actual persons as opposed to algorithms. "Brain modeling is a different field and, though we sometimes describe biological analogies, our prime concern is with what the artificial networks can do, and why." When we say we consider something 'innate', what this means is that we assume a learning algorithm that has its search space restricted ab initio. This way, explanatory adequacy

becomes the issue of how an algorithm, as opposed to a person, can acquire human language, and all we can promise is to keep appeals to 'innate' material within the same bounds for algorithms as are routinely assumed for humans.

Recall that the elementary building blocks of `4lang`, the vertices of a graph, correspond to words or morphemes. There is a considerable number, about 10^5, of these, and we add an empty (unlabeled) node $\cdot$. These are connected by three types of directed edges: '0' (is, isa); '1' (subject); and '2' (object). Our theory of types is rather skeletal, especially when compared to what is standard in cognitive linguistics (Jackendoff, 1983) or situation theory (Barwise and Perry, 1983; Devlin, 1991), theories we share with a great deal of motivation, especially in regards to common-sense reasoning about real world situations. When we say that a node is (defeasably) typed as Location or Person, this simply means that a 0-edge runs from the node in question to the `place/1026` or `man/659` node (see 2.1). This applies not just to nodes assumed present already, but also to hypernodes (set of nodes with internal structure, see Definition 5 in 1.5) created during text understanding.

There can be various relations obtaining between objects but, importantly, relations can also hold between things *construed as* objects, such as geometrical points with no atomic content. Consider *the corner of the room is next to the window* – there is no actual physical object 'the corner of the room'. Relational arguments may also include complex motion predicates, as in *flood caused the breaking of the dam*, and so on. To allow for this type-theoretical looseness, arguments of relations will be called *matters*, without any implication that they are material. We use edges of type 1 and 2 to indirectly anchor such higher relations, so the subject of causing will have a 1-edge running from the vertex `cause/3290` to the vertex `flood/85`, and the object, the bursting of the dam, will have a 2-edge running from the `cause/3290` node to the head of the construction where *dam* (not in `4lang`) is subject of `burst/2709`. For ditransitive and higher arity relations, which are tangential to our main topic here, we use decomposition (see 2.4).

In general, we define *valuations* as partial mappings from graphs (both from vertices and from edges) to some small linear order L of scores. There is no analogous 'truth assignment' because in the inner models that are central to the theory, everything is true by virtue of being present. On occasion we may be able to reason based on missing signifiers, the dog that didn't bark, but this is atypical and left for later study. Learning, therefore, requires three kinds of processes: the learning of nodes, the learning of edges, and the learning of valuations. We discuss each in turn.

Learning new vertices We assume a small, inborn set of nodes roughly corresponding to cardinal points of the body schema (Head and Holmes, 1911) and cardinal aspects of the outside world such as the gravity vertical (Campos, Langer, and Krowitz, 1970), to which further nodes are incrementally adjoined (see 3.1). This adjunction typically happens in one shot, a single exposure to a new object like a `boot/413` is sufficient to set up a permanent association between the word and the object, likely including sensory snapshots from smell to texture and a prototypical image (Rosch, 1975). The association is thus between a phonologically marked point, something that, by virtue of being so

marked, is obtained by projecting the entire thought vector on the persistent linguistic subspace L (see 2.3).

As the child is repeatedly exposed to new instances of the category, or even pre-existing instances but seen from a different perspective, against a different background, etc. they gradually obtain a whole set of vectors in L, together forming a point cloud that is generally (but not always, see *radial categories* below) describable by a probability distribution with a single peak, the prototype. This model is very well suited for the Probably Approximately Correct (PAC) theory of learning (Valiant, 1984), and is commonly approximated in machine learning by Gaussian density models. This is not to say that Gaussians are the only plausible model – density estimation offers a rich variety, and remarkably, many of the approaches are directly implementable on artificial neural networks.

