Simulating Sense Extension

Rowan Hall Maudslay[1,2] **Simone Teufel**[2]
[1]École Nationale des Chartes | Université Paris Sciences et Lettres (PSL)
[2]Dept. of Computer Science and Technology | University of Cambridge
rowan.maudslay@psl.eu sht25@cam.ac.uk

Abstract

Humans have an uncanny ability to push words beyond their limits. This ability manifests in many phenomena, including metaphor, metonymy, semantic drift, slang, jargon, conversion, and overextension. Generative models of these phenomena are uncommon, because there is no robust methodology that can be used to train and evaluate models of this nature. To address this, we introduce a new task, *novel sense formulation*, in which a model is exposed to a multimodal representation of an unseen concept and must use an existing word creatively to describe it. We create seven datasets corresponding to the phenomena above, and evaluate a perceptron, a transformer, and an influential cognitive model. For most phenomena, a multimodal variant of the perceptron performed best. The cognitive model underperformed, suggesting that its description of the mechanism behind sense extension is incomplete.

1 Introduction

There are many situations in which a person encounters a new concept, and extends the semantics of a word to describe it. For example, a scientist might observe a new phenomenon and characterise it using metaphor (Hofstadter and Sander, 2013; Ortony, 1975), as was the case with the *planetary* model of the atom (Perrin, 1901) and with Darwin's (1859) evolutionary *tree*. Children regularly perform similar feats of abstraction when they overextend words to compensate for gaps in their vocabularies (Clark, 1973). For instance, a child may call the cover of a jar a *door* if they do not yet know the word *lid*. In each of these examples, a word is used to refer to a concept with which it is not conventionally associated, but which is semantically related to its existing senses. This phenomenon is known as *sense extension* (Pustejovsky, 1995; Briscoe and Copestake, 1991; Copestake and Briscoe, 1995).

Sense extension allows humans to generalise (Rossen Skadhauge, 1995; Pustejovsky and Rumshisky, 2010) and to apply familiar knowledge in unfamiliar settings. Despite significant interest in generalisation in both cognitive science (e.g. Tenenbaum et al., 2011; Lake et al., 2017) and AI research (e.g. Ito et al., 2022; Freiesleben and Grote, 2023), there is currently no standard methodology that is used to train or evaluate computational models of sense extension. For some types of sense extension, such as slang, bespoke task formulations have been proposed (e.g. Sun et al., 2021; Pinto and Xu, 2021; Yu, 2023). However, the majority of these task formalisations do not support multimodal modelling, and many suffer from methodological flaws. Moreover, the wide variety of task formalisms makes it difficult to evaluate a single model across phenomena.

To address this, we introduce a new task called *novel sense formulation*, in which the goal is to select the best word to describe an unseen concept. In this task, a model is trained using an incomplete representation of the world: a pruned universe, or *pruniverse*. Concepts in the pruniverse are encoded using multimodal data. During evaluation, the model is exposed to concepts that were not included in its pruniverse, and must perform sense extension to describe them. In the real world, these concepts are already associated with words, but in the simulation they are not. We construct seven different pruniverses, each designed to evaluate a different type of sense extension.

Our task makes it possible to investigate multimodal sense extension across phenomena. In our experiments, we evaluate three different models which exploit four different representational modalities. We find that a multimodal model performs best for all types of sense extension save one. An influential cognitive model was outper-

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1826–1850, 2026. https://doi.org/10.1162/TACL.a.780
Action Editor: Dilek Hakkani-Tur. Submission batch: 7/2025; Revision batch: 12/2025; Published 8/2026.

formed substantially by a simple perceptron, suggesting that the cognitive model does not adequately explain how humans perform sense extension. This view is supported by an additional post-hoc experiment, in which we investigate whether the models exhibit prototype effects.

2 Sense Extension in Word Production

The process of converting a concept into a linguistic signal is called *production*; the inverse process is called *comprehension* (Clark and Clark, 1977). Word production is commonly theorised to involve three independent stages: conceptualisation, formulation, and articulation (Fromkin, 1971; Garrett, 1975; Dell, 1986; Levelt, 1989). In *conceptualisation*, a speaker determines the concept that they wish to express; in *formulation*, this concept is translated into a wordform; and finally in *articulation* this wordform is converted into muscle movements (e.g. of the tongue). In our work, we build models of formulation.

During production, a speaker will sometimes be faced with a *lexical gap* (e.g. Myers-Scotton and Jake, 1995), which is an absence of any word for a particular concept (Lehrer, 1974). Lexical gaps can be filled using existing words via sense extension (Rabagliati et al., 2010; Ortony, 1975). A sense extension which addresses a lexical gap is sometimes called a *catachresis* (Black, 1962).

In the present work, we build models that emulate sense extension. We focus on seven types of word production which involve catachresis. We do not claim that this collection is definitive, nor that these phenomena are independent.

Metaphor A *metaphor* is a piece of language in which one entity is viewed as another (Lakoff and Johnson, 1980; Lakoff, 1993). For example, the word *attack* can be used metaphorically to refer to intense criticism (e.g. "they *attack* the policy"). In a metaphorical extension, some semantic features of the literal sense are kept, whilst others are discarded (Hofstadter, 1985; Veale and Hao, 2008): like a literal *attack*, a metaphorical *attack* is an act of hostility between two parties, but unlike a literal *attack*, no physical aggression is involved.

Metonymy A *metonym* is a piece of language in which one wordform stands in for another (Fass, 1988). An example of metonymy is when a place name is used to refer to a government situated in that place (e.g. "*Beijing* announced tariffs"). One factor that differentiates metaphor and metonymy is that metaphor involves resemblance, whereas metonymy does not (Gibbs, 1994): while a verbal attack is "like" a physical attack, the Chinese government is not "like" Beijing. Metonymic sense extensions often adhere to productive schemas (Pustejovsky, 1995). For example, many words for containers (e.g. *bottle*, *box*, *punnet*, etc.) can be used metonymically to refer to the contents of such containers (e.g. "I drank a *bottle*").

Semantic Drift Within a language community the semantics of words can change over time. This process is known as *semantic drift* (Bybee, 2015). For example, the word *follower* is used to refer to someone who espouses another's ideas, but since the advent of social media it has also been used to refer to an online subscriber. Two common causes of semantic drift are metaphor and metonymy (Bloomfield, 1933; Ullmann, 1962; Blank, 1999).

Jargon Experts use specialist terminology to be precise when conferring with other experts (Cabré, 1999). This terminology is commonly called *jargon* (e.g. Wilkinson, 1992). For example, rock climbing practitioners use words like *T-stack* or *chickenwings* to refer to specific climbing moves (Whittaker, 2020). Jargon facilitates efficient in-group communication, but it can also hinder understanding for out-groups (Martín Rojo, 2010).

Slang Informal non-technical language is called *slang* (Lighter, 2001). Some slang involves *neologism*, which is the creation of a new wordform such as *hangxiety*, a portmanteau that combines *hangover* and *anxiety*. Many slang terms, however, are sense extensions (Sun et al., 2024). For example, in British English slang the verb *gas* refers to the bestowal of excessive praise which unduly inflates someone's ego. Like jargon, slang can be used to signal group membership (Drake, 1980) or for deliberate obfuscation (Green, 2016).

Conversion *Conversion* is the process through which a word is created with a different part of speech (POS) to an existing word, but the same surface form (Sweet, 1891). For example, the verb *drink* ("to *drink*") and the noun *drink* ("a *drink*") are related by conversion. Like metonymy, conversion is highly productive (Marchand, 1969).

Overextension In language acquisition, between the ages of 12 and 30 months, children commonly use words in unconventional contexts to re-

fer to concepts whose actual names they do not know (Clark, 1973). This phenomenon is known as *overextension* (Rescorla, 1980). For example, a child might use the word *dog* to refer to any four-legged animal (Bloom, 1973). Overextension is an example of *assimilation*, which is the process through which children integrate new information into existing cognitive schemas (Piaget, 1952).

3 Previous Computational Treatments

Several computational analyses of sense extension patterns have been performed (e.g. Dang et al., 1998; Rossen Skadhauge, 1995; Boleda et al., 2012). Attempts to produce predictive models of sense extension, however, are less common. The most common modelling task in this area is word sense disambiguation (WSD). The goal in WSD is to determine which sense of a word is evoked in a sentence, based on the options in a computational lexicon such as WordNet (Miller, 1995; Fellbaum, 1998). In WSD, only conventional senses are treated. The focus of our work is on novel senses.

Several alternative tasks exist in which the goal is to identify novel senses. These tasks include metaphor detection (e.g. Leong et al., 2018), semantic change detection (e.g. Tahmasebi et al., 2021), slang detection (Pei et al., 2019), and unknown sense detection (Erk, 2006). Also related is word sense induction (WSI), in which the goal is to extract a sense inventory from a corpus; Lau et al. (2012) have used WSI models to identify novel senses. In these tasks, a model must identify whether or not a sense is novel, but it does not need to make any prediction about what these senses mean or how they arose.[1] Our interest is in models which make predictions of this nature.

To study the interpretation of novel senses, a variety of figurative language interpretation tasks have been proposed, based on paraphrasing (Shutova et al., 2013; Tong et al., 2024) or textual entailment (Saakyan et al., 2022). A major limitation of these tasks is that the training material is not controlled, and may include the sense extensions which are used in evaluation. If a model performs well, it may only have done so because it was shown the paraphrase or entailed information during training. It is not necessarily the case that the model will be able to interpret sense extensions which it had not previously encountered.

[1] Some of the aforementioned tasks involve not only novel senses, but also conventionalised sense extensions.

An alternative task which addresses this limitation is word sense extension (WSE; Yu and Xu, 2023). WSE formulates sense extension as a language modelling task. In WSE, models are trained on a corpus that has been modified to have certain senses replaced with artificial tokens. The goal is to paraphrase these tokens back to their original forms, thereby recreating sense extensions which were concealed during training. This methodology ensures that models are correctly evaluated on unseen examples. In this respect, WSE can be framed as a historical reconstruction task, a method that has previously been used for drug discovery (e.g. Swanson and Smalheiser, 1997), taxonomy enrichment (e.g. Nikishina et al., 2020), and citation prediction (e.g. Yu et al., 2012). In our work, we use a similar methodology to ensure that models are evaluated on unseen examples.

The design of WSE has several limitations. The decision to treat sense extension through language modelling makes it difficult to diagnose model failures: if a WSE model performs poorly, it is not clear if this performance deficit reflects an inability to perform sense extension, or a more general inability to understand sentences. Moreover, sense extension is theorised to involve interactions between different representational modalities (Aldrich, 1968; Forceville and Urios-Aparisi, 2009), but WSE does not support multimodal modelling. By contrast, the task we design is multimodal, and better isolates sense extension from other aspects of language understanding.

