Chapter 2

Background

In this chapter, I argue for a new perspective on comparative concepts in linguistic typology, which grounds operationalizations of complex hybrid concepts in empirical measures of underlying linguistic dimensions. I present the major goals of typology and the challenges of defining comparative concepts for typology, and review existing approaches to these challenges and how they compare to my proposed approach. I then review the development of deep learning models of language, describing how they provide new avenues for defining empirical measures of semantic dimensions of language, and the richness of the semantic and conceptual information they acquire. I then provide a review of the study and application of comparative concepts in *computational* typology, highlighting how building rich empirical models of underlying semantic and perceptual spaces have been key to successful computational typological research, and the parallels between these approaches and my proposed approach. I also highlight the shortcomings of current discrete approaches to semantics in computational typology. Together, this motivates the empirical grounding approach to comparative concepts and linguistic categories that I take in this thesis.

Finally, I provide a high level overview of the lexicality spectrum, defining

formal and functional dimensions of lexicality, and describing their interrelationships. This sets the stage for the remainder of the thesis, which focuses on defining empirical measures of these dimensions, leveraging deep learning models of language, and investigating how these dimensions relate to existing lexicality-related distinctions in multilingual databases.

2.1 Typology and Comparative Concepts

Linguistic typology is the study of variation across the world's languages. Typologists perform cross-linguistic comparisons with the aim of making generalizations about this variation. Such generalizations may consist of identifying and classifying languages into a small set of types (typological classification) or identifying cross-linguistically consistent patterns in variation. By studying this variation, typologists aim to identify the limits on and universals of human languages, and, often, to identify simple, language-neutral explanations of these limits.

To make cross-linguistic comparisons and identify cross-linguistic variation, typological research has explicitly or implicitly had to identify a frame of *alignment* between languages–typically taking the form of shared concepts identified across languages. Take, for example, the study of basic word order typology:

(2.1) *Paul* *kisses* *Peter.*
SUBJ VERB OBJECT

(2.2) *pooru-wa* *piitaa-wo* *kisu-shiteiru*
Paul-TOPIC Peter-ACC kiss-DO-PRES.CONT
SUBJ OBJECT VERB

English and Japanese, then, vary in their basic word orders. In English, the verb is proceeded by the subject and followed by the object ("SVO order"), while in Japanese, the default ordering is subject-object-verb ("SOV"). This comparison, however, relies on the consistent cross-linguistic identification of

the categories of SUBJECT, VERB, and OBJECT. Such concepts over which cross-linguistic comparisons can be made have been termed COMPARATIVE CONCEPTS (Haspelmath, 2010; Croft, 2016).

Many of these comparative concepts present serious methodological challenges. It is well known that many categories in linguistics are semantically *motivated*, but not semantically *defined*. Take, for example, the category of subject. While subject's across languages are typically the agents of an action, in the specifics of individual languages, there is additional complexity. While English has "SVO" order, in passive constructions, it is the patient and not the agent which appears in this initial position:

(2.3) *Paul is kissed by Peter.*
SUBJ VERB OBJECT

Defining the subject in terms of the obvious distributional commonality between *Paul* in (2.1) and (2.3) would make the claim "English is an SVO language" circular. While getting around this particular problem is relatively straightforward (through the deployment of additional constructional tests), this type of issue is pervasive in typological analysis, and careless or inconsistent application of categories cross-linguistically can lead to generalizations or even debates which are vacuous.

For example, frequent debates have occurred over whether a particular language has the category ADJECTIVE: a syntactic category covering property words, distinct from nouns and verbs. The typical structure of such debates involves identifying the behaviour of words which denote properties in a particular language in various constructions, and comparing their behaviour to members of other classes. For example, in Korean, both adjectives and verbs (but not nouns) inflect for tense, leading some to argue that Korean lacks a class of adjectives, and to claim that in Korean, adjectives are a type of stative verb. On the other hand, some have argued that because adjectives in Korean are somewhat

restricted in terms of the tense–aspect–mood constructions they can appear with, they are better analysed as a distinct class. However, cross-linguistically, adjectives rarely inflect for tense or aspect, so this distinction is being made on a very language-particular basis.

Croft (2001) calls this type of syntactic argumentation METHODOLOGICAL OPPORTUNISM: the application of arbitrary language-particular criteria to identify distinctions between supposedly universal categories. This approach cannot lead to consistent generalizations across languages. If we consider a generalization like "adjectives do not inflect for tense", then Korean is a counterexample if adjectives are not a type of verb. If they are a type of verb, then Korean is a counterexample to the generalization "adjectives require some kind of copula-like element in prediction". To understand what the actual generalizations in typology are and whether a particular language is or is not a counterexample, we need to base these comparisons on cross-linguistically consistent criteria.

What the best comparative concepts are for a given problem is an empirical question, based on their predictive power in terms of generalizations about language variation. All the comparative concepts I have discussed so far are HYBRID CONCEPTS (Croft, 2016): they combine aspects of formal distribution with semantics. As an alternative, we might consider FUNCTIONAL[1] compar-

[1]By simultaneously addressing both issues of comparative concepts and the lexical–functional distinction in this thesis, I am trapped into a very confusing overload of the term *functional;* it has two, almost diametrically opposed meanings in the literature. As discussed in Chapter 1 the context of the lexical–functional distinction, *functional* refers to a pole of a continuum of linguistic behaviour, where "functional" items/elements/units are those which serve primarily to organize and clarify relationships between other elements. These elements are often described as "grammatical" or "lacking meaning". In the context of comparative concepts and typological theory, however, linguistic *function* refers to the communicative content of linguistic expressions, as contrasted with its *form*: the specific linguistic realization which conveys a function. In this sense, the function of an adjective is basically its intension: the property it denotes, while the function of a verbal tense marker is the temporal and aspectual information it conveys about the event describe by the verb. Thus, a "functional comparative concept" in this sense is one which is defined in terms of the communicative content of the linguistic expression, rather than its formal distribution. Here, the terms "function" and "functional" are preferred to "semantics" and "semantic" because the later terms are often taken to refer only to truth-conditional content, while function is inclusive information structure.

Of course, these senses are related, but in a quasi-antonymic manner: more functional elements have more abstract, relational, and language-internal functions. In this thesis, when I use the

ative concepts: concepts that invoke only the communicative function of the linguistic expressions studied, regardless of their formal distribution. Returning to the adjective example, this would be making comparisons of all expressions of property meanings across languages, as suggested by Haspelmath (2012). However, as noted by Croft (2016), such broad functional comparative concepts may fail to capture important cross-linguistic distinctions that are relevant for typological generalizations. Different types of properties may have different distributions within a single language: the expression of colour properties may pattern differently from emotions, for example. Thus, increasingly fine-grained functional comparative concepts are often necessary to capture cross-linguistic variation. Even with purported "functional" comparative concepts, something of the formal tends to creep in. Ultimately, after all, the typological generalizations are about the behaviour of linguistic expressions, which are formal entities. There are therefore two major dimensions of challenge in identifying useful, valid comparative concepts for typology: selecting the right functions, and formalizing categories of linguistic expressions which align with these functions. In the next section, I review different meta-approaches to these challenges, and describe the general way I tackle these issues in the thesis, which diverges from prior work in important ways.