On rare occasions, children may learn abstract nodes, such as `color/2207`, based on explicit enumerations 'red isa color, blue isa color,...', but on the whole we don't have much use for post hoc taxonomic categories like *footwear*. Many of these taxonomies are language- and culture-dependent, for example Hungarian has a category *nyílászáró szerkezet* 'closure device' which in English is overtly conjunctive: *doors and windows*. In this particular case, the conjuncts are explicitly nameable, but cognitive semantics considers many other cases that Lakoff (1987) calls *radial* categories, where no single prototype can be identified. Here we illustrate the phenomenon based on (Hanks, 2000), where the homonymy/polysemy distinction is considered from the perspective of the lexicographer, using a standard example:

`bank/227`	`bank/1945`
is an institution	is land
is a large building	is sloping
for storage	is long
for safekeeping	is elevated
of finance/money	situated beside water
carries out transactions	
consists of a staff of people	

Hanks, much as Lakoff and Wittgenstein before him, pays close attention to the fact that radial categories may be explained in terms of a variety of conditions that may or may not be sufficient. The actual `4lang` definitions are more sparse `bank bank argentaria bank 227 u N institution, money in` versus `bank part ripa brzeg 1945 u N land, slope, at river`. We could use defaults to extend this latter definition with `<long>` or perhaps even *elevated*, though we do not at all see how to derive this latter condition for submerged banks such as the famous Dogger Bank.

`bank/227`
`bank/1945`

More important than the details of this particular definition are the fact of common 'metaphorical usage' (snow bank, fog bank, cloud bank) which, many would argue, are present in `bank/1945` as well, with metonymic usage of the institution for the building, or perhaps conversely, using the building metonymically for the institution, as in *The*

Pentagon decided not to deploy more troops. One way or another, this is the key issue for radial categories: surely a large building does not consist of a staff of people.

In a plain intersective theory of word meaning we would simply have a contradiction: as long as `bank/227` is defined as the intersection of the *building* set and the *carries out transactions* set, we obtain as a result the empty set, since buildings don't carry out transactions. We will illustrate how `4lang` solves the problem on the definition of *institution* `inte1zme1ny institutio instytucja 3372 e N organize at, institution work at, has purpose, system, society/2285 has, has long(past), building, people in, conform norm`. This is a lot to unpack, but we concentrate on the seeming contradiction between `system` and `building`. Our understanding of real-life institutions is assumed to be encoded in very high-dimensional thought vectors, and the word *institution* is only the projection of these vectors on the permanent (stable) linguistic subspace L given to us as the eigenspace of the largest eigenvectors (see our discussion of Little (1974) in 2.3). Within L there is a whole subspace S, spanned by vectors (words) related to systems such as *machine, automatism, process, behavior, period, attractor, stability, evolution*, and so on. There is also a subspace B devoted to buildings, spanned by words such as *wall, roof, room, cellar, corridor, brick, mortar, concrete, window, door* and so on and so forth. By accident, there may be some highly abstract words such as *component* that are applicable in both S and B, but we may as well assume that the two subspaces are disjoint. However, thought vectors can have non-zero projection on both of these subspaces at the same time, and our claim is that this is exactly what is going on with `institution`. Since by definition `bank/227 is_a institution`, the word sense `bank/227` just inherits this split without any special provision.

This has nothing to do with the homonymy between `bank/227` and `bank/1945`: we have two disjoint polytopes for these, rather than one polytope with a rich set of projections. There is no notion of 'bank' of which `bank/227` and `bank/1945` could be obtained by projection as there is a single sense of *institution* of which both the building and the system are projections. Importantly, humans perform contextual disambiguation effortlessly: Hanks (2000) makes this point using real life examples

> people without bank accounts; his bank balance; bank charges; gives written notice to the bank; in the event of a bank ceasing to conduct business; high levels of bank deposits; the bank's solvency; a bank's internal audit department; a bank loan; a bank manager; commercial banks; High-Street banks; European and Japanese banks; a granny who tried to rob a bank

on the one hand, and

> the grassy river bank; the northern bank of the Glen water; olive groves and sponge gardens on either bank; generations of farmers built flood banks to create arable land; many people were stranded as the river burst its banks; she slipped down the bank to the water's edge; the high banks towered on either side of us, covered in wild flowers

on the other. Compare this to the case *the bank refused to cash the check*. The victim is typically quite unable to say whether it was the system that is to blame or the staff, actually acting against the system in an arbitrary and capricious manner. There may be some resolution based on a deeper study of financial regulations and the bank's bylaws, but this takes 'slow thinking', what Kahneman (2011) calls 'System 2', as opposed to the 'fast thinking' (System 1) evident in the `227/1945` disambiguation process, and requires access to a great deal of non-linguistic (encyclopedic) knowledge.