WSE does not evaluate a specific kind of sense extension, but Yu (2023) has modified this task to treat figurative language. Other computational tasks have been designed to study slang (Sun et al., 2021), semantic drift (Habibi et al., 2020; Grewal and Xu, 2021; Yu and Xu, 2021), and conversion (Yu and Xu, 2022). Like WSE, the majority of these tasks do not support multimodal modelling. Moreover, the creation of unique task formulations for each type of sense extension makes it difficult to evaluate models across phenomena. In our work, we create a standard task formulation which can be used to treat many types of sense extension.

One existing multimodal treatment of sense extension was performed by Pinto and Xu (2021), who built models of overextension using resources linked to WordNet. Another multimodal method has been proposed by Brochhagen et al. (2023), who treat multiple phenomena (semantic drift and

overextension). In this task, the input is a pair of concepts, and models predict whether these concepts are expressed using the same wordform; concepts of this nature are said to *colexify* (François, 2008). Our goal, by contrast, is to evaluate models which produce new sense extensions.

4 New Tasks for Sense Extension

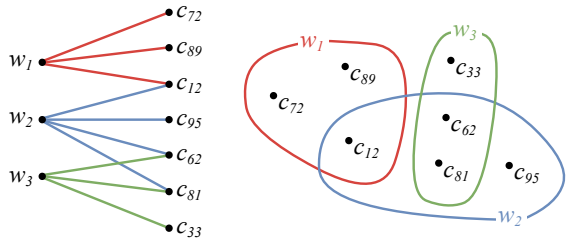
We introduce a pair of task formulations that can be used to evaluate multimodal models for any kind of sense extension. To define these tasks, let $\mathcal{W}$ be the set of all wordforms in a language, and $\mathcal{C}$ be the set of all the concepts which are lexicalised. The distinction between wordforms and concepts corresponds to Saussure's (1916) notion of signifiers and signifieds. We formalise a sense as a tuple $\langle w, c \rangle \in \mathcal{S}$ consisting of a wordform $w \in \mathcal{W}$ and a concept $c \in \mathcal{C}$. Concepts are commonly described in lexica using definitions, but they can be represented by other data modalities, such as images.

4.1 Novel Sense Formulation

Novel sense formulation (NSF) mimics the challenge faced by a speaker who needs to extend the semantics of a word in order to refer to a new concept. In NSF, the input is a concept c which does not appear in the training lexicon. The output is a wordform $w \in \mathcal{W}$ which can be used to refer to this concept. NSF is achieved using a model of $p(w \mid c)$. In generating w for c, the model implicitly creates a novel sense, $\langle w, c \rangle$.

NSF can be framed as a graph enrichment task. This can be achieved by construing the lexicon as a bipartite incidence graph, with disjoint vertex sets $\mathcal{W}$ and $\mathcal{C}$. This graph is illustrated in Figure 1a. Unlabelled edges connect wordforms $w \in \mathcal{W}$ to concepts $c \in \mathcal{C}$. Each edge corresponds to a word sense, $\langle w, c \rangle \in \mathcal{S}$. Many-to-one mappings between wordforms and concepts correspond to examples of synonymy, and one-to-many mappings correspond to examples of polysemy. In NSF, a new concept c is added to this graph. A model must generate a new edge $\langle w, c \rangle$ which connects this concept to a suitable wordform w.

NSF can alternatively be framed as a cluster assignment problem. To do this, the lexicon can be construed as a hypergraph. A *hypergraph* is a generalisation of a graph in which edges can connect to more than two vertices. Figure 1b shows a hypergraph that is equivalent to the bipartite graph in



(a) A bipartite graph (b) A hypergraph

Figure 1: Two formulations of the lexicon

Figure 1a. Adding a new edge $\langle w, c \rangle$ to Figure 1a is equivalent to assigning c to the cluster of concepts denoted by w in Figure 1b. Under this view, the goal in NSF is to choose the cluster to which a new concept c should be assigned.

NSF is based on two key assumptions. The first assumption is that concepts can be represented as discrete entities. This assumption is reflected in the formulations of the lexicon illustrated in Figure 1. The second assumption is that the process of concept creation is separate from the process of concept naming. This assumption is discussed further in §10 (under "Predefined Concepts").

The task formulations in the literature that are most similar to NSF are those by Pinto and Xu (2021) and Sun et al. (2021). To treat overextension, Pinto and Xu construct models of $p(c' \mid c)$, which determine whether a new concept c could colexify with a familiar concept c'. These models are used to treat $p(w \mid c)$ by first finding the c' which best colexifies with c, and then using a lookup to find the word used for c'. This approach is only possible because Pinto and Xu make the simplifying assumption that there is a one-to-one correspondence between wordforms and concepts prior to sense extension. Whether or not this assumption holds for children is debatable, but in an adult's lexicon the relationship is certainly many-to-many (see Figure 1a). Unlike Pinto and Xu's task, NSF supports arbitrary lexical structure prior to sense extension.

To treat slang, on the other hand, Sun et al. (2021) construct models of $p(w \mid c, \mathbf{w})$, where $\mathbf{w}$ is a sentence with a masked token in the location of w. Each concept c is represented using a gloss from a slang dictionary, such as Green's Dictionary of Slang. This use of glosses is problematic, because a gloss sometimes includes the word with which it is associated. For example, consider this sense of *drink* from Green's Dictionary of Slang:

> a bribe, a sum of money that would supposedly purchase 'a drink'

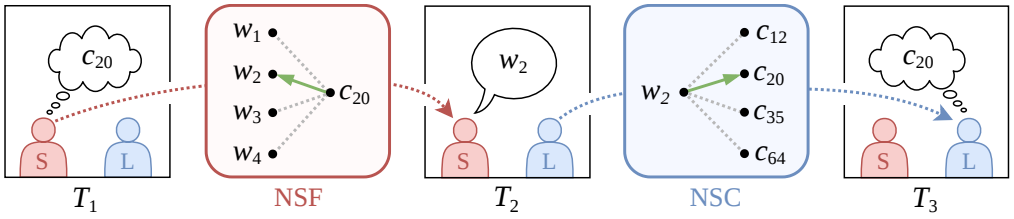


Figure 2: A new thought travels between minds

This gloss includes an explanation about why the word *drink* is used for this concept. In Sun et al.'s work, a model needs to determine which word goes with each gloss, but glosses like this one contain the solutions. Sun et al.'s use of glosses therefore introduces circularity. In NSF, we avoid circularity by creating concept representations using non-textual data, described in §6. We additionally do not provide a linguistic context (**w**) as input, and instead integrate information about the usage of a concept (e.g. syntactic constraints) into the concept representations themselves.

4.2 Novel Sense Comprehension

It is possible to define the reverse task of NSF, which we call *novel sense comprehension* (NSC). NSC mimics the challenge facing a listener who needs to interpret the sense of a word that has been used in a new context. We define this task below, but we do not treat NSC in this work.

The input for NSC is a wordform w, and the output is the concept c to which w refers. The output concept is selected from a set of options $\mathcal{C}'_w \subset \mathcal{C}$. The word w does not have a previous association with any of these options (i.e. $\forall c \in \mathcal{C}'_w : \langle w, c \rangle \notin \mathcal{S}$). While NSF is achieved using a model of $p(w \mid c)$, NSC is achieved using a model of $p(c \mid w)$.

The relationship between NSF and NSC is illustrated by the production–comprehension paradigm in Figure 2. On the left, a speaker (S) wants to refer to a new concept (at timestamp T_1), so chooses the best word for it (NSF) and produces that word (T_2). On the right, a listener (L) hears the word just uttered by the speaker (T_2), chooses which concept it refers to (NSC), and therefore interprets the sense that the speaker just created (T_3).

In NSF, the output is selected from all wordforms $\mathcal{W}$, while in NSC the output is selected from a small set of concepts, $\mathcal{C}'_w$. This design choice reflects an asymmetry facing speakers and listeners in the real world. When naming a concept, speakers choose from any wordform in their vocabulary. When interpreting which concept a word refers to, however, listeners choose only from options available in the context of an utterance.

To evaluate NSC models, we need to construct a set of option concepts $\mathcal{C}'_w$ for every word w, consisting of concepts appearing in contexts where w could be used. It is not obvious how these sets of concepts could be created. For this reason, we leave NSC to future work.

5 Dataset Construction

To evaluate NSF, examples of novel senses $\langle w, c \rangle$ are needed. The performance of a model can be measured by eliciting wordform predictions for c, and comparing these predictions to w.

Lexica only contain senses which are widely used. However, many conventional senses were once creative sense extensions (Bowdle and Gentner, 2005). For example, the sense *neck*$_4$ in WordNet (defined as "a narrow part of an artifact that resembles a neck") was once a novel extension of *neck*$_2$ ("the part of an organism that connects the head to the body"). Lexica can therefore be used to simulate NSF using historical reconstruction.

More specifically, instead of training models using an entire lexicon $\mathcal{S}$, models can instead be trained using an incomplete lexicon. We call this incomplete lexicon a pruned universe, or *pruniverse*. Let $\langle w, c \rangle \in \mathcal{P}$ be the senses in a particular pruniverse, which are a subset of the senses in a lexicon, $\mathcal{P} \subset \mathcal{S}$. Additionally, let $\langle w, c \rangle \in \mathcal{U}$ be unseen senses which were included in the original lexicon but excluded from the pruniverse: $\mathcal{U} \subseteq \mathcal{S} \setminus \mathcal{P}$. Models can be trained using senses in $\mathcal{P}$ and then evaluated using unseen senses in $\mathcal{U}$. With pruniverses, any conventional sense can be treated as though it were novel.

An illustration of a pruniverse is given in Figure 3. This pruniverse was created by ablating the concept c_{72} from the lexicon in Figure 1. To reintegrate this concept, there are three words which an NSF model must choose between (w_1, w_2 and w_3). The assignment of c_{72} to w_1 corresponds to the regeneration of the pruned sense $\langle w_1, c_{72} \rangle$.

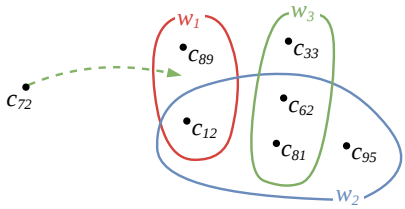


Figure 3: Example of a pruniverse

We construct seven pruniverses, each designed to isolate a different type of sense extension. Each pruniverse is created by first identifying a set of senses in WordNet which exemplify a particular phenomenon (e.g. metaphor). Unless mentioned otherwise, the unseen set for that phenomenon is created by randomly sampling 10% of these senses, and the corresponding pruniverse is created by removing these unseen senses from WordNet. 90% of the target senses remain in the pruniverse; we refer to these senses as illustrative senses $\langle w, c \rangle \in \mathcal{I}$, where $\mathcal{I} \subset \mathcal{P}$. It is likely that these senses will be useful for training models.

The methodology we use to construct pruniverses is similar to a methodology used in WSE, where training data is created by removing occurrences of certain senses from a sense-annotated corpus (Yu and Xu, 2023; see §3). In WSE, models are trained on sentences from this modified corpus, but in NSF, models are trained directly on concept representations. Moreover, while Yu and Xu select unseen senses randomly, we select unseen senses which exemplify certain phenomena.