2.1.1 Defining Hybrid Comparative Concepts

Constructions and strategies **?** provides a useful terminological and conceptual framework for describing how typological research can and has often implicitly combined form and function into useful cross-linguistic generalizations, which I will use to frame the discussion in this section. We will follow **?** to define a FUNCTIONAL CONSTRUCTION (p. 17):

word "functional" to describe words, morphemes, items, elements, or units, I mean it in the lexical–functional sense. Other uses of the words "function" or "functional" generally refer to the communicative content sense (function as opposed to form), unless otherwise specified.

> any pairing of form and function in a language (or any language) used to express a particular combination of semantic content and information packaging[2]

This type of construction is defined only by *what* it expresses. Croft contrasts this with the narrower STRATEGY **?**, p. 19:

> a construction in a language (or any language), used to express a particular combination of semantic structure and information packaging function (the *what*), that is further distinguished by certain characteristics of grammatical form that can be defined in a crosslinguistically consistent fashion (the *how*).

The separation of strategies from functional constructions is a useful conceptual tool for approaching conflicting cross-linguistic formal data about the expression of a particular function. For example, one argument that Korean Adjectives are a type of verb is that Korean Nouns require a copula to be predicated, while Adjectives do not. Rather than saying "Korean lacks adjectives", we can say that Korean uses different strategies for predicating nouns and adjectives, with nouns using a copula strategy, and verbs using a zero-copula strategy. This method of description foregrounds the actual distributional data that needs to be explained. Indeed, empirically there are many interesting questions about the cross-linguistic distribution and co-occurrence of different strategies for particular linguistic functions.

Semantic maps An extremely influential method for relating form to function in linguistic typology is the use of SEMANTIC MAPS (Haspelmath, 2003; Croft, 2002, pp. 133–139).[3] As I discussed, a major problem with replacing traditional comparative concepts like "adjective" with broad functional comparative concepts like "properties" is that languages may have different formal behaviour

[2]Here, "information packaging" refers to the discourse organization of semantic content within an utterance. It is not of central importance to the present discussion.

[3]These sources summarize the emerging literature around semantic maps and standardize terminology; the method was developed over gradually over a few decades by a number of linguists, as described in the referenced passages.

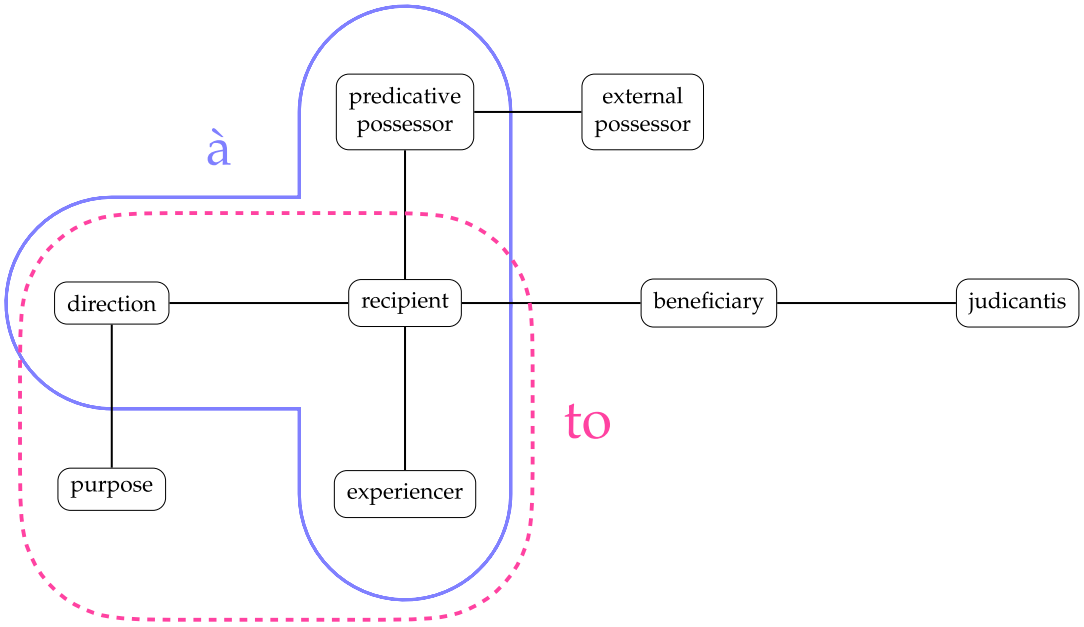

Figure 2.1: An example semantic map for the dative domain, adapted from Haspelmath (2003). Nodes represent different functions which "dative-like" elements can express. The boundaries for English *to* and French *à* are shown in pink and blue, respectively. Both terms cover contiguous regions of the map, satisfying the Semantic Map Connectivity Hypothesis.

for different properties. As such, we may need to define more fine-grained functional comparative concepts. But what if we are still interested in a broader function like "properties"?

Semantic maps offer a solution by decomposing broad functional domains into a network of finer-grained functions, which can then be related to one another based on their observed co-expression[4] across languages. Figure 2.1 shows an example of a semantic map. To construct such a map, the functional categories are first selected on the basis of whether at least one pair of languages differ in their expression of a function. English uses *to* both for direction ("I'm go-

[4]That is, using the same form or construction in a language. English co-expresses singular and plural second person as *you*.

ing to the store") and purpose ("I'm leaving early to be on time"), while French uses *à* for direction but *pour* for purpose. Thus, "purpose" must be a distinct functional category. The functional categories are then organized into a graph structure based on co-expression, with the aim to make language-particular strategies correspond to connected subgraphs of the semantic map. This desideratum has been termed the Semantic Map Connectivity Hypothesis (Croft, 2001, p. 96), and has also been claimed as a universal property of human language. Designing maps to satisfy this property has the following corollary for the resulting map: if function A and function B are co-expressed in a language, and there is no path between A and B on the map that does not pass through function C, then C must also be co-expressed with A and B in that language. Croft (2001) calls this resulting graph the *conceptual space*, and this is the resulting language universal that is claimed by a particular semantic map analysis. On top of the conceptual space, we draw semantic maps for particular languages, which show the subgraphs of the conceptual space that are co-expressed in that language. The conceptual spaces, then, represent cross-linguistic *constraints* on the application of strategies to express particular functions cross-linguistically, thereby providing a comparative bridge between form and function. Semantic maps can be very useful for expressing the generalizations about problematic, broad hybrid comparative concepts like "dative" or "adjective."