In fact, the learning of nouns corresponding to the core case of concrete objects is now solved remarkably well by systems such as YOLO9000 (Redmon et al., 2016) and subsequent work in this direction, lending credence to the insight of Jackendoff (1983) taking "individuated entities within the visual field" as the canonical case for these. Outside this core, the recognition of abstract nouns like *treason* or attitudes like *scornful* are still in their infancy, though sentiment analysis is making remarkable progress.

Learning new edges Again, we assume a small, inborn set of edges (0,1,2), and an inborn mechanism of spreading activation. The canonical edge types are learned by a direct mechanism. Let us return to `boot/413` for a moment and assume a climate/cultural background where the child has already learned `shoe/377` first. Now, seeing the boot on a foot, and having already acquired the notion of `shoe/377`, the child simply adds a '0' edge 'boot isa shoe' i.e. a '0' edge to the graph view of their inner model. In vector semantics, it is the task-specific version of Eq. 2.6 that is added to the system of equations that characterizes the inner model:

$$P_R(t+1) = P_R(t) + s|\texttt{boot}\rangle\langle\texttt{shoe}| \tag{5.6}$$

The case of '1' edges, the separation of subject from predicate, is a bit more complex, especially as two-word utterances are initially used in a variety of functions that the adult grammar will treat by separate construction types such as possessives *dada chair* 'daddy's chair'; spatials *ricky floor* 'Ricky is on the floor'; imperatives *papa pix* 'daddy, fix (this)' and so on. Subjects/subjecthood may not fully emerge until the system of pronouns is firmed up, but our central point here is that what Tomasello (1992) calls "second-order symbols" (for him including not just nominative and accusative linkers but all case markers) are learnable incrementally, on top of the system of what he calls first order symbols (typically, nouns). What is learned by learning verbs is not just some actions, but an entire Fillmorean frame with roles, and markers for these roles. Remarkably, machine learning systems such as Karpathy and Li (2014) are now capable of recognizing and correctly captioning action shots with verbs like *play, eat, jump, throw, hold, sit, stand,...*, see Karpathy's old webpage for some examples.

In 3.1 we speculated that subjects and objects are initially undifferentiated, and it is the same action that we see either performed by the body *John turns* or on something within arms reach *John turns the wheel* for a large class of motion verbs showing intransitive/transitive alternation. But the same incrementality applies to all adpositions/case markers that act as second order entities, e.g. that the dative would indicate the recipient. Let us consider the situations `boot on foot`, which has direct visual support, and

`boot for_ excursion`, which also has strong contextual support, but outside the visual realm.

If the parents are skinheads, the association 'boot for excursion' may never get formed, since the parents wear the boots on all occasions. But if the boots are only worn for excursions (or construction work, or any other specific occasion already identified as such by the child) we will see the *boot* and the *excursion* or *construction work* nodes jointly activated, which will prompt the creation of a new purposive link between the two, just as a joint visual input would trigger the appropriate locative linker.

Again we emphasize that the gradual addition of links described here is not intended as a replacement for actual child language acquisition work such as (Jones, Gobet, and Pine, 2000), but rather as an indication of how such a mechanism, relying on training data of the same sort, can proceed. We note that the ab initio learning of semantic frames (Baker, Ellsworth, and Erk, 2007) is still very hard, but the less ambitious task of semantic role labeling is by now solved remarkably well (Park, 2019).

Learning valuations Probabilities are by no means the only valuation we see as relevant for characterizing human linguistic performance, and using a seven point scale $s = \{0, \ldots, 6\}$ is clearly arbitrary. Be it as it may, similar scales are standardly used in the measurement and modeling of all sorts of psychological attitudes since Osgood, Suci, and Tannenbaum, 1957, and there is an immense wealth of experimental data linking linguistic expressions to valuations. Perhaps the simplest of these would be the GOOD/BAD scale we discussed in 5.2. For improved modeling accuracy, we may want to consider this a three-point scale *good, neutral, bad*, since most things, in and of themselves, are neither particularly good nor particularly bad.