Statistics for the pruniverses are shown in Table 1.[2] On the right of Table 1 there is an example of an unseen concept for each phenomenon, with abbreviated glosses used to represent concepts. Details about how unseen senses are identified for each phenomenon are given below, with additional information in App. A.

Metaphor Metaphorical senses in WordNet are identified using ChainNet (Maudslay et al., 2024), a dataset that identifies metaphorical and metonymic connections between WordNet senses. We retrieve every sense in ChainNet that is derived from another sense via a metaphor edge.

Metonymy Metonymy data is constructed using the same procedure as for metaphor, based on metonymy edges in ChainNet.

Semantic Drift For semantic drift, an unseen set is constructed consisting of nominal senses which arose after a certain cutoff in time. The cutoff we chose is 1960, which is 25 years before WordNet started development. WordNet does not include obsolete senses, and therefore cannot be used to study lexical developments that precede Modern English. The age of senses is identified using the Oxford English Dictionary (OED), based on the WordNet–OED alignment of Maudslay and Teufel (2022). The illustrative set consists of all the senses that arose between 1925 and 1960. The cutoff of 1925 is chosen because it results in an illustrative set which is comparable in size to several other illustrative sets. Choosing an earlier year would yield a larger set, but it is likely that these older senses would be less useful for model training than senses which arose close to 1960.

Jargon For jargon, we identify every sense related to linguistics, computer science, politics, economics, and psychology, according to the WordNet Domains Hierarchy (Bentivogli et al., 2004). We chose these domains because they consist of a large number of senses, and because many of these senses are based on sense extension rather than neologism.

Slang We construct a set of slang senses by identifying every WordNet sense whose equivalent sense in the OED is labelled as colloquial.

Conversion We identify every nominal sense in WordNet that is a conversion of a verbal sense using WordNet morphology annotation produced by Fellbaum and Miller (2003).

Overextension For unseen overextensions, we use the data of Pinto and Xu (2021), who manually aligned examples of overextension with concepts in WordNet. Unlike for the other phenomena, these unseen senses do not appear in WordNet (i.e. $\mathcal{U} \cap \mathcal{S} = \emptyset$).[3] The overextension pruniverse should correspond to the lexicon of a child, who will know a small fraction of the senses in a language. Accurate information about a child's lexicon is impossible to access. To estimate a child's vocabulary, we follow Pinto and Xu, and extract every wordform from Wordbank (Frank et al., 2017), a dataset of age-of-acquisition norms

[2] While some of these datasets seem small by the standards of deep learning, each sense is itself associated with many datapoints (e.g. images in ImageNet or tokens in SemCor).

[3] Although each unseen overextension is not a sense in WordNet, it nevertheless consists of a wordform w and a concept c which appear in WordNet separately: $\mathcal{U} \subset \mathcal{W} \times \mathcal{C}$.

Pruniverse	$\lvert\mathcal{P}\rvert$	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{U}\rvert$	Example Unseen Sense $\langle w, c\rangle \in \mathcal{U}$
Metaphor	131,564	6,565	748	⟨*neck*, "a narrow part of an artifact that resembles a neck"⟩
Metonymy	131,872	5,427	611	⟨*dish*, "the quantity that a dish will hold"⟩
Semantic Drift	132,113	2,408	257	⟨*cable*, "television that is transmitted over cable"⟩
Jargon	132,491	4,043	449	⟨*worm*, "a software program capable of reproducing itself"⟩
Slang	132,688	2,270	252	⟨*bomb*, "an event that fails badly"⟩
Conversion	132,331	2,323	265	⟨*leap*, "a self-propelled movement upwards or forwards"⟩
Overextension	5,259		228	⟨*ball*, "the bulb of an onion plant"⟩

Table 1: Pruniverse statistics and examples

for children up to 30 months old. We then construct a pruniverse consisting of every sense of these words. A child is unlikely to know every sense of the words they produce, so this pruniverse is not a realistic approximation of a child's lexicon, but we are not aware of any other method which is better. Children do not encounter positive examples of overextension, so there are no illustrative senses in this pruniverse.

6 Concept Representations

To produce models of NSF, we need a way to represent the semantics of the concepts in WordNet. One way that this can be achieved is by constructing a concept embedding space, in which a vector is provided for each $c \in \mathcal{C}$.[4]

Concept embeddings are commonly produced by computing sentence embeddings of each concept's gloss (e.g. Chen et al., 2015; Oele and van Noord, 2018; Huang et al., 2019; Blevins and Zettlemoyer, 2020). This method has been employed in previous work on sense extension (e.g. Sun et al., 2021). However, as discussed in §4.1, there is a danger that this method can introduce circularity. In NSF, the goal is to reproduce the word associated with an unseen concept, but sometimes the gloss of that concept will include the word which is used to refer to it. For example, in WordNet the definition of the metaphorical sense $neck_4$ ("a narrow part of an artifact that resembles a neck") includes the word *neck*.[5] To avoid circularity, we use specially designed concept embeddings which are constructed without being exposed to the words in a language. We use the term *lexicon-agnostic* to describe these representations, because they are agnostic to the mapping between wordforms and concepts in a language's lexicon.

6.1 Embedding Construction

We construct four lexicon-agnostic embedding spaces. The first two are constructed using standard techniques, based on images and semantic relations (Figure 4b and 4c). The last two are constructed using non-standard techniques which extract lexicon-agnostic information from sense-annotated corpora (Figure 4d). These spaces are chosen as they are likely to include information that is useful for NSF, and because they can be constructed using existing resources. All four spaces are outlined below, with details in App. B.

Visual Embeddings To build visual representations, we exploit the hidden states of image classification models, which have been shown to encode abstract visual properties of images (Zhang et al., 2018). More specifically, we pass images of WordNet concepts through a high-performing image classification model, ConvNeXT (Liu et al., 2022), and extract the final hidden state that results from each image. We construct concept embeddings by calculating the mean hidden state for 256 images per concept, following Pinto and Xu (2021). The images we use are randomly sampled from ImageNet (Deng et al., 2009) and BabelPic (Calabrese et al., 2020b).

Taxonomic Embeddings In WordNet, concepts are linked to each other by lexical relations such as IS A or HAS A. We use the knowledge-graph embedding method TransE (Bordes et al., 2013) to construct embeddings based on these relations.

Syntagmatic Embeddings Words that occur close to each other are said to be *syntagmati-*

[4]Concept embeddings are not the same as sense embeddings: most WordNet-based sense embeddings (e.g. Calabrese et al., 2020a; Scarlini et al., 2020a,b) provide n different embeddings per synset, where n is the number of wordforms in the synset, while a concept embedding space provides a single representation per synset.

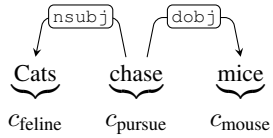
[5]It may be possible to modify glosses to remove the word that is used for that concept. However, this would not solve the problem, because the definition might contain a synonym or other words that are otherwise related to the target word.

any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails.

(a) A gloss

(b) An image

(c) A relation



(d) A sentence

Figure 4: The meaning of the concept c_{mouse} in different data modalities

cally related. Word2Vec (Mikolov et al., 2013) is a word embedding method which gives similar embeddings to words that are syntagmatically related, such as *hogwarts* and *dumbledore* (Levy and Goldberg, 2014). To create these embeddings, Word2Vec is trained to predict whether wordforms occur together in a sentence $\mathbf{w} \in \mathcal{W}^*$. Word2Vec can be adapted to build concept embeddings by instead training it using lists of concepts that occur together in text, $\mathbf{c} \in \mathcal{C}^*$. These lists can be created by extracting concept annotation from sense-tagged sentences. To construct data in this form, we extract concept co-occurrences from SemCor (Miller et al., 1994), NTU-MC (Tan and Bond, 2014), the WordNet Gloss Corpus, and every SensEval and SemEval dataset (Raganato et al., 2017). We also extract concept co-occurrences from Wikipedia, using the WSD model EWISER (Bevilacqua and Navigli, 2020) to produce sense annotation. Combined, these datasets yield a total of nearly one billion concept-annotated tokens. Word2Vec is trained on this data, resulting in an embedding space that encodes the syntagmatic relations between concepts.

Paradigmatic Embeddings Words that can be substituted for each other without violating a language's grammar are said to be *paradigmatically related*. Dep2Vec (Levy and Goldberg, 2014) is a variant of Word2Vec which produces similar embeddings for paradigmatically related words (e.g. embedding *hogwarts* close to other schools). It is trained using dependencies extracted from parse trees, such as $\langle \textit{chase}, \texttt{dobj}, \textit{mice} \rangle$. We can adapt Dep2Vec to build an embedding space that covers $\mathcal{C}$ rather than $\mathcal{W}$ by training it instead using dependencies between concepts, such as $\langle c_{\text{pursue}}, \texttt{dobj}, c_{\text{mouse}} \rangle$, as shown in Figure 4d. To identify concept dependencies, we parse Wikipedia using SpaCy, and determine the concept that corresponds to each word using the tags predicted by EWISER. Dep2Vec is trained using these dependencies in order to create an embedding space that captures paradigmatic relations between concepts.

6.2 Embedding Space Analysis

Each embedding space has a different coverage of $\mathcal{C}$, because they were constructed using resources which do not cover every WordNet concept. A breakdown of the coverage is shown in Table 2. In terms of nouns, the taxonomic embeddings cover every WordNet concept. The other three embedding spaces each cover about 60% of WordNet's nominal concepts, but the concepts which they cover are different. Figure 5 is a Venn diagram that shows the overlap between these spaces. In total, 31% of the nominal concepts in WordNet are covered by all three spaces. The paradigmatic and syntagmatic spaces both cover an additional 25%, while the visual space covers 24% that are not covered by either of those spaces.

To test whether the different spaces provide complementary information, we adapt a method used by Pinto and Xu (2021). This method evaluates whether similarity measures that are computed using different embeddings are correlated with each other. Low correlations indicate that different embeddings encode different information. We sample 10,000 pairs of concepts, and compute the cosine similarity between them in each space. We then compute the Spearman correlations between the similarity scores for each pair of spaces. These correlations are shown in App. B. The two spaces which are derived from overlapping data—the syntagmatic and paradigmatic spaces—are the

POS	Concept Embedding Space				WordNet
	Visual	Taxon.	Syntag.	Paradig.	
Noun	47,160 (57%)	82,115 (100%)	48,068 (59%)	45,918 (56%)	82,115
Verb	320 (2%)	13,542 (98%)	11,760 (85%)	11,742 (85%)	13,767
Adj.	0 (0%)	0 (0%)	14,832 (82%)	14,364 (79%)	18,156
Adv.	0 (0%)	0 (0%)	3,018 (83%)	2,878 (79%)	3,621
Total	47,480 (40%)	95,657 (81%)	77,678 (66%)	74,902 (64%)	117,659

Table 2: Concept embedding coverage

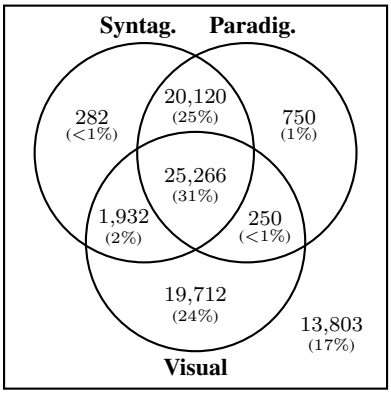


Figure 5: Overlap of nominal concepts

most strongly correlated ($\rho = 0.499$). The other correlations are all low, which indicates that the different embedding spaces capture different nuances of concept semantics.