The semantic map method and the semantic map connectivity hypothesis underlying it have been extremely fruitful in identifying cross-linguistic co-expression patterns and universals. While conceptual spaces are not based on semantic similarity *per se*, rather facts of co-expression, in many domains where the semantic map method has been applied, the resulting conceptual spaces align closely with semantic similarity. In Section 2.3.4, I will describe how computational methods have built on the theory and practice of semantic maps, as well as some of the limitations of existing approaches.

Retro-definitions Haspelmath (2021) proposes a radical approach to hybrid comparative concepts, which he terms RETRO-DEFINITION. In this approach, common but potentially problematic comparative concepts ("traditional comparative concepts") are maintained as terms, but re-defined in a way that is cross-linguistically straightforward to operationalize and closely matches the traditional term. In this way, it represents a radical acceptance that comparative concepts need not relate to any "true" categories of language. As an example, Haspelmath proposes retro-defining adjectives as "property roots", regardless of their syntactic behaviour in a given language. Even more radically, he proposes retro-defining "inflection" as morphemes that express a fixed set of meanings cross-linguistically, and "derivation" as any kind of word-formation process that expresses any other type of meaning (Haspelmath, 2024). This has the effect of allowing the precise usage of these terms in cross-linguistic comparison, but whether these definitions provide the most *useful generalizations* about language is an open question.

Prototype theory and fuzzy categories Another conceptual approach to dealing with the challenges of defining comparative concepts is to embrace the idea that categories are inherently fuzzy and gradient, and organized around a central *prototype*. This viewpoint was popularized in cognitive science by a series of seminal works by Eleanor Rosch, which demonstrated clear effects that people both agree which members of categories like "vegetable" or "furniture" are most prototypical, and that prototypicality influences processing (**?**). This finding was influential on early work in the development of cognitive linguistics, with theorists like Ronald Langacker and George Lakoff arguing that linguistic categories also have a prototype structure **??**, and has been proposed as a solution to conflicting cross-linguistic data about categories: unusual distributional behaviour is associated with less prototypical members of a category.

However, the approach has come under fire for failing to account for apparent category boundaries or being unfalsifiable when applied to distributional data (Newmeyer, 1999; Haspelmath, 2024).

In many instances, prototype models are like a crude version of a semantic map model, because the semantic map connectivity hypothesis has similar implications: the longer the path between functions in conceptual space, the less likely they are to be co-expressed. However, rather than providing a fine-grained map of functions, prototype models focus on identifying central features, and suppose that less prototypical members should be less likely to be co-expressed.

My approach: grounding comparative concepts The main approach in this thesis does not fall neatly into any of the above widely-discussed approaches to hybrid comparative concepts, but takes a heterogeneous set of inspirations from them. The approach of the thesis is to define empirical measures which capture continuous dimensions of formal and functional variations, and relate them to *existing* category operationalizations. While this approach will be described in more detail later in the thesis, especially in **??**, I will here briefly review how it relates to these existing approaches. Similar to the semantic map approach, I aim to take a complex, hybrid concept and decompose it into finer-grained dimensions. Because my measures are continuous, it shares with many versions of the prototype approach the idea that category membership is gradient. My approach takes from retro-definitions the idea that comparative concepts come from a linguistic tradition which needs to be questioned. However, rather than re-defining existing terms, I take an existing operationalization as a starting point, and investigate how well my empirical measures align with these operationalizations. In the thesis, I focus on standard, theory-neutral operationalizations from databases like UniMorph and Universal Dependencies. Because the relationship between my empirical measures and the operationalizations

is the object of study, the claim does not rest on the a-priori validity of the databases–we aim to study the relationship of these comparative concepts *as they are used*. Future work can and should see if different operationalizations align better with the empirical measures I define. This empirical grounding approach also takes inspiration from literature in computational typology, which I will review in Section 2.3, but to my knowledge this literature has not been directly invoked in the context of comparative concepts.

Debates around comparative concepts often focus on problematizing existing attempts (Croft, 2016, p. 379). With this approach, I aim to reverse the discussion, by quantifying the consistency of existing operationalizations with respect to my empirical measures. In the context of this thesis's focus on *lexicality*, this means relating distinctions among word classes or between inflection and derivation to highly *abstract* empirical measures; in contrast to the semantic map approach or many of Haspelmath's retro-defintions, which focus on the *functions themselves*. In this way I aim to provide a new perspective on whether such distinctions really relate to the abstract properties which are often invoked to explain them, but have been difficult to measure directly. To do so, I rely heavily on emergent computational techniques for learning linguistic representations, which I will now describe in the next section.

2.2 Finding Meaning in Computational Models

Over the past two decades, deep learning models have revolutionized the field of natural language processing. These models come out of a history of brain-inspired cognitive modelling called *connectionism* (Rumelhart et al., 1986), which posited that important aspects of human cognition were best understood as the emergent behaviour of large parallel and distributed networks of simple processing units. In natural language processing, deep learning models have

demonstrated the ability to perform linguistic tasks at a level that would have once been thought to require human-level linguistic competence, like translating sentences or summarizing documents (Brown et al., 2020). While these models do not in-and-of-themselves provide a theory of human language processing, they provide a useful test bed for gradient and usage based theories of language (Futrell and Mahowald, 2025). In this section, I discuss several types of models, and the evidence that they acquire rich semantic and conceptual representations, and the importance of large-scale pretraining for this acquisition.

2.2.1 Type-level Distributional Embeddings

Much of the research on deep learning models of language is heavily indebted to the Distributional Hypothesis, which posits that a word's meaning is determined by the contexts in which it occurs (Harris, 1954; Joos, 1950; Firth, 1957). This hypothesis led to the development of distributional semantic models, which represent the meaning of linguistic units (typically, words) as a function of their co-occurrence patterns with other words in large corpora. A distributional semantic model typically represents each word as a high-dimensional vector (a point in high-dimensional Euclidean space), often referred to as the word's embedding. The application of neural networks to learn these embeddings led to a revolution in the field of distributional semantics in the early 2010s. The SkipGram or Word2Vec approaches (Mikolov et al., 2013a) learn vector representations of words by training a neural network to predict the context words surrounding a target word. The embeddings trained on this task are used as the word representations. With training on large corpora (consisting of millions or billions of words), these embeddings were quickly recognized to capture *semantic* similarity through their geometric properties: words which humans judge as similar tend to have embeddings which are close in vector space (Mikolov et al., 2013b). Further, these embeddings were found to capture

various types of semantic relationships through vector arithmetic: for example, the vector difference between the vector for *queen* is approximately equal to the vector for *king* minus the vector for *man* and plus the vector for *woman* (Mikolov et al., 2013b). These properties were improved upon by subsequent models like GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017), the latter of which incorporated distributional subword information to provide a back-off which is especially useful for representing rare words.