Another valuation of great practical interest would be TRUST. For this we can assume a set of fixed (or slowly changing) sources like people, newspapers, etc., and a set of nonce propositions coming from these. Sometimes the source of a proposition is unclear, but quite often we have information on which proposition comes from which source. By a trusted source we mean one where we positively upgrade our prior on the trustworthiness of the propositions coming from them, and by a distrusted one we mean one that triggers a downgrade (negative upgrade) in the trustworthiness of the proposition. As (dis)confirmation about particular propositions comes in, we can gradually improve our model of sources in the obvious manner, by backpropagating the confirmation values to them. This can be formulated in a continuous model using probabilities, but the essence of the analysis can be captured in terms of discrete likeliness just as well.

Of particular technical interest is the ACTION POTENTIAL valuation taking values in $A = \{-1, 0, 1, 2\}$, where -1 means 'blocked' or 'refractory', 0 means 'inactive', 1 means 'active', and 2 means 'spreading'. These can be used to keep track of the currently active part of the graph and implement what we take to be the core cognitive process, spreading activation (Quillian, 1969; Nemeskey et al., 2013). Here we will not pursue this development (see 7.4 for further details), but note that we don't see this valuation as formally different from e.g. the probability valuation, except that innateness is plausible for the

former but not the latter. To paraphrase Dedekind's famous quip, spreading activation was created by God, the other valuations are culturally learned.

Unlike the lexicon itself, valuations are not permanent. The inputs to a valuation are typically nonce hypernodes 'death at Snowbird' and the linguistic subspace L only serves as a basis for computing the mapping from the hypernodes in question to the scale s. We assume that the activation mechanism is unlearned (innate), but this still leaves open the question of how we know that forces of nature are a likely cause of death in Reykjavík but not in Istanbul? Surely this knowledge is not innate, and most of us have not studied mortality tables and statistics at this level of specificity, yet the broad conclusion, that death by natural forces is more likely in Reykjavík than in Istanbul, is present in rational thinking at the very least in a defeasible form (we will revise our naive notions if confronted with strong statistical evidence to the contrary).

Part of the answer was already provided in 5.2, where we described the mechanism to compute these values. Aside from very special cases, we assume that such valuations are always computed afresh, rather than stored. What is stored are simpler building blocks, such as 'volcano near Reykjavík', 'volcano isa danger' from which we can easily obtain 'danger near Reykjavík'. A great deal of background information, such that *danger* is connected to *death*, must be pulled in to compute the kind of valuations we described in Table 5.1, but this does not alter the main point we are making here, that inner models are small information objects (the entire mental lexicon is estimated to be about 1.5MB, see Mollica and Piantadosi, 2019).

From the foregoing the reader may have gathered the impression that learning of nodes is relatively easy, learning of edges is harder, and learning of valuations is the hardest, something doable only after the nodes and edges are already in place. The actual situation is a bit more complex: any survey of the lexicon will unearth nodes that are learnable only with the aid of valuations. The strict behaviorist position that learning is simply a matter of stimulus-response conditioning has been largely abandoned since Chomsky, 1959. Whether the alternative spelled out by Chomsky, an innate Universal Grammar (UG) makes more sense is a debate we need not enter here beyond noting the obvious, that lexical entries are predominantly language-particular. This is no doubt the main reason why Chomsky places the lexicon in the "marked periphery", outside "core grammar".

Children acquiring a language acquire its lexicon, and there is no reason to believe that this process relies on innate knowledge of concepts (nodes) for the most part. In keeping with our approach to consider the entire lexicon, we begin with a brief survey of the semantic fields used by Buck, 1949:

1. Physical World
2. Mankind
3. Animals
4. Body Parts and Functions
5. Food and Drink
6. Clothing and Adornment
7. Dwellings and Furniture
8. Agriculture and Vegetation
9. Physical Acts and Materials
10. Motion and Transportation
11. Possession and Trade
12. Spatial Relations
13. Quantity and Number
14. Time
15. Sense Perception
16. Emotion
17. Mind and Thought
18. Language and Music
19. Social Relations
20. Warfare and Hunting
21. Law and Judgment
22. Religion and Beliefs

The list offers whole semantic fields like 4, 12, 14, and 15, where we have argued (see 3.1) that the best way to make sense of the data is by reference to embodied cognition, a theory that comes very close to UG in its insistence of there being an obviously genetically determined component of the explanation. The same approach can be extended to several other semantic fields: we discuss this on 6: Clothing, Personal Adornment, and Care.