7 Model Architecture

We design three NSF models. These models make predictions of $p(w \mid c)$ based on $\mathcal{C}_w^{\mathcal{P}}$, which are the concepts associated with w in a pruniverse $\mathcal{P}$:

$$\mathcal{C}_w^{\mathcal{P}} = \{c : \langle w, c\rangle \in \mathcal{P}\} \tag{1}$$

7.1 Cognitive Model

A *cognitive model* is a computer model that mimics a human cognitive process (Lake et al., 2017). Many previous works on sense extension have applied the same family of Bayesian cognitive models (e.g. Ramiro et al., 2018; Habibi et al., 2020; Pinto and Xu, 2021). We adapt these models to NSF. To make predictions, Bayes' theorem is applied to $p(w \mid c)$:

$$p(w \mid c) \propto p(c \mid w)\, p(w) \tag{2}$$

The prior can be estimated using frequency data with add-one smoothing ($\text{freq} : \mathcal{W} \mapsto \mathbb{N}$):

$$p(w) = \frac{1 + \text{freq}(w)}{|\mathcal{W}| + \sum_{w' \in \mathcal{W}} \text{freq}(w')} \tag{3}$$

The likelihood term can be defined in a variety of ways, dependent on different theories of categorisation. Let $\mathbf{e}_c \in \mathbb{R}^d$ be an embedding of c, and sim be a function that compares two embeddings ($\mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$). Three widely used variants are:

1. An "exemplar" model (e.g. Ramiro et al., 2018; Grewal and Xu, 2021; Habibi et al., 2020; Yu and Xu, 2023), based on the exemplar theory of categorisation (Medin and Schaffer, 1978; Nosofsky, 1986). This model computes the mean similarity between c and each concept associated with w:

$$p_{\text{exemplar}}(c \mid w) \propto \underset{c' \in \mathcal{C}_w^{\mathcal{P}}}{\text{mean}}\, \text{sim}(\mathbf{e}_c\,, \mathbf{e}_{c'}) \tag{4}$$

2. A "chaining" model (e.g. Sloman et al., 2001; Xu et al., 2016; Ramiro et al., 2018), based on the theory that word senses form sequential connections (Lakoff, 1987; Austin, 1961).[6] This model computes the maximum similarity between c and each concept associated with w:

$$p_{\text{chaining}}(c \mid w) \propto \max_{c' \in \mathcal{C}_w^{\mathcal{P}}} \text{sim}(\mathbf{e}_c\,, \mathbf{e}_{c'}) \tag{5}$$

3. A "prototype" model (e.g. Ramiro et al., 2018; Habibi et al., 2020; Grewal and Xu, 2021; Yu and Xu, 2023), based on prototype theory (Rosch, 1973). This model constructs an embedding that represents the prototypical sense of w, and then computes the similarity between c and this prototype:

$$p_{\text{prototype}}(c \mid w) \propto \text{sim}\Big(\mathbf{e}_c\,, \underbrace{\underset{c' \in \mathcal{C}_w^{\mathcal{P}}}{\text{mean}}\, \mathbf{e}_{c'}}_{\text{prototype}}\Big) \tag{6}$$

The similarity function is defined based on an exponential decay function, following Shepard (1987). We alter this function to have one free parameter h_m per modality $m \in \mathcal{M}$ (e.g. visual or taxonomic). Suppose that $\mathbf{e}_c$ is a concatenation of embeddings from different spaces, and $\mathbf{e}_c^m$ is a subvector of $\mathbf{e}_c$ that corresponds to m:

$$\text{sim}(\mathbf{e}_c\,, \mathbf{e}_{c'}) = \exp \sum_{m \in \mathcal{M}} \frac{-\,\text{dist}(\mathbf{e}_c^m\,, \mathbf{e}_{c'}^m)}{h_m} \tag{7}$$

Cosine distance is used for the distance function.

7.2 Perceptron Model

The cognitive model makes rigid assumptions about how conceptual similarity and word frequency influence sense extension. To avoid making these assumptions, we use a multi-layer perceptron (MLP). For each concept $c' \in \mathcal{C}_w^{\mathcal{P}}$, we construct a vector $\mathbf{d}_{\langle c, c'\rangle}$ that consists of the distance between c and c' in each embedding modality $m \in \mathcal{M}$. This vector is appended with the

[6] In some work, all three variants are called "chaining" models, and this variant is called the "nearest-neighbour" model (e.g. Habibi et al., 2020; Grewal and Xu, 2021).

unigram probability $p(w)$ and passed through an MLP with one output channel. The goodness of fit between a word w and a concept c (score : $\mathcal{W} \times \mathcal{C} \mapsto \mathcal{R}$) is calculated by finding the maximum output of the MLP for any $c' \in \mathcal{C}_w^{\mathcal{P}}$:

$$\text{score}(w, c) = \max_{c' \in \mathcal{C}_w^{\mathcal{P}}} \text{MLP}\left(\mathbf{d}_{\langle c, c' \rangle} \oplus [p(w)]\right) \quad (8)$$

Probabilities are calculated using softmax:

$$p(w \mid c) = \underset{\mathcal{W}}{\text{softmax}}\, \text{score}(c, w) \quad (9)$$

7.3 Transformer Model

The perceptron is able to learn how conceptual similarity and word frequency influence sense extension. However, it is likely that the factors that govern whether a new concept belongs to a word are more complex than geometric measures of similarity. We use a transformer (Vaswani et al., 2017) to test whether more sophisticated inter-concept relationships can be inferred. The input to a transformer's encoder is n vectors in $\mathbb{R}^d$. When a transformer is applied to language modelling, each vector corresponds to a token in a sentence. For our NSF transformer, on the other hand, each vector corresponds to a concept.

More specifically, to compute $p(w \mid c)$, an input is constructed that consists of a set of embeddings: one for each concept in $\mathcal{C}_w^{\mathcal{P}}$, and one for the new concept c. To give the model a signal about which of these is the new concept c, we concatenate a fixed encoding to each of these embeddings. A vector $\mathbf{h}_{\text{new}} \in \mathbb{R}^d$ is concatenated to the embedding of the new concept, and a different vector $\mathbf{h}_{\text{old}} \in \mathbb{R}^d$ is added to the embeddings of the concepts already associated with the word.[7] We call the resultant set of embeddings $\mathcal{E}_{\langle w,c \rangle}$:

$$\mathcal{E}_{\langle w,c \rangle} = \left\{\mathbf{e}_c \oplus \mathbf{h}_{\text{new}}\right\} \cup \left\{\mathbf{e}_{c'} \oplus \mathbf{h}_{\text{old}} : c' \in \mathcal{C}_w^{\mathcal{P}}\right\}$$

These embeddings are used as the input to a transformer's encoder. The output from this encoder is a new set of n vectors in $\mathbb{R}^d$. We extract the output vector which corresponds to the candidate concept c, and denote this vector $\mathbf{t}_{\langle w,c \rangle}$:

$$\mathbf{t}_{\langle w,c \rangle} = \text{Transformer}\left(\mathcal{E}_{\langle w,c \rangle}\right)_c \quad (10)$$

To compute $p(w \mid c)$, $\mathbf{t}_{\langle w,c \rangle}$ is passed through a linear layer with a single output, and then converted to a probability using softmax.

[7] These fixed encodings are analogous to the position encodings used in sequence modelling.

8 Evaluation

We evaluate each NSF model on each pruniverse. We also compare multimodal and unimodal model variants, in order to test the hypothesis that sense extension involves interactions between representational modalities (e.g. Aldrich, 1968; Forceville and Urios-Aparisi, 2009).

8.1 Experimental Setup

Models We experiment with unimodal and multimodal variants of each NSF model; the choice of the best embedding combination is treated as a hyperparameter. For the cognitive model, the choice of likelihood model is also treated as a hyperparameter. These models are trained on NSF using a negative log-likelihood objective. Two baselines are used. The first baseline predicts wordforms randomly with uniform probability; the second ranks wordforms according to their frequency in Wikipedia. Additional detail is in App. C.

Data Models are trained exclusively using senses within a given pruniverse, and evaluated on the corresponding unseen senses. Only nominal and verbal senses in each pruniverse are used, because our concept embedding spaces have considerably better coverage for these POS than for adjectives and adverbs (see Table 2).

Ranking Metric For each unseen concept c, the probability predictions $p(w \mid c)$ of a model can be computed for all n wordforms in the pruniverse. NSF can be evaluated by analysing the rank which a model assigns to the correct wordform. Let $r_{\langle w,c \rangle} \in \{1, \cdots, n\}$ be the rank that a model assigns to the correct wordform w when it is input with c. For an evaluation metric, we use the mean reciprocal rank (MRR). MRR is defined as:

$$\text{MRR}(\mathcal{U}) = \underset{\langle w,c \rangle \in \mathcal{U}}{\text{mean}} \frac{1}{r_{\langle w,c \rangle}} \quad (11)$$

This metric strictly penalises a model if it does not include the correct wordform among its top predictions: rank ten achieves a reciprocal rank of only 0.1, even though there are about 60,000 wordforms to choose from in most pruniverses.