The rich semantic properties of distributional embeddings have been shown to have linguistic import. For example, Ettinger and Linzen (2016) showed that semantic priming effects can be predicted by embedding vector similarity, indicating that distributional information captured by these embeddings is predictive of human semantic processing. Distributional embeddings have also been shown to encode gradededness of properties of concepts–they encode the relative prototypical sizes of objects, as well as many other properties like speed, temperature, and gender (Grand et al., 2022).

A core hypothesis of the semantic map approach is that there is such a thing as universal conceptual similarity which is reflected through cross-linguistic co-expression patterns, and this literature has produced a rich body of findings supporting universal conceptual structure across a wide range of concepts (Youn et al., 2016; Haspelmath, 2003; Rogers, 2016; Regier et al., 2013). Distributional embeddings provide support for this view; the structure of embedding spaces learned from these co-occurrence patterns are similar enough cross-linguistically that embeddings trained on sufficient data can be largely aligned across languages using only simple linear translations (rotations, reflections, and scaling), even in an unsupervised manner (Vulić et al., 2020b; Hartmann et al., 2019). While of course, being word level representations, differences in lexicalization and polysemy patterns across languages limit the degree of alignment that can be achieved, these results nevertheless provide additional evidence for universal

conceptual structure underlying language.

The nature of SkipGram and related approaches implies that the embeddings capture *type-level* semantic information: each word has a single embedding, shared across contexts (this is often referred to as a STATIC embedding). This makes them especially well-suited for comparing to human experiments that don't provide utterance contexts, like the ones described above, or for modelling type-level data. However, this type-level nature prevents them from providing natural ways to capture *token-level* semantic variation. In Chapters 3–4, I use distributional embeddings because my data, like most morphological data, is type-level. Further, because at the word level, morphology can be thought of as a *relationship* between two types, I am able to study morphemes in combinations with many different roots, providing a diversity of contexts for each morpheme type, but not each word type.

2.2.2 Contextual Embeddings and Language Models

Researchers sought to overcome the type-level limitations of type-level embeddings like Word2Vec by developing CONTEXTUAL EMBEDDINGS, which provide a unique embedding for each token in context. Through developments in deep learning, larger and more complex models have been created to learn these embeddings, utilizing the transformer architecture (Vaswani et al., 2017). The most prominent approach to learning contextual embeddings is BERT, which uses a masked language modelling objective, learning to predict missing words in context (Devlin et al., 2019). This objective is used for PRETRAINING: fitting the model on an extremely large corpus for a task. From pretraining BERT model learns an embedding for each token in context, which provides a solid basis for FINE-TUNING on a range of downstream tasks; that is to say, training further on a smaller task-specific dataset and objective to create a better task-specific model. Examples of linguistically-oriented tasks where BERT embeddings

achieve new levels of performance include part-of-speech tagging and word sense disambiguation (Tenney et al., 2019b; Wiedemann et al., 2019; Chronis and Erk, 2020).

However, the transition to token-level embeddings produced new challenges for interpreting representations. Today's models are *deep*, meaning they have many *layers*, each of which provides a different representation of a token. It is not always clear which layer is the *right* one for a given task—representations become more contextualized at deeper layers (Ethayarajh, 2019), but features like parts of speech are better represented in early layers (Tenney et al., 2019a) Today's models are also *subword-level*—they represent rare words as sequences of multiple tokens, each representing a subword selected by a statistical learning algorithm (usually byte-pair encoding or a similar method; Sennrich et al., 2016), so the units which have vector representations are not always aligned with linguistic structure. Finally, distance between embeddings is dominated by *rogue dimensions*, a handful of dimensions of large magnitude and obscure semantic similarity (Timkey and van Schijndel, 2021). After better understanding these issues, researchers were able to extract rich semantic information from contextual embeddings by pooling across contexts and subwords (Bommasani et al., 2020; Eyal et al., 2022), standardizing dimensions (Timkey and van Schijndel, 2021), and selecting layers based on task. All this is to say, as modelling approaches have become more complex, it requires more care to identify and extract the semantic information encoded in these models. After such discoveries, contextual embeddings have been shown to capture interesting semantic information, about, e.g., different senses of *break* in English (Petersen and Potts, 2023) and constructions like *a beautiful five days* or *day by day* (Scivetti and Schneider, 2025; Rozner et al., 2025)

For the next class of models I discuss, identifying a semantic space is still an area of extremely active research. These are the so-called AUTOREGRESSIVE

LANGUAGE MODELS (or simply language models), which are trained to predict the next token in a sequence, given all prior tokens. These models have received enormous attention recently due to their ability to generate fluent and coherent text and solve complex tasks "in-context", meaning by sequentially generating to complete a sequence which includes instructions for the task (Brown et al., 2020). In line with these capabilities, these models are typically very large, with billions to hundreds of billions of parameters, and are trained on massive corpora of billions or trillions of tokens. Their impressive capabilities have led to a surge of interest in understanding what kinds of linguistic and world knowledge they acquire during training (Futrell and Mahowald, 2025).

Despite impressive performance, we are a ways off from a clear understanding of how these models represent meaning. Their autoregressive nature means the vectors "representing" a particular token are only conditioned on the *preceeding* context—which means sense information may be distributed onto later disambiguating tokens (e.g. in *break the law* vs *break the news*). Nevertheless, these models have been shown to be impressive predictors of incremental processing in humans: both of psycholinguistic performance through reading times (Staub, Forthcoming; Wilcox et al., 2023), and of neural activity during language processing (Schrimpf et al., 2021; AlKhamissi et al., 2025). Further, the nascent literature on interpreting the internal representations of these models has shown evidence for rich conceptual representations. Using feature and circuit extraction techniques like sparse autoencoders, researchers have identified highly abstract features, like an eye feature that responds to mentions of eyes across languages, and in code and ASCII drawings of faces and eyes (Tarng et al., 2025). Other work has identified circuits which copy abstract concepts even when words or tokens differ (Feucht et al., 2025) and shown linear gradability effects with vectors representing properties like BEAUTY (Kozlowski et al., 2025). Many of these models are *multilingual*, being able to generate text in a range

of languages, and there is some evidence that they use *shared* representations across languages, for relations among lexical concepts (Lindsey et al., 2025), and for traditionally "grammatical" or "morphological" concepts like tense, case, and number (Brinkmann et al., 2025).