We start from an embodied portion, 4, and proceed by defining *shoe* as 'clothing, worn on foot'; *leggings* as 'clothing, worn on legs'; *shirt* as 'clothing, worn on trunk'; etc. We begin by noting that *clothing* 'the things that people wear to cover their body or keep warm' is already available in `4lang` as `cloth, on body, human has body, cause_ body[warm]`. Using this, a good number of Buck's keywords fit this scheme: 6.11 *clothe, dress*; 6.12 *clothing, clothes*; 6.21 *cloth*; 6.41 *cloak*; 6.412 *overcoat*; 6.42 *woman's dress*; 6.43 *coat*; 6.44 *shirt*; 6.45 *collar*; 6.46 *skirt*; 6.47 *apron*; 6.48 *trousers*; 6.49 *stocking, sock*; 6.51 *shoe*; 6.52 *boot*; 6.53 *slipper*; 6.55 *hat, cap*; 6.58 *glove*; and 6.59 *veil*. For some of these case our definition of clothing would need a bugfix to include the 'modesty' aspect (which is actually culture-specific) by merging in our definition of *cover* `=agt on =pat, protect, cause_[lack{gen see =pat}]` to yield an additional clause e.g. for *veil* `cause_ [lack{gen see face}]`.

cover

This analysis illustrates the point about highly abstract units we made in 1.2: obviously *boot* means different things for different cultures, and the Roman legionnaire would not necessarily recognize the caligae in the skinhead's DMs. But the conceptual relatedness is clearly there, and as we discussed above, the word can be learned as a node in a network composed of abstract units such as `cover` and *foot* `organ, leg has, at ground` which we need anyway.

foot

This is not to say that the 54 main headings covered in Chapter 6 of Buck, 1949 are all automatically covered in `4lang`, especially as many are listed in this chapter only because the lexicographer had to put them somewhere, and this seemed the best place. In addition to the core entries discussed so far, we have a wide variety of clothing materials: 6.22 *wool*; 6.23 *linen, flax*; 6.24 *cotton*; 6.25 *silk*; 6.26 *lace*; 6.27 *felt*;

wool 6.28 *fur*; 6.29 *leather*. For the most part we treat these as genus `material`, e.g. *wool* `material, soft, sheep has`, but sometimes we place them under other genera.
fur e.g. *fur* `hair/3359, cover skin, mammal has`.

Buck also lists here some professions (6.13 *tailor*; 6.54 *shoemaker, cobbler*); activities characteristic of cloth- and clothing-making (6.31 *spin*; 6.33 *weave*; 6.35 *sew*; 6.39 *dye (vb.)*); professional tools (6.34 *loom*; 6.32 *spindle*; 6.36 *needle*; 6.37 *awl*; 6.38 *thread*). We have discussed professions like *cook* in 2.2, and for a typical tool we of-
needle fer *needle* `artifact, long, thin/2598, steel, pierce, has hole, <sew ins_>`.

More challenging are the 'accessories' or 'adornments' which are not, strictly speaking, items of clothing in and of themselves (6.57 *belt, girdle*; 6.61 *pocket*; 6.62 *button*; 6.63 *pin*; 6.71 *adornment (personal)*; 6.72 *jewel*; 6.73 *ring (for finger)*; 6.74 *bracelet*; 6.75 *necklace*; 6.81 *handkerchief*) as well as culture-specific items that are associated to clothing and adornment only vaguely (6.82 *towel*; 6.83 *napkin*; 6.91 *comb*; 6.92 *brush*; 6.93 *razor*; 6.94 *ointment*; *6.95 soap*; 6.96 *mirror*). First, we need to consider what is an *accessory* 'something such as a bag, belt, or jewellery that you wear or carry because it is attractive'. This is easily formulated in `4lang` as `person wear, attract`. Similarly with *adornment* 'make something look more attractive by putting something pretty
attract on it'. The key idea is to define *attract* as `=agt cause_ {=pat want {=pat near =agt}}`. Once this is done we are free to leave it to non-linguistic (culturally or genetically defined) mechanisms to guarantee that nice-smelling ointments and pretty jewelry will be attractive. The example highlights the need for a realistic theory of acquiring highly abstract concepts. In S19:2 we wrote:

> the pattern matching skill deployed during the acquisition of those words denoting natural kinds cannot account for the entirety of concept formation. People know exactly what it means to betray someone or something, yet it is unlikely in the extreme that parents tell their children "here is an excellent case of betrayal, here is another one". Studies of children's acquisition of lexical entries such as McKeown and Curtis (1987) have made it clear that natural kinds, however generously defined so as to include cultural kinds and artifacts, make up only a small fraction of the vocabulary learned, even at an early age, and that children's acquisition of abstract items "but not concrete word learning, appears to occur in parallel with the major advances in social cognition" (Bergelson and Swingley, 2013).

While our remarks on the subject must remain somewhat speculative, it seems clear that *attract* is learned together with *attraction, attracting, attractive* i.e. without special reference to the fact that the root happens to be verbal. In fact, there is every reason to suppose that abstract terms are root-like, and it is only the syntax that imposes lexical category
responsible on them. Consider *responsible* `has control, has authority, has blame`. The Hungarian version proceeds from a verb *felel* 'respond' through an adjective *felelős* 'is responsible' to a noun *felelősség* 'responsibility'. In Chinese, we begin with a noun

ze2ren4 责任, form a verb *fu ze2ren4* 负责任, and proceed to the adjectival *fu4 ze2ren4 de* 负责任的.

We also have roots that are neutral between verbal and adjectival forms, for example *open, cool, warm*. Common to them is the ability to treat the adjective as the result state of the verb, for example *open* `move[can/1246], move through, lack shut/2668` and `after(=pat open/1814)`; *cool* `temperature, normal er_, er_ cold` and `after(cold)`; *warm* `temperature(er_ gen)` and `after(warm/1655)`. The nominal forms *warm/warmth, join/joint, cool/cold, heat/hot, ...* are remarkably close, but perhaps not close enough without making recourse to the kind of stratal morphology that makes a strong distinction between stem-level and word-level morphology (Kiparsky, 2016).

`open/1814`
`open/1815`
`cool/1103`
`cool/1101`
`warm/1655`
`warm/878`

The use of `=agt` and `=pat` in the definition of *attract* makes clear that it is essential to have two items in attraction, i.e. the relation is binary. Our theory of learning must start with an elementary act of recognizing attraction, just as we recognize nearness, one thing being on top of another, and a host of other relations. Clearly, there is huge evolutionary advantage to recognizing nearness to us, as this will be a primary signal of whether something can attack us and/or whether we can manipulate it to our advantage. We consider proximity marking (`near`, see 3.1) to be a reasonable candidate for universality.

We also consider the other two components of our definition, a naive theory of needs and wants (6.2), and a naive theory of causation (2.4) as evolutionarily highly motivated: clearly, being able to model what other actors in the environment are likely to do, based on their needs/wants, will hugely improve our own chances for surviving and thriving, and a theory, even a naive theory, of causation has a similar salutary effect. The challenge here is to put together three highly abstract theories to produce a fourth one for *attraction*.

While our remarks must remain somewhat speculative, what we believe is that the putting together is driven by valuations. We recognize attraction by first seeing increasing nearness, `after(=agt nearer =pat)`, next by attributing this change to the desire of the patient, `=pat want {=pat near =agt}`, and finally the desire itself as being aroused by the property of attractiveness lodged in the agent. At some point, the `after` clause is converted to the `cause_` by means of the naive analysis of causation.

The micro-analysis of how these steps are built on one another during language acquisition requires further study, and will clearly involve polarity: nearness is GOOD for good things, but BAD for bad things. Since we expect all beings, not just `self`, to want good things, we must assume that attracting something is itself a good thing (as it is, unless `=pat` is bad). In certain cases, we can expect these valuations to play out on primary linguistic and sensory data: to the extent something is pleasant to the touch a baby may want to touch it and thereby bringing it near, or even put it in their mouth, to bring it even nearer. Signaling of attractiveness is already evident in flowers, and signaling of badness, aposematism, is common in the animal kingdom. Where earlier generations of researchers may have seen the theory of causation as something only the properly trained

mind is capable of (Kant even assumed some innate human capacity), we see a perfect Darwinian continuity connecting humans to far simpler life forms.