Similarity Metric Another way that NSF predictions can be evaluated is by analysing how similar a model's top prediction is to the correct word. For each unseen sense, we compute the cosine

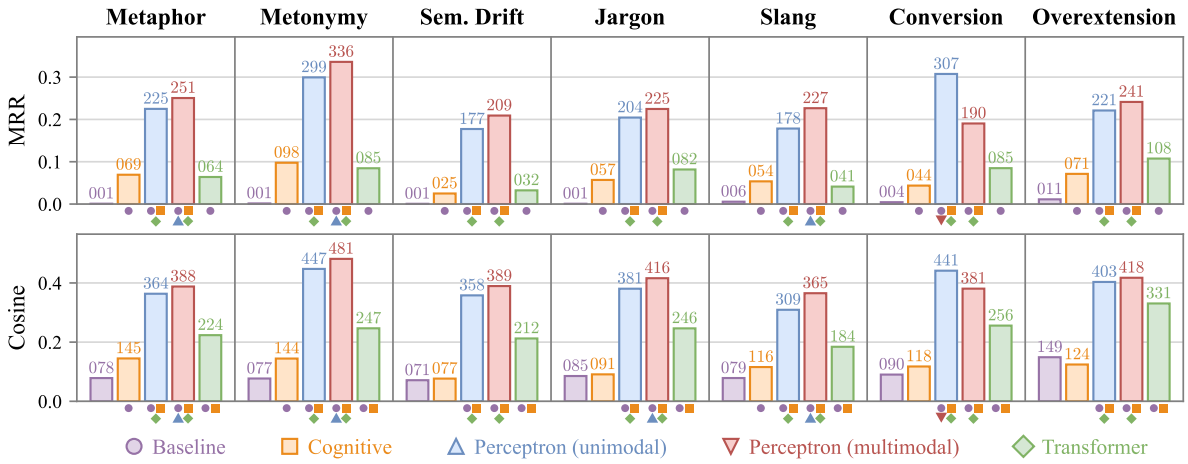


Figure 6: Results for ranking (top) and for similarity (bottom)

similarity between w and the model's top prediction w^*, using the pretrained Word2Vec embeddings of Mikolov et al. (2013). The overall score is taken as the mean.

Significance Testing To compute pairwise significance, we apply a two-tailed Monte Carlo permutation test at significance level $\alpha = 0.01$, computed with $R = 100{,}000$ permutations. The Benjamini–Hochberg procedure is used to correct against false discoveries.

8.2 Results and Discussion

Figure 6 presents the results for the ranking metric (top) and the similarity metric (bottom). If there is a symbol below a bar, this means that the model associated with that bar is significantly better than the model associated with the symbol. For example, if the result for model a has the symbol for model b underneath it, this indicates that model a is significantly better than model b. We report the results of unimodal and multimodal variants of the perceptron model separately, and only the highest-scoring baseline for each pruniverse is shown. Full baseline results are provided in App. E.

We can see from Figure 6 that both perceptron variants were significantly better than all of the other models, in all seven pruniverses. This is true for both metrics, meaning that the perceptron was not only likely to include the correct wordform in its top predictions (high MRR, top), but it was also likely to predict other words which were similar to the correct word (high similarity, bottom).

The perceptron achieved the highest overall performance on metonymy (0.336 MRR) and conversion (0.307 MRR). These high scores may reflect the fact that metonymy and conversion are highly regular processes (Pustejovsky, 1995). Relatively harder were slang (0.227 MRR) and jargon (0.225 MRR). The perceptron performed the least well on semantic drift, where it achieved 0.209 MRR; this score indicates that the correct wordform was predicted in about the 5$^{\text{th}}$ rank on average. The score differences between phenomena may reflect differences in the score ceiling that it is possible to obtain for each pruniverse. For example, some examples of semantic drift may have arisen because of specific events in the real world, and may be impossible to predict without additional context.

The multimodal variant of the perceptron model was significantly better than the unimodal variant for four pruniverses (metaphor, metonymy, slang, and jargon) in terms of the similarity metric. For example, for slang the multimodal perceptron attained a similarity score of 0.365, while the unimodal variant scored 0.309. The same pattern occurs in the MRR results, albeit with the exception of jargon, where the two variants were indistinguishable. These results support the claim that sense extension may involve interactions between representational modalities. The only setting where the unimodal variant outperformed the multimodal variant was on conversion. This result was achieved using the syntagmatic embeddings, which implies that predicting conversions may be possible from distributional properties alone.

The transformer performed less well than both perceptron variants in every pruniverse under both metrics, and the difference was usually large. For instance, the transformer scored 0.064 MRR for metaphor, while the perceptron achieved 0.251. Transformers require a large amount of data to train. It may be possible to improve the transformer's performance by utilising the data linked

	Pruniverse	Example Unseen Concept	Word	Top Predictions	Rank	Sim
1	Metaphor	"a miscalculation that recoils on its maker"	*backfire*	*backlash recoil* ***backfire*** *...*	3	.45
2	Metonymy	"a playing card with nine pips on the face"	*nine*	*seven eight five ten pitch...*	111	.94
3	Sem. Drift	"the basic unit of money in Brazil"	*real*	*peso escudo* ***real*** *florin...*	3	.08
4	Jargon	"a firm that transports people or goods"	*carrier*	*airline airway operator...*	180	.62
5	Slang	"a room that is comfortable and secluded"	*den*	*parlor poolroom study* ***den*** *...*	4	.34
6	Conversion	"a metal joint formed by softening with heat"	*weld*	*welding weldment* ***weld*** *...*	3	.66
7	Overext.	"[*watch*] a small portable timepiece"	*clock*	*necklace* ***clock*** *purse shoe...*	2	.11

Table 3: Example NSF predictions of the multimodal perceptron

to pruniverses in a more efficient manner. For example, a transformer could be trained directly on images, rather than on visual embeddings.

The cognitive model was also significantly worse than both perceptron variants in every setting. In terms of the similarity results, it was significantly worse than the transformer in all pruniverses, and was indistinguishable from the baselines in three pruniverses (semantic drift, jargon, and overextension). The large gap between the cognitive model and the perceptron suggests that the cognitive model is only partially able to explain how humans perform sense extension.

Table 3 shows examples of mistakes made by the multimodal perceptron.[8] In the table, each unseen concept is represented using a gloss, but note that models receive lexicon-agnostic representations instead of glosses. In these examples, the perceptron's top prediction is usually similar to the correct word, even in the cases where the correct word appeared at a low rank. Common mistakes were predicting a different member of the same category (e.g. predicting *seven* instead of *nine*, row 2), predicting a morphological variant of the correct word (e.g. *welding* instead of *weld*, row 6), and predicting a word in the wrong register (e.g. *parlor* instead of *den*, row 5).

Sometimes the perceptron attained a low similarity score despite the fact that its top prediction appears reasonable upon inspection. For example, to describe a Brazilian currency the model generated the word *peso*, when the correct answer was *real* (row 3). This prediction yielded a low similarity score (0.08), even though the *peso* and the *real* are both currencies in the Iberophone world. This low score reflects a limitation of the metric, not an issue with the model: the most common sense of *real* is unrelated to currencies, and it is this sense that will be represented by the word embeddings used in the metric. Similarly, to describe a watch, a child overextended the word *clock* (row 7), while the model predicted *necklace*. This prediction is reasonable, as a watch and a necklace are both metal accessories. However, although a clock and a necklace are both similar to a watch, they are not similar to each other, and therefore this prediction also attained a low similarity.

9 Cognitive Plausibility

If a model performs well on NSF, it could do so in a manner that is similar to how a human would perform the same task, or in a manner that is different. In this section, we analyse the cognitive plausibility of the different models.

9.1 Model Complexity

One factor that is used to gauge the cognitive plausibility of a model is model complexity. In cognitive science, there is often a preference for models with few parameters (e.g. Cibelli et al., 2016; Xu et al., 2017; Pinto and Xu, 2021). Table 4 shows the number of parameters in the models we evaluate. The best model was the perceptron, which has between 41 and 71 free parameters.[9] The cognitive model is heavily constrained, with one parameter per embedding space, while the transformer has many parameters. The poor performance of the transformer suggests that the amount of data may have been insufficient to fit these parameters.

One argument given to support the preference for cognitive models with few parameters is that it prevents models from overfitting (Roberts and Pashler, 2000; Pitt and Myung, 2002). Proponents of this view commonly utilise metrics such as the Akaike information criterion (AIC; 1973) or the Bayesian information criterion (BIC; Schwarz, 1978), which heavily penalise models with more

[8] App. F contains details of how examples were selected.

[9] The exact number of parameters in each model depends on the number of modalities. Parameters belonging to the models used to produce embeddings are not included.

Model	# Parameters
Cognitive	1 – 4
Perceptron	41 – 71
Transformer	342,017 – 457,217

Table 4: Parameters in NSF models

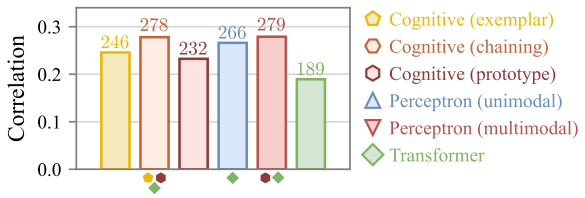


Figure 7: Prototypicality correlations

parameters. We agree that overfitting is problematic, but chose instead to control for overfitting using a train–test datasplit, sometimes absent from previous work on sense extension (e.g. Pinto and Xu, 2021). This enabled us to use NSF metrics which are more interpretable than AIC or BIC.

Another argument given to motivate the application of simple models is the claim that they are inherently more biologically plausible. For example, Gigerenzer and Brighton (2009) have argued that the brain employs inductive biases like those captured by simple models, and Yu and Xu (2025) have argued that storage limitations in the brain create natural complexity constraints. Opponents of this view, however, maintain that simple models have no basis in the neural processes that give rise to linguistic abilities (McClelland et al., 2010). Indeed, the brain has about 100 trillion synaptic connections (Stevens and Sullivan, 1998), so it is far less constrained than all the models we compare.

9.2 Prototype Effects

An alternative way the cognitive plausibility of a model can be evaluated is by investigating whether it exhibits emergent behaviours which correspond to observations made in psycholinguistic experimentation (McClelland et al., 2010). To this end, we conduct a post-hoc analysis to evaluate whether the models exhibit prototype effects. The hypothesis we test is that the "prototype" variant of the cognitive model accurately captures prototypicality, a tacit assumption in previous work that uses this model (e.g. Ramiro et al., 2018; Habibi et al., 2020; Grewal and Xu, 2021).

To conduct this study, we use the prototypicality norms of Uyeda and Mandler (1980), which consist of 28 categories, each with 30 different members rated for prototypicality.[10] For each category word w in this data (e.g. *fruit*), and for each of the concepts c which is an instance of that category (e.g. apple or tomato), we compute $p(w \mid c)$.[11] To investigate whether the models' predictions contain evidence of prototype effects, we calculate the Spearman correlation between each model's probability predictions and the human prototypicality ratings. Correlation scores are reported for each model, averaged over all pruniverses and all categories in the data from Uyeda and Mandler.

Results are shown in Figure 7. The results of the cognitive model variants are presented separately. Significance is computed using the same methodology as in the main experiment ($\alpha = 0.01$), and is again reported using symbols beneath the bars.

The multimodal perceptron and the chaining variant of the cognitive model attained the largest correlations numerically ($\rho = 0.279$ and 0.278). These models were both significantly better than the transformer ($\rho = 0.189$). The prototype variant of the cognitive model did not make predictions that correlated closely with human judgements of prototypicality ($\rho = 0.232$), and was significantly worse than the chaining variant of the same model, suggesting that it is not an accurate realisation of prototype theory.