Together, these results provide evidence that deep learning models of language acquire rich semantic and conceptual representations, at both the type and token levels. These representations can be extracted and studied to provide insights into human language processing, as well as the nature of linguistic meaning. These representations are also naturally gradient, making them well-suited for studying gradient theories of language and meaning, something that has been challenging with the traditional approaches in linguistics (Petersen and Potts, 2023; Futrell and Mahowald, 2025).

2.2.3 Vision-and-language models

While language models and embedding models have demonstrated impressive capabilities in acquiring semantic representations from text alone, their basis in a strong form of the distributional hypothesis makes fully segregating form and function difficult. Because the aim of typology is to identify cross-linguistic generalizations about how different languages express the same communicative functions, typologists usually aim to identify functions which are as form-agnostic as possible (functional constructions) for at least some types of cross-linguistic comparison. In the second part of this thesis, I aim to bring this spirit into computational approaches to typology, by utilizing recent advances in VISION-AND-LANGUAGE MODELS (VLMs). These models can process a combination of visual and textual input, allowing them to model scenarios where language use is grounded in and modulated by visual perception–such as visual storytelling, visual question answering, and image captioning (Lin et al., 2014; Antol et al., 2015; Huang et al., 2016). In this way, VLMs provide a

way to combine the computational power of language models with a language- and form-agnostic source of meaning/function: the visual modality.

While VLMs have been the subject of less interpretability[5] research and linguistic analysis than pure language models (due to the rapidly changing nature of the field), the evidence so far suggests a similarly rich picture to the other models of language discussed in this section. VLMs have been shown to exhibit better understanding of hypernyms and categorical structure than pure language models (Qin et al., 2025), to possess units that respond to the same concept presented in different visual forms (e.g. images, text, and drawings; Goh et al., 2021), and to better predict hippocampal activity for "concepts" than unimodal models (Choksi et al., 2022). VLMs have also been shown to share representations *both* across languages *and* across the visual and textual modalities (Wu et al., 2025), and to acquire representations of lexical similarity that align with human judgements (Yun et al., 2021). These results suggest that VLMs have rich internal representations which combine information from both modalities, and that these representations can be studied to provide insights into human language processing and meaning. I will now provide a high-level overview of how these models are constructed to provide context for their use in Part II.

Making a vision-and-language model

Today's state-of-the-art VLMs are typically built out of two key components: a VISION ENCODER, which processes the visual input, and a LANGUAGE MODEL, which produces text output conditioned on the visual input. Figure 2.2 shows a schematic of this architecture, in the process of generating a caption for an input image. In this thesis, I use the PaliGemma VLM (Beyer et al., 2024), which follows this general approach.

[5]Interpretability is the field of research which aims to understand the representations and processing of machine learning models.

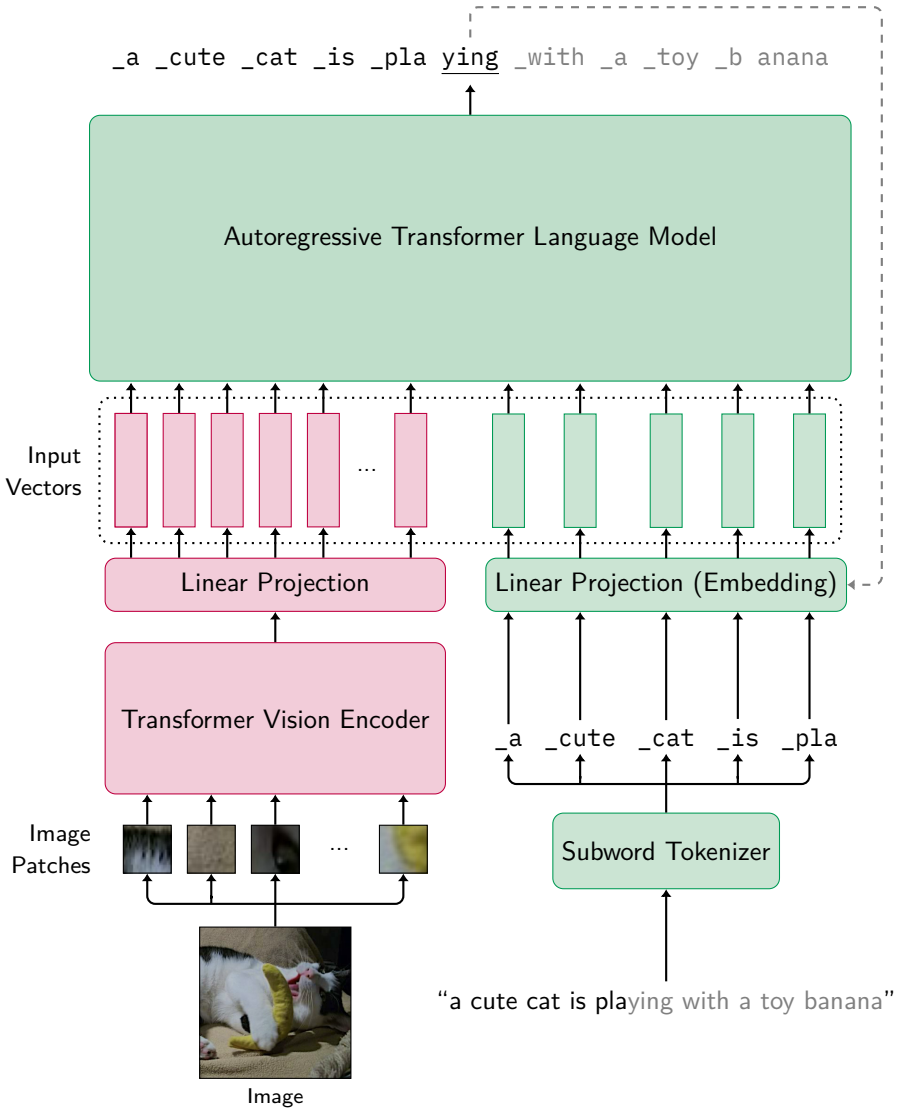

Figure 2.2: A typical vision-and-language model architecture, in the process of captioning an image. A vision transformer produces a representation of the input image, which is linearly projected into embedding vectors for an autoregressive transformer language model. The model generates text one subword token at a time based on the image and the preceding tokens. Here, it has generated the token `ying` as the next token after `_a _cute _cat _is _pla`. The token `ying` will be added to the input at the next time step to continue generating a caption.