In previous work, the different variants of the cognitive model have been used to evaluate the theories of categorisation that inspired their design. If one variant outperformed another, this result was interpreted as evidence of the superiority of one theory over another. For example, in their study of Chinese classifiers, Habibi et al. (2020) found that the exemplar model outperformed the prototype model, and based on this they conclude that classifiers may be "better captured by exemplar theory than by prototype theory". Our results suggest that this type of interpretation could be problematic, because the models may not correspond with the theories that inspired their design. There are many ways that a particular cognitive theory can be implemented. Instead of operationalising prototype theory via the use of prototype embeddings, as in eq. (6), an alternative operationalisation might be more appropri-

[10] This data has not previously been linked to WordNet, so we create a manual alignment; see App. G.

[11] Note that the directionality of conditionality is the inverse to that used in human recall experiments, in which a person is presented with a category w and asked to produce exemplar concepts c (e.g. Rosch et al., 1976).

ate. Indeed, while Rosch (1975) initially theorised that the brain encoded categories using prototype representations, she later discarded this position (Rosch, 1978, 1988, see also Lakoff, 1987, p. 43).

10 Limitations and Assumptions

Active vs. Passive Vocabulary A person's *active vocabulary* is the set of wordforms that they use in language production; their *passive vocabulary* is the set of wordforms that they understand but do not regularly use. NSF does not make this distinction: an NSF model is able to generate any wordform from its entire vocabulary.

Representational Leakage To avoid circularity, we used lexicon-agnostic embeddings. In order to build these embeddings, it was necessary to use datasets based partly on automatic annotation. Annotation errors in these datasets could introduce circularity. For example, if an image of a computer mouse in ImageNet is erroneously labelled as an occurrence of the rodent concept c_{mouse} (Figure 4), then the visual embedding of c_{mouse} might include a signal that these concepts colexify.

Isolation of Phenomena It may be possible for a model to reproduce an unseen sense without using the process which a pruniverse is intended to evaluate. This is because each sense can be related to many other senses in different ways. For example, a single sense could be a metaphorical extension of one sense, metonymically related to another, and morphologically related to a third. Because this sense is metaphorical, it could be included in the metaphor unseen set, even though an NSF model may be able to produce it using another process. This situation is difficult to avoid, because there is no dataset that systematically identifies every inter-sense relation.

Predefined Concepts Our formulation of sense extension is based on the standard model of word production (see §2), and we inherit its assumptions. That is to say, we assume that word production can be decomposed into discrete stages, including the formation of concepts (conceptualisation) and the assignment of concepts to words (formulation). NSF only treats formulation, but the brain also needs to perform conceptualisation. For example, when a scientist extends the word *flow* to refer to the movement of electrons, the concept of electron movement is not given to them—they create this concept themselves. Treating formulation in isolation from conceptualisation may be problematic, because these processes may be interlinked. For example, the concept of electron flow may only have arisen in the scientist's mind *because* they had the lexicalised concept of *flow* (e.g. Clark, 1996, ch. 10). This issue is highly contentious (see Lakoff, 1987, ch. 18).

11 Conclusion

We introduced a new task formulation for the study of sense extension: novel sense formulation (NSF). This task makes it possible to study different kinds of sense extension in a standardised manner, and also enables the use of multimodal resources for sense extension modelling. We created seven datasets, called pruniverses, which evaluate different types of sense extension, as well as four concept embedding spaces, which capture different aspects of the semantics of these sense extensions. Using these resources, we evaluated three models of novel sense formulation. The highest-performing model was a multimodal perceptron. It was able to achieve the best results on metonymy, but performed less well on semantic drift, suggesting that some kinds of sense extension pose unique challenges. A widely used cognitive model underperformed, suggesting that its explanation of sense extension is inadequate. Our code and data are made available online.[12]

Future work could explore the construction of models that are tailored to the challenges posed by specific phenomena. For example, a model designed for slang could employ a heuristic to estimate the humour of a sense, to reflect the observation that slang senses are often humorous (Danesi, 2010); a model for metaphor, on the other hand, could incorporate ratings of concreteness or emotivity. Pruniverses could also be produced for other phenomena, or for other languages. NSF could additionally be employed in research that investigates whether large language models (LLMs) generalise in a human-like manner (e.g. Zhuang et al., 2024). For example, an LLM could be trained on a corpus consisting of documents which precede a certain date, and evaluated using the semantic drift pruniverse. Alternatively, an LLM could be trained on a corpus that approximates the language that a child is exposed to during early development (e.g. Warstadt et al., 2023), and evaluated using the overextension pruniverse.

[12] https://github.com/rowanhm/pruniverse

Acknowledgements

The first author would like to thank all those with whom he discussed this work during its development, including Laurence Midgley, Tiago Pimentel, Ryoko Uno, Silke Mentchen, Josef Valvoda, Rami Aly, Cecily Whiteley, Siân Gooding, and Melanie Mitchell. He would also like to thank Agostina Calabrese, Michele Bevilacqua, and Riccardo Orlando for generating a BabelNet–WordNet mapping for us, as well as Emilia Butters who came up with the name *pruniverse*.

This work has received support under the Major Research Program of PSL Research University "CultureLab" launched by PSL Research University and implemented by ANR with the reference ANR-10-IDEX-0001.

References

Hirotogu Akaike. 1973. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 267–281. Akademiai Kiado.

Virgil C. Aldrich. 1968. Visual metaphor. *Journal of Aesthetic Education*, 2(1):73–86.

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. 2021. Pykeen 1.0: A python library for training and evaluating knowledge graph embeddings. *Journal of Machine Learning Research*, 22(82):1–6.

John L. Austin. 1961. The meaning of a word. *Journal of Symbolic Logic*, 35(4):23–43.

Thomas Bayes. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the Wordnet domains hierarchy: Semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, pages 94–101, Geneva, Switzerland. Association for Computational Linguistics.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 2854–2864, Online. Association for Computational Linguistics.

Max Black. 1962. *Models and Metaphor*. Cornell University Press.

Andreas Blank. 1999. Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. In Andreas Blank and Peter Koch, editors, *Historical Semantics and Cognition*, pages 61–90. De Gruyter Mouton.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 1006–1017, Online. Association for Computational Linguistics.

Lois Bloom. 1973. *One Word at a Time: The use of Single Word Utterances before Syntax*. Mouton.

Leonard Bloomfield. 1933. *Language*. Allen & Unwin, New York.

Gemma Boleda, Sebastian Padó, and Jason Utt. 2012. Regular polysemy: A distributional model. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 151–160, Montréal, Quebec. Association for Computational Linguistics.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26, pages 2787–2795, Lake Tahoe, Nevada. Curran Associates, Inc.

Brian F. Bowdle and Dedre Gentner. 2005. The career of metaphor. *Psychological Review*, 112(1):193.

Ted Briscoe and Ann Copestake. 1991. Sense extensions as lexical rules. In *Proceedings of the*

IJCAI Workshop on Computational Approaches to Non-Literal Language, pages 12–20, Sydney, Australia. University of Colorado at Boulder.

Thomas Brochhagen, Gemma Boleda, Eleonora Gualdoni, and Yang Xu. 2023. From language development to language evolution: A unified view of human lexical creativity. *Science*, 381(6656):431–436.

Joan Bybee. 2015. *Language Change*. Cambridge University Press.

Maria Teresa Cabré. 1999. *Terminology: Theory, Methods, and Applications*. John Benjamins, Amsterdam.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020a. EViLBERT: Learning task-agnostic multimodal sense embeddings. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI 2020)*, pages 481–487, Online. International Joint Conferences on Artificial Intelligence (IJCAI) Organization.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020b. Fatality killed the cat or: BabelPic, a multimodal dataset for non-concrete concepts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 4680–4686, Online. Association for Computational Linguistics.

Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. 2015. Improving distributed representation of word sense via WordNet gloss composition and context clustering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*, volume 2, pages 15–20, Beijing, China. Association for Computational Linguistics.

Emily Cibelli, Yang Xu, Joseph L. Austerweil, Thomas L. Griffiths, and Terry Regier. 2016. The Sapir-Whorf hypothesis and probabilistic inference: Evidence from the domain of color. *PLOS ONE*, 11(7):1–28.

Andy Clark. 1996. *Being There: Putting Brain, Body, and World Together Again*. The MIT Press.

Eve V. Clark. 1973. What's in a word? On the child's acquisition of semantics in his first language. In Timothy E. Moore, editor, *Cognitive Development and Acquisition of Language*, pages 65–110. Academic Press, San Diego.

Herbert H. Clark and Eve V. Clark. 1977. *Psychology and Language: An Introduction to Psycholinguistics*. Harcourt Brace Jovanovich.

Ann Copestake and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics*, 12(1):15–67.

Marcel Danesi. 2010. The forms and functions of slang. *Semiotica*, 2010(182):507–517.

Hoa Trang Dang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig. 1998. Investigating regular sense extensions based on intersective Levin classes. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING 1998)*, volume 1, pages 293–299, Montreal, Quebec. Association for Computational Linguistics.

Charles Darwin. 1859. *On the Origin of Species*. John Murray, London.

Gary S. Dell. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological review*, 93(3):283–321.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 248–255, Miami Beach, Florida. IEEE Computer Society.

Glenden Frank Drake. 1980. The social role of slang. In Howard Giles, Willam Peter Robinson, and Philip M. Smith, editors, *Language*, pages 63–70. Pergamon, Amsterdam.

Katrin Erk. 2006. Unknown word sense detection as outlier detection. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006)*, pages 128–135, New York City, New York. Association for Computational Linguistics.

Dan Fass. 1988. Metonymy and metaphor: What's the difference? In *Proceedings of the 12th International Conference on Computational Linguistics (COLING 1988)*, volume 1, pages 177–181, Budapest, Hungary. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Christiane Fellbaum and George A. Miller. 2003. Morphosemantic links in WordNet. *Traitement Automatique des Langues*, 44(2):69–80.

Charles Forceville and Eduardo Urios-Aparisi, editors. 2009. *Multimodal Metaphor*. Mouton de Gruyter.

Michael C. Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A. Marchman. 2017. Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, 44(3):677–694.

Alexandre François. 2008. *Semantic maps and the typology of colexification: Intertwining polysemous networks across languages*, pages 163–215. John Benjamins Publishing Company.

Timo Freiesleben and Thomas Grote. 2023. Beyond generalization: A theory of robustness in machine learning. *Synthese*, 202(4):109.

Victoria A. Fromkin. 1971. The non-anomalous nature of anomalous utterances. *Language*, 47(1):27–52.

Merrill F. Garrett. 1975. The analysis of sentence production. *Psychology of Learning and Motivation*, 9:133–177.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, volume 70, pages 1243–1252, Sydney, Australia. Proceedings of Machine Learning Research.

Raymond W. Gibbs. 1994. *The Poetics of Mind: Figurative Thought, Language, and Understanding*. Cambridge University Press.

Gerd Gigerenzer and Henry Brighton. 2009. Homo Heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1):107–143.

Jonathon Green. 2010. *Green's Dictionary of Slang*. Chambers.

Jonathon Green. 2016. *Slang: A Very Short Introduction*. Oxford University Press.

Karan Grewal and Yang Xu. 2021. Chaining algorithms and historical adjective extension. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*, pages 189–218. Language Science Press, Berlin.

Amir Ahmad Habibi, Charles Kemp, and Yang Xu. 2020. Chaining and the growth of linguistic categories. *Cognition*, 202:104323.

Douglas R. Hofstadter. 1985. *Metamagical Themas: Questing for the Essence of Mind and Pattern*. Basic Books.

Douglas R. Hofstadter and Emmanuel Sander. 2013. *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. Basic Books.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Takuya Ito, Tim Klinger, Doug Schultz, John Murray, Michael Cole, and Mattia Rigotti. 2022. Compositional generalization through abstract representations in human and artificial neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pages 32225–32239. Curran Associates, Inc.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

George Lakoff. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, 1st edition. University of Chicago Press.

George Lakoff. 1993. The contemporary theory of metaphor. In Andrew Ortony, editor, *Metaphor and Thought*, pages 202–251. Cambridge University Press.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 591–601, Avignon, France. Association for Computational Linguistics.

Adrienne Lehrer. 1974. *Semantic fields and lexical structure*. Elsevier.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.

Willem J.M̃. Levelt. 1989. *Speaking: From intention to articulation.* The MIT Press.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, volume 2, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Jonathan E. Lighter. 2001. Slang. In John Algeo, editor, *The Cambridge History of the English Language*, pages 219–252. Cambridge University Press.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 11976–11986, New Orleans, Louisiana. IEEE Computer Society.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019)*, New Orleans, Louisiana. Curran Associates, Inc.

Hans Marchand. 1969. *The Categories and Types of Present-Day English Word Formation: A Synchronic-Diachronic Approach*. München.

Luisa Martín Rojo. 2010. Jargon. In Mirjam Fried, Jan-Ola Östman, and Jef Verschueren, editors, *Variation and Change: Pragmatic perspectives*, pages 155–170. John Benjamins Publishing Company.

Rowan Hall Maudslay and Simone Teufel. 2022. Homonymy information for English WordNet. In *Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference*, pages 90–98, Marseille, France. European Language Resources Association.

Rowan Hall Maudslay, Simone Teufel, Francis Bond, and James Pustejovsky. 2024. ChainNet: Structured metaphor and metonymy in WordNet. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2984–2996, Torino, Italy. European Language Resources Association and International Committee on Computational Linguistics.

James L. McClelland, Matthew M. Botvinick, David C. Noelle, David C. Plaut, Timothy T. Rogers, Mark S. Seidenberg, and Linda B. Smith. 2010. Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8):348–356.

Douglas L. Medin and Marguerite M. Schaffer. 1978. Context theory of classification learning. *Psychological Review*, 85(3):207–238.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the Human Language Technology Conference*, pages 240–243, Plainsboro, New Jersey. Association for Computational Linguistics.

Carol Myers-Scotton and Janice L. Jake. 1995. Matching lemmas in a bilingual language competence and production model: Evidence from intrasentential code switching. *Linguistics*, 33(5):981–1024.

Irina Nikishina, Varvara Logacheva, Alexander Panchenko, and Natalia Loukachevitch. 2020. Studying taxonomy enrichment on diachronic WordNet versions. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 3095–3106, Online and Barcelona, Spain. International Committee on Computational Linguistics.

Robert M. Nosofsky. 1986. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39.

Dieke Oele and Gertjan van Noord. 2018. Simple embedding-based word sense disambiguation. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, pages 259–265, Singapore. Global Wordnet Association (GWA).

Andrew Ortony. 1975. Why metaphors are necessary and not just nice. *Educational Theory*, 25(1):45–53.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035, Vancouver, British Columbia. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. Slang detection and identification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL 2019)*, pages 881–889, Hong Kong, China. Association for Computational Linguistics.

Jean Perrin. 1901. Les hypothèses moléculaires. *Revue Scientifique*, 15:449–461.

Jean Piaget. 1952. *The Origins of Intelligence in Children*. W. W. Norton & Co.

Renato Ferreira Pinto and Yang Xu. 2021. A computational theory of child overextension. *Cognition*, 206:104472.

Mark A. Pitt and In Jae Myung. 2002. When a good fit can be bad. *Trends in Cognitive Sciences*, 6(10):421–425.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

James Pustejovsky and Anna Rumshisky. 2010. Mechanisms of sense extension in verbs. In Gilles-Maurice de Schryver, editor, *A Way with Words: Recent Advances in Lexical Theory and Analysis*. Mehna Publishers.

Hugh Rabagliati, Gary F. Marcus, and Liina Pylkkänen. 2010. Shifting senses in lexical semantic development. *Cognition*, 117(1):17–37.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework

and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, volume 1, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Christian Ramiro, Mahesh Srinivasan, Barbara C. Malt, and Yang Xu. 2018. Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, 115(10):2323–2328.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. European Language Resources Association.

Leslie A. Rescorla. 1980. Overextension in early language development. *Journal of Child Language*, 7(2):321–335.

Seth Roberts and Harold Pashler. 2000. How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107(2):358–367.

Eleanor Rosch. 1973. Natural categories. *Cognitive Psychology*, 4(3):328–350.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192.

Eleanor Rosch. 1978. Principles of categorization. In Eleanor Rosch and Barbara B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Routledge.

Eleanor Rosch. 1988. Coherences and categorization: A historical view. In Frank S. Keller, editor, *The Development of Language and Language Researchers: Essays in Honor of Roger Brown*, pages 373–392. Erlbaum Mahwah.

Eleanor Rosch, Carol Simpson, and R. Scott Miller. 1976. Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, 2(4):491–502.

Peter Rossen Skadhauge. 1995. Sense extension functions in lexical semantics. In *Proceedings of the 10th Nordic Conference of Computational Linguistics (NODALIDA 1995)*, pages 59–68, Helsinki, Finland. Department of General Linguistics, University of Helsinki, Finland.

Arkadiy Saakyan, Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 2022. A report on the FigLang 2022 shared task on understanding figurative language. In *Proceedings of the 3rd Workshop on Figurative Language Processing*, pages 178–183, Online and Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ferdinand de Saussure. 1916. *Cours de Linguistique Générale*. Payot, Paris.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 8758–8765, New York City, New York. AAAI Press.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 3528–3539, Online. Association for Computational Linguistics.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Roger N. Shepard. 1987. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Steven A. Sloman, Barbara C. Malt, and Arthur Fridman. 2001. Categorization versus similarity: The case of container names. In *Similarity and Categorization*. Oxford University Press.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Charles F. Stevens and Jane Sullivan. 1998. Synaptic plasticity. *Current Biology*, 8(5):151–153.

Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. Toward informal language processing: Knowledge of slang in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2024)*, volume 1, pages 1683–1701, Mexico City, Mexico. Association for Computational Linguistics.

Zhewei Sun, Richard Zemel, and Yang Xu. 2021. A computational framework for slang generation. *Transactions of the Association for Computational Linguistics*, 9:462–478.

Don R. Swanson and Neil R. Smalheiser. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91(2):183–203.

Henry Sweet. 1891. *New English Grammar, Logical and Historical*. Clarendon Press, Oxford.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. In *Computational Approaches to Semantic Change*, pages 1–91. Language Science Press, Berlin.

Liling Tan and Francis Bond. 2014. NTU-MC toolkit: Annotating a linguistically diverse corpus. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 86–89, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, volume 1, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.

Stephen Ullmann. 1962. *Semantics: An Introduction to the Science of Meaning*. Blackwell, Oxford.

Katherine M. Uyeda and George Mandler. 1980. Prototypicality norms for 28 semantic categories. *Behavior Research Methods & Instrumentation*, 12(6):587–595.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, Long Beach, California. Curran Associates, Inc.

Tony Veale and Yanfen Hao. 2008. A fluid knowledge representation for understanding and generating creative metaphors. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 945–952, Manchester, England. COLING 2008 Organizing Committee.

John Venn. 1880. On the diagrammatic and mechanical representation of propositions and reasonings. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 10(59):1–18.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2018. UFSAC: Unification of sense annotated corpora and tools. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1027–1034, Miyazaki, Japan. European Language Resources Association.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM*

Challenge at the 27th Conference on Computational Natural Language Learning, pages 1–34, Singapore. Association for Computational Linguistics.

Pete Whittaker. 2020. *Crack Climbing: Mastering the Skills and Techniques*. Vertebrate Publishing, Sheffield.

Andrew M. Wilkinson. 1992. Jargon and the passive voice: Prescriptions and proscriptions for scientific writing. *Journal of Technical Writing and Communication*, 22(3):319–325.

Yang Xu, Terry Regier, and Barbara C. Malt. 2016. Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, 40(8):2081–2094.

Yang Xu, Terry Regier, and Nora S. Newcombe. 2017. An adaptive cue combination model of human spatial reorientation. *Cognition*, 163:56–66.

Lei Yu. 2023. Systematic word meta-sense extension. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 10953–10966, Singapore. Association for Computational Linguistics.

Lei Yu and Yang Xu. 2021. Predicting emergent linguistic compositions through time: Syntactic frame extension via multimodal chaining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pages 920–931, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lei Yu and Yang Xu. 2022. Noun2Verb: Probabilistic frame semantics for word class conversion. *Computational Linguistics*, 48(4):783–818.

Lei Yu and Yang Xu. 2023. Word sense extension. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, volume 1, pages 3281–3294, Toronto, Ontario. Association for Computational Linguistics.

Lei Yu and Yang Xu. 2025. Infinite mixture chaining: An efficiency-based framework for the dynamic construction of word meaning. *Open Mind*, 9:1–24.

Xiao Yu, Quanquan Gu, Mianwei Zhou, and Jiawei Han. 2012. Citation prediction in heterogeneous bibliographic networks. In *Proceedings of the 2012 SIAM International Conference on Data Mining (SDM 2012)*, pages 1119–1130. Society for Industrial and Applied Mathematics.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pages 586–595, Salt Lake City, Utah. IEEE Computer Society.

Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. 2024. Visual grounding helps learn word meanings in low-data regimes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2024)*, volume 1, pages 1311–1329, Mexico City, Mexico. Association for Computational Linguistics.

A Pruniverse Details

To build pruniverses, we retrieve every sense from the Princeton WordNet version 3.0 (Fellbaum, 1998), convert all wordforms to lowercase, and filter out wordforms which contain either a dash or whitespace (to avoid multi-word phrases). With the exception of the overextension pruniverse, each pruniverse is formed by identifying an unseen set and then removing these senses from this sense inventory.