The vision encoder in a VLM typically uses a vision transformer to provide a high-dimensional representation of an input image. A vision transformer is a variant of the transformer architecture (Vaswani et al., 2017) which can process images by dividing them into patches which are embedded using a learned linear transformation and then processed similarly to tokens in a text transformer (Dosovitskiy et al., 2020). This model is usually trained together with a transformer text encoder on an extremely large dataset of image–caption pairs. In training, the model maximizes the similarity between embeddings of images and their corresponding captions, while minimizing the similarity of embeddings of non-corresponding image–caption pairs (Radford et al., 2021; Zhai et al., 2023). After training, the vision encoder can be used independently to produce high-dimensional vector representations of images without the need for captions.[6]

The language model component is typically initialized from a pretrained autoregressive transformer language model of the kind discussed in Section 2.2.2. The vision encoder is connected to the language model in the following way: the output of the vision encoder (a high-dimensional vector) is linearly projected into a series of vectors which are used as "prefix" tokens to condition the language model's text generation (Tsimpoukelli et al., 2021; Karpathy and Fei-Fei, 2017).

With continued training on image–text pairs for a variety of tasks (image captioning, optical character recognition, visual question answering, etc.) the output of the vision encoder is adapted to condition the language model's text generation. In this way, the language model can generate text which is grounded in the visual input. This approach also means that most VLMs *include* a language model which can be used independently of the vision encoder, a fact which I exploit in Chapter 5.

[6]Vision encoders used to be trained without text encoders, using objectives like image classification (Deng et al., 2009), but the use of contrastive learning with text has been shown to produce substantially better representations for a diverse range of target tasks (Radford et al., 2021).

2.2.4 Rich representations from large-scale pretraining

In this section, I have reviewed several classes of deep learning models of language, highlighting key similarities and difference among them as pertains to their representation of semantic information. Distributional embeddings like Word2Vec and FastText leverage collocational patterns to learn a high-dimensional vector space where geometric relationships between vectors correspond to semantic similarity. Contextual embeddings like BERT extend this approach to provide token-level representations which can capture contextual nuances and constructional meanings, but require more careful interpretation to extract conceptual semantic information. Autoregressive language models like the GPT family learn to predict the next token in a sequence, acquiring representations that solve complex tasks and are especially useful for modelling human language processing, but their internal representations are still not well understood. Finally, VLMs combine visual and textual information to ground language in perceptual content, preserving the strengths of language models while providing a form-agnostic source of meaning.

The major theme of all this research, and indeed of most of the research in natural language processing in the twenty-first century, is the value of *large-scale pretraining* on massive datasets (Saphra et al., 2024). In between each of the approaches discussed here, there were many intermediate models and data-efficient approaches proposed, but these models achieved their success and long-standing impact through pretraining on large corpora. The advent of the fine-tuning and few-shot learning paradigms further cemented the hegenomy of pretrained models, as increasingly even small-data tasks are best solved by leveraging large pretrained models (Brown et al., 2020; Devlin et al., 2019).

Not only has large-scale pretraining been the key to improving performance on natural language processing tasks, but it has also paralleled findings in typology about the universal conceptual structure driving differences and similarities

across languages. In the same way as coexpression patterns reveal a conceptual space respected across languages, sufficiently large-scale pretraining on datasets induces isomorphic word vector spaces across languages (Vulić et al., 2020b), and produces convergent representations across different model architectures, languages, and modalities (Jha et al., 2025; Huh et al., 2024; Brinkmann et al., 2025). Indeed, a growing body of work suggests that representational alignment between deep learning models and *the human brain* is driven primarily by the discovery of a shared, universal conceptual structure (Kauf et al., 2024; Hosseini et al., 2024; Chen and Bonner, 2025; Antonello and Huth, 2024). And on top of this, the models that result from large-scale pretraining have been shown to be sensitive to extremely fine -grained semantic and distributional distinctions (Petersen and Potts, 2023; Rozner et al., 2025; Goldberg, 2024), making them exceptionally rich sources of linguistic representations.

Therefore, I argue that the representations learned by large-scale pretrained models provide a promising avenue for studying function in typology. While the large-scale pretraining paradigm presents challenges and limitations for typology (as most of the world's languages lack sufficient data for current methods), the linguistic capabilities of these models will be essential for bringing a stronger empirical basis to the notion of semantic contentfulness in this thesis, and presents exciting avenues for future work in typology more broadly. In the following section, I will review how comparative concepts have been used in computational typology to date, and how the representations learned by large-scale pretrained models can provide new avenues for defining, studying and operationalizing functional comparative concepts.

2.3 Approaches to Comparative Concepts in Computational Typology

This section reviews the roles that comparative concepts have played in computational approaches to linguistic typology. In this section, I aim to highlight how technological advances can enable new types of comparative concepts, and how the choice of comparative concepts can shape the types of questions that can be asked and generalizations that can be formed. In particular, I will argue that computational modelling thrives when fine-grained distance metrics are defined, be this in a high-dimensional discrete space or especially a continuous space.

2.3.1 Comparative concepts in multilingual databases

The predominant approach to comparative concepts in computational typology has simply been to follow the lead of whatever annotation scheme is used in the typological or cross-linguistic database used in the study. For example, influential works like Futrell et al. (2015), which demonstrated minimization of dependency length as a universal pressure on word order, used the data of Universal Dependencies (**?**), as-is. Similarly, cross-linguistic studies of morphological complexity and inflectional paradigms have relied on a combination of feature values from databases like the World Atlas of Language Structures (WALS) (**?**), or the encodings of grammatical features in UniMorph (Batsuren et al., 2022).

Databases like UniMorph and Universal Dependencies attempt to use a single cross-linguistic annotation scheme for their comparative concepts, but these schemes usually much more closely resemble the hybrid categories of language-particular analyses. For example, UniMorph's feature set includes values like Tense=PAST, with no information about what constructions that

form is used in. Chapter 3 includes a detailed discussion of some limitations of UniMorph annotations, which are largely based on language-specific grammatical traditions rather than typological best practices. Similarly, Universal Dependencies part-of-speech tags are cross-linguistically "universal", but they are deployed with the "methodological opportunism" described by Croft (2002), where categories are defined in a language-particular way. Very recently, Universal Dependencies has begun a process of aligning their representational scheme with more fine-grained and cross-linguistically valid comparative concepts (**??**), but this process is still ongoing.

That the annotation of these databases is *not* consistently based on cross-linguistically valid universal comparative concepts actually provides an interesting opportunity, however, and one which this thesis exploits. A major aim of this thesis is to define computational comparative concepts of dimensions hypothesized to underlie grammatical category distinctions cross-linguistically, and then investigate how these dimensions relate to the categories used in these databases. In this way, I can investigate the extent to which these databases distinctions, while flawed, nevertheless align with the underlying dimensions of meaning that motivate these distinctions.

2.3.2 Phonological typology

Phonology offers the earliest and clearest example of how empirical, continuous measures can advance typology. It has long been recognized that language-specific categories like phonemes are ill-suited for cross-linguistic comparison ("phonemes are not fruitful universals"; **?**) because they are defined by language-internal contrasts. To address this, **?** introduced a *feature-based* account of phonetic universals, later refined by **?** through the notion of *natural classes*.

Vowels have long been central to typological inquiry. **?** proposed early implicational universals about vowel systems within his proto-featural frame-

work. Yet within this featural approaches, universals of vowel systems were understood as complex, varying substantially on the number of vocalic contrasts within a language. The true generalizations, which turn out to be extremely simple when properly understood in a continuous space, required substantial developments in the acoustic theory of vowels. The decisive shift came with the formant theory (**??**), which linked vowel quality to acoustic resonances (F1–F3). The development of better technologies for measuring and recording formants through the twentieth century ultimately provided a real-valued acoustic representation (**?**) that allowed vowels to be compared across languages in a shared empirical space, enabling the development of theories that made quantitative, testable predictions. The quantal theory (**?**Stevens, 1989) proposed that languages prefer perceptually stable regions of this space, while the dispersion theory (Liljencrants et al., 1972) modelled vowel inventories as systems maximizing perceptual distance. By simulating optimal vowel systems and comparing them to attested inventories, these models offered a precise computational account of typological tendencies. Subsequent dispersion–focalization models (**??**) and probabilistic analyses of entire vowel-system distributions (**?**) further refined these predictions, revealing the relative influence of competing pressures such as distinctiveness and perceptual stability.

While vowel typology concerns formal acoustic dimensions, its trajectory exemplifies a broader lesson: defining an empirical, continuous underlying space can transform typological theory. Just as formant space enabled precise and falsifiable generalizations about vowel systems, deep learning models may provide an analogous empirical grounding for meaning. Linking model-derived semantic spaces to human cognition could allow typology of linguistic function to achieve the same level of quantitative precision. In the next section, I turn to evidence from semantic category systems supporting this view.

2.3.3 Semantic category systems

The domain most closely paralleling vowel typology in its treatment of function is the study of semantic category systems—cross-linguistic analyses of how languages partition a shared underlying semantic space. As with vowels, researchers model these systems as optimizing trade-offs among universal pressures such as simplicity, communicative efficiency, and learnability. Once an underlying space is defined, these pressures can be formalized, simulated, and quantitatively tested against attested systems.

The case of colour terms provides the clearest illustration. In their seminal study, **?** identified robust implicational hierarchies—two-term systems distinguishing light from dark, three-term systems adding red. Berlin and Kay couched these generalizations in terms of an implicational hierarchy of colour terms; however, their methodology could offer little insight into *why* these hierarchies exist. The breakthrough came with the development of a precise underlying perceptual space for colour. Studying the relationship between the colour space in terms of frequency and perceptual distance, Later work showed that a continuous perceptual space was key. Building on the CIEL*a*b* colour space (**?**), which aligns physical and perceptual properties of colour, Regier et al. (2007) modelled how systems maximize within-category similarity and between-category distinctiveness. These simulations closely predicted attested colour inventories and revealed that real systems are significantly more optimal than chance. This initial effort has expanded into a rich literature, modelling trade-offs among different pressures such as communicative need, perceptual structure, and learnability. While this remains a more rapidly evolving area of research than vowel system typology, studies continue to refine hypotheses and distinguish between pressures with increasingly fine-grained predictions about colour systems.

Comparable approaches have since been applied to other semantic do-

mains—kinship, number, quantifiers, modals, and pronouns—where discrete conceptual structure allows for tractable mapping. More challenging are domains with continuous meaning spaces, such as spatial terms (**?**) and tense–aspect marking (**?**), which have required simplifying the space into coarse discrete categories or low-dimensional projections (e.g. multidimensional scaling of usage data; **?**). Therefore, work in these frameworks has primarily focused on domains where an underlying conceptual space can be more straightforwardly defined.

Overall, both the study of vowel systems and semantic category systems show that by creating an operationalization of an underlying space, we can test and discover parsimonious and predictive theories of the fundamental data of linguistic typology. In each case, substantial typological progress was able to be made without access to the "true" space—we are still learning about the perceptual dimensions of vowels (), and CIEL*a*b* has known shortcomings ()–but the development of some empirical model of the underlying space that could be coded across languages was critical for this progress. However, the limited scope of functions studied in these types of computational frameworks points to the challenge of defining an underlying space for more complex semantics. In this thesis, I argue that recent advances in deep learning models of language are likely an early step in this direction, analogous to the early stages of development of the formant theory of vowels.

2.3.4 Multidimensional scaling

The multidimensional scaling (MDS) approach to semantic maps proposed by Croft and Poole (2008) represents the most well-developed technique for modelling a continuous functional space for typological comparison. This approach was developed to address the challenge of translating large typological

datasets into a traditional semantic map following the methodology of **?**,[7] and to provide a stronger mathematical basis for the semantic map theory. These methods take as input a high-dimensional discrete matrix of linguistic data, and produce a low-dimensional continuous representation of the data, using either optimal unfolding, or in some studies, matrix decomposition techniques. The resulting map provides an approximation where Euclidean distance approximates the frequency with which two functions are expressed with the same form. In this way, an underlying semantic/conceptual space is being inferred from typological data about form—representing a move away from the discrete identification of functional comparative concepts to a richer emergent empirical representation of the complexities of meaning, towards the desiderata of this thesis. Studies vary primarily in the way they construct the input dissimilarity matrix, and thereby in how much they allow for an emergent representation of function. van der Klis and Tellings (2022) provide a recent overview of the different methods used in the literature, which I briefly summarize here to illustrate how differing approaches rely on different comparative concepts. They provide a three-way typology of approaches to constructing the input for an MDS analysis.

First, there is the classical input representation, which Croft and Poole (2008) used to recreate **?**'s analysis of indefinites.[8] This type of map takes a set of N linguistic forms $\mathcal{F}$, and a set of K underlying functions $\mathcal{M}$. The binary input matrix $\mathbf{I} \in \{\mathrm{Y},\mathrm{N}\}^{K\times N}$ is constructed such that

$$\mathbf{I}_{i,j} = \begin{cases} \mathrm{Y} & \text{if form } f_i \text{ from language } l \text{ conveys (or is used to express) function } m_j, \\ \mathrm{N} & \text{otherwise.} \end{cases}$$

to which the optimal unfolding technique is applied. In this type of analysis, the functions are entirely manually posited by the typologist and abstracted

[7] While an algorithm for producing graph-based maps from large-scale data was later introduced by **?**, the MDS techniques retain a number of advantages.

[8] Other examples of this type of map include...

away from the constructions on which the functional claim is based. As a result, the MDS analysis cannot capture fine-grained variations around the prototypes of a given abstract function, but can capture differences in the closeness of two functions in a more fine-grained way than the classical approach to semantic maps—frequency of form-function co-occurences is modelled. Nevertheless, in terms of comparative concepts, the functions here retain the traditional approach to function in typology, with its known shortcomings.