When constructing unseen sets, we ensure that every unseen sense $\langle w, c\rangle \in \mathcal{U}$ meets two criteria. The first criterion is that the word w has at least one sense in the corresponding pruniverse (i.e. $\forall\langle w, c\rangle \in \mathcal{U}, \exists c' \in \mathcal{C} : \langle w, c'\rangle \in \mathcal{P}$). This criterion ensures that no unseen sense is a neologism. The second criterion is that there are no synonyms for the concept c in the pruniverse (i.e. $\forall\langle w, c\rangle \in \mathcal{U}, \nexists w' \in \mathcal{W} : \langle w', c\rangle \in \mathcal{P}$). This criterion ensures that each unseen sense simulates exposure to a new concept, rather than the use of novel language to describe a familiar concept. To meet the second criterion, we additionally remove any sense from a pruniverse that covers the same concept as an unseen sense. Every unseen

sense is nominal, because several of the resources which are used to identify unseen senses only cover nouns.

For the metaphor and the metonymy pruniverses, a sense is only used as an unseen sense if it is not a parent of any further metaphor or metonym in ChainNet. This is done to reduce the chance of a sense being put in the unseen set which is critical to the understanding of other senses of a word.

For the jargon pruniverse, any sense which is connected to another sense in WordNet by the instance hypernymy relation is removed, as these correspond to proper nouns. Any sense which corresponds to either a natural language or a political orientation is also filtered out, because these constitute a disproportionate number of the senses in the data. This filtering is achieved by removing hyponyms of the natural language concept and the political orientation concept in WordNet.

For the conversion pruniverse, a nominal sense is only used as an unseen sense if it derives from a verbal sense, and not vice versa. In their annotation of morphological derivation in WordNet, Fellbaum and Miller (2003) do not identify the directionality of derivations. To account for this, we approximate this information using corpus data. More specifically, the direction of the derivation between a pair of senses is assumed to go from the sense which is more frequently used to the sense which is less frequently used. This approximation does not attempt to recreate the etymological reality of the derivation, but instead captures the fact that a speaker is more likely to encounter senses which are more frequent before they encounter senses which are less frequent. Therefore, a nominal sense is only included in the unseen set if it belongs to a wordform that occurs more frequently as a verb than as a noun. To estimate these frequencies, statistics are extracted by POS-tagging Wikipedia using SpaCy (Honnibal et al., 2020).

B Embedding Details

When constructing embedding spaces, the default hyperparameters of all models are used unless mentioned otherwise. The embedding size is set to 300 in all cases, for parity between the spaces. For the visual embeddings, the 768-dimensional ConvNext embeddings are projected into a 300-dimensional space using the Scikit-Learn (Pedregosa et al., 2011) implementation of SVD. For the taxonomic embeddings, TransE is trained using the implementation in PyKeen (Ali et al., 2021). For the syntagmatic embeddings, the Gensim implementation (Řehůřek and Sojka, 2010) of Word2Vec is used, and UFSAC versions of sense-annotated corpora (Vial et al., 2018) are used when available. For the paradigmatic embeddings, Levy and Goldberg's (2014) original implementation of Dep2Vec is used, with the negative sampling rate set to 15.

Details of the correlations between the embedding spaces are shown below:

	Visual	**Syntag.**	**Paradig.**
Taxon.	.029	.009	.005
Visual		.247	.137
Syntag.			.499

C Model Implementation

Models are implemented in PyTorch (Paszke et al., 2019). To train multimodal variants, embeddings from the four embedding spaces are concatenated together in different combinations. Some concepts are not represented in all four spaces. When a concept is missing from a space it is represented using the mean embedding from that space. For the perceptron model, whenever either c or c' is missing from a particular space, the distance value corresponding to that space is set to 0 in $\mathbf{d}_{\langle c,c' \rangle}$.

To compute $p(w)$, word type frequency counts are computed using Wikipedia. Before computing these counts, Wikipedia is filtered to only include senses in a particular pruniverse. This ensures that the frequency counts do not signal the existence of unseen senses. To filter Wikipedia, sense annotation produced by EWISER is used.

For the cognitive model, to ensure numerical stability we use a softplus function to clamp the minimum value of each h_m to $|\mathcal{M}|/10$. For the transformer, the PyTorch implementation is used, but the input embeddings are first passed through a linear layer to rescale them to be a consistent dimensionality d regardless of how many modalities were concatenated to each other. For the binary position encodings $\mathbf{h}_{\text{new}}$ and $\mathbf{h}_{\text{old}}$, the sinusoidal encoding of Vaswani et al. (2017) is used, with the dimensionality set to 300 to match the concept embedding spaces; another option would be to use learned embeddings (e.g. Gehring et al., 2017).

D Training and Hyperparameterisation

Separate models are trained for each phenomenon, using the senses in the phenomenon's pruni-

Pruniverse	Model		
	Cog.	Perc. (Uni)	Perc. (Multi)
Metaphor	p, P	S	V+T+S+P
Metonymy	x, S+P	S	V+T+S+P
Sem. Drift	x, P	S	S+P
Slang	x, S+P	S	S+P
Jargon	x, P	S	V+T+S+P
Conversion	x, S	S	V+T+S+P
Overext.	p, P	P	S+P

Likelihood models: [x]exemplar [c]chaining [p]prototype
Embeddings: [V]Visual [T]Taxonomic [S]Syntagmatic [P]Paradigmatic

Table 5: Highest performing model variants

verse $\mathcal{P}$. For each $\langle w, c\rangle \in \mathcal{P}$, the concept c is input to a model, and $p(w \mid c)$ is computed for 256 wordforms. One of these wordforms is the correct solution, and the other 255 wordforms are selected randomly from the pruniverse. Models are optimised using a negative log-likelihood objective with AdamW (Loshchilov and Hutter, 2019). The cognitive model is trained using only the nominal senses in the illustrative sets; the perceptron and transformer have more parameters, so they are trained using every nominal sense in each pruniverse. Of the nominal senses used for model training, 10% are set aside for a development set. Early stopping is used to end training when the loss on the development set fails to reduce for 20 consecutive steps. For the transformer, early stopping was performed after 80 steps, because transformers benefit from longer training.

We performed a grid search across different representational spaces (for all models) and likelihood model variants (for the cognitive model). Training every model on every pruniverse with every combination of embedding space is not feasible. We therefore chose to train four unimodal variants of each model (each using one embedding space) and three multimodal variants (paradigmatic + syntagmatic; paradigmatic + syntagmatic + visual; paradigmatic + syntagmatic + visual + taxonomic). These multimodal combinations were chosen as they led to the best performance in preliminary experimentation. For the cognitive model, we trained variants with each of the three likelihood models. The transformer was very computationally expensive to train, so we only trained it on a single embedding space combination (paradigmatic + syntagmatic + visual), as we found that it performed the best with this combination in preliminary study. For each pruniverse, we therefore trained 21 cognitive models, 7 perceptrons, and 1 transformer, for a total of 203 models when all pruniverses are considered. The highest performing variant of each model for each pruniverse was chosen based on loss on a development set; the variant that was chosen is in Table 5.

Model	b	γ	n	h	x	a	d
Cog.	128	1×10^{-2}					
Perc.	32	1×10^{-4}	2	10	0.1		
Trans.	64	5×10^{-6}	2	256	0.1	4	128

Table 6: Base model hyperparameters

Models have several additional hyperparameters. These are listed below:

- The batch size, b.
- The learning rate, γ.
- The number of layers in the perceptron and transformer models, n.
- The size of the hidden state in each feedforward layer in the perceptron and transformer models, h.
- The dropout (Srivastava et al., 2014) for the perceptron and transformer models, x.
- The number of attention heads in the transformer model, a.
- The dimensionality of the encoder in the transformer model, d.

Performing a parameter search over all of these parameters for all 203 model variants was too computationally expensive, so we treated these hyperparameters as fixed. The fixed hyperparameters are shown in Table 6. These were chosen based on results in preliminary experimentation.

E Result Processing Details

When extracting the top predictions from models (w^*), wordforms are excluded that are not in the vocabulary of the Word2Vec space used to calculate similarity. This is necessary as otherwise it would not be possible to compute the similarity between the model's top prediction w^* and the correct word w.

When calculating the rank $r_{\langle w,c\rangle}$ that a model assigns to the correct word w for concept c, synonyms which also have a sense for c in WordNet are excluded. This is to avoid penalising a model which produces an alternative correct solution. When extracting rank predictions from models, we additionally need to account for the possi-

Pruniverse	$\lvert\mathcal{W}\rvert$
Metaphor	59,634
Metonymy	59,674
Semantic Drift	59,721
Slang	59,796
Jargon	59,796
Conversion	59,694
Overextension	600

Table 7: Pruniverse vocabulary sizes

bility that a model assigns the same probability to multiple wordforms. In these instances, we assign each wordform in the tied group the lowest rank in the ordered ranking. Thus, if a model predicts two wordforms with equal probability as its top prediction, both of these wordforms is assigned rank 2. This ranking scheme is sometimes known as "1334 ranking".

The size of the vocabulary which models need to choose outputs from is shown in Table 7. A breakdown of the results of both baselines is shown in Table 8.

F Example Selection

Here we comment on how examples were selected for Table 3. Examples 2 and 5 were chosen because they attained a high similarity score but low reciprocal rank. Example 7 was chosen because it had a high reciprocal rank but low similarity score. Examples 3 and 6 were chosen to illustrate common mistake patterns. Finally, examples 1 and 4 were chosen to represent "average" cases, namely those between between the 25$^{\text{th}}$ and 50$^{\text{th}}$ percentile in terms of rank prediction.

G Prototypicality Data

We manually aligned Uyeda and Mandler's (1980) prototypicality norms with WordNet. More specifically, every category name was aligned with the closest wordform in the models' vocabularies

Pruniverse	MRR		Cosine	
	Rand.	Freq.	Rand.	Freq.
Metaphor	.000	**.001**	**.078**	.066
Metonymy	.000	**.001**	**.077**	.072
Sem. Drift	.000	**.001**	**.071**	.061
Jargon	.000	**.001**	**.085**	.064
Slang	.000	**.006**	.068	**.079**
Conversion	.000	**.004**	.080	**.090**
Overext.	**.011**	.005	**.149**	.074

Table 8: Breakdown of baseline results

Category	WordNet Term
kind of cloth	*cloth*
kind of money	*money*
member of the clergy	*cleric*
type of human dwelling	*dwelling*
type of reading material	*reading material*
weather phenomenon	*weather condition*

Table 9: Prototypicality norms alignment

(e.g. *tree*), and each member of these categories was aligned with the closest concept in WordNet. Some of the categories in the prototypicality data of Uyeda and Mandler (1980) are based on a multi-word phrases that do not appear in the WordNet vocabulary, such as "member of the clergy". To handle these cases, we selected the word in the WordNet vocabulary that is closest in meaning. The categories that are affected are shown in Table 9. In eight of the 28 categories in Uyeda and Mandler's data we found no reasonable corresponding word in the vocabulary of models, so we excluded these from the data. The 20 included categories are *article of clothing*, *article of furniture*, *bird*, *cleric*, *cloth*, *color*, *country*, *dwelling*, *fruit*, *metal*, *money*, *musical instrument*, *reading material*, *sport*, *toy*, *tree*, *vegetable*, *vehicle*, *weapon*, and *weather condition*.