Croft and Poole (2008) also introduce a second method for producing an MDS map, which allows more fine-grained study of the prototype structure of functions. This second map relies on **?**'s tense-aspect data. Here, rather than manually determining in a binary manner whether a particular linguistic form can or cannot encode a particular function, a range of specific constructions are included in the analysis. In **?**, informants across languages translated sentences in a specific temporal and observational context (e.g., you saw someone writing a letter yesterday). These data were assigned to tense-aspect prototypes, so the input matrix $\mathbf{I}$ now has K sentential contexts $c_j \in C$, belonging to a smaller number of abstract function prototypes, and takes the following form:

$$\mathbf{I}_{i,j} = \begin{cases} \text{Y} & \text{if form } f_i \text{ was used for sentential context } c_j \text{ in language } l, \\ \text{N} & \text{otherwise.} \end{cases}$$

The lessened reliance on manually posited functions allows for a the prototype structure to emerge from the data. However, the contexts are still manually selected by the typologist, and the semantic information still only comes from co-occurrence with forms in the sample itself–so the study necessarily cannot represent the full complexity of the studied forms across the vast space of possible meanings, nor can it fully capture frequency effects. This type of study has also been applied to other aspectual constructions (**?**), and to verb-specific semantic roles (**?**).

Finally, the third major type of MDS analysis relies on a fully bottom-up

approach to function, using parallel corpora rather than reference grammars or elicited survey data. In this type of study, relevant parallel clauses for a particular phenomenon are identified in a parallel corpus, and the input matrix **I** is constructed such that each row contains the construction used in that clause in each language studied, producing a K-tuple where K is the number of languages. To compute distance in this type of study, the Hamming distance between the tuple for two clauses is computed, yielding a similarity matrix based on the number of languages that use the same construction in both clauses. This type of analysis removes the manual positing of functions and contexts entirely, proceeding bottom-up, and is thus the most in the spirit of our present inquiry. However, the approach still only captures similarity based on translations in corpora. The fact that contemporary deep learning models capture extremely fine-grained semantic distinctions is not leveraged in this approach, and so the semantic space captured is only as good as the evidence directly given by the co-occurrence of translations, rather than language-internal evidence about meaning.[9]

Overall, the MDS approach allows us to both study the rich gradient structure underlying linguistic function, and decrease the dependence on manually posited functions. However, the state of the art still relies entirely on parallel co-occurence data. In the next section, I will discuss the small literature that leverages recent advances in deep learning to provide a rich representation of function in typology.

2.3.5 Deep learning models of comparative concepts

Despite the rich body of evidence that deep learning models of language capture fine-grained semantic distinctions, there has been relatively little work leveraging these models to provide empirically grounded comparative concepts

[9] A wide range of domains and phenomena have been studied with this approach:

for typology. Recently, Gregorio et al. (2025) used multilingual BERT (Devlin et al., 2019) and Aya (**?**) to study animacy cross-linguistically. Specifically, they identify which syntactic roles and clausal positions are most associated with animacy of the referent. However, the role of the models used here is not truly gradient, nor is the function emergent–the models are used to produce a 3-way classification (human, animate, and inanimate) based on an annotated corpus, and the analysis is conducted over these discrete categories. While the rich representations of the models are critical for creating an accurate classifier, the comparative concepts are still discrete and manually posited.

Papadimitriou et al. (2021) study grammatical subjecthood with a less discrete approach. Specifically, they train a multi-layer perceptron classifier on multilingual BERT representations to distinguish between the embeddings of transitive subjects and objects, then examine the classifier's categorization of intransitive subjects, finding that intransitive subjects are categorized as more subject-like than object-like, and that classifiers transfer across languages, including languages with different morphosyntactic alignment (e.g. ergative-absolutive vs. nominative-accusative languages). However, they found that animate non-subjects and passive subjects were more likely to be classified as subjects and objects respectively, indicating a semantic dimension to this cross-linguistically robust representation of subjecthood.

Another technique for obtaining a gradient representation of a semantic dimension is to use *semantic projection* (Grand et al., 2022), which has been shown to capture human judgements about object features. This technique uses exemplars at extremes of a semantic dimension (e.g. "huge" and "tiny"), using a deep learning model to embed them in a Euclidean space. All possible embedding pairs across the two sets of exemplars are subtracted from each other, and these difference vectors are averaged to produce a single vector representing the semantic dimension. This vector can then be used to project other words

onto this dimension by taking the dot product of their embedding with the dimension vector. Li (2025) uses this technique to study models' representations of animacy. The authors claim that their results show that models represent animals as more animate than humans, in line with psychological findings in humans (). They suggest that this indicates inductive biases in humans that shape grammatical animacy by focusing on certain constructions. However, their operationalization of animacy is questionable, as the exemplars they use to define high animacy are exclusively non-human animals. Nevertheless, the techniques here show how deep learning models can be used to provide a gradient representation of a semantic dimension which can be used to study cross-linguistic patterns in form-function mappings.

Altogether, the results in this nascent literature are promising, but there are still many dimensions of deep learning representations that have not been explored. In this thesis, I will focus on how deep learning models help provide new and better models of lexicality, which has so far not been directly addressed in any of this literature.

2.4 Formal and functional dimensions of lexicality

> ...we may be quite sure of the analysis of the words in a sentence, and yet not succeed in acquiring that inner "feel" of its structure that enables to tell infallibly what is "material content" and what is "relation"
>
> — Edward Sapir (**?**)

Some units of language are more meaningful than others. This basic insight is almost as old as the study of language itself. In the Greek tradition, Aristotle distinguished *phōnē sēmantik* (sign-bearing sounds) from *phōn ásēmos* (non-sign-bearing sounds), such as the class of *árthron* which includes prepositions and preverbs (**?**). This distinction was not limited to the proto-linguistics of Indo-European languages: in the 12th century the *Wén zé ( )* of Chen Kui ( )

catalogued *zhùchí* ⊠⊠ (lit. "helping words")–corresponding to what we would today call function words. Across the world's languages, we see asymmetries between elements that express content and those that express grammatical function. It is little wonder then that the distinction between contentful and functional elements continues to have relevance across linguistic theories and domains. Yet boundary cases abound and the nature of the distinction has made it challenging to formalize. In Chapter 1, I sketched the idea of a LEXICALITY SPECTRUM as a general way of conceptualizing the correlation between formal expression and (degree of) semantic content. In this section, I describe the formal and functional dimensions of this spectrum in more detail.

2.4.1 The formal dimension

As I argued in Chapter 1, the lexicality spectrum is operant at multiple levels of formal linguistic structure, with different names being used for similar distinctions at different levels. But the distinctions between "words", "clitics", "morphemes", and "affixes" are all theoretically tenuous at best (Zwicky, 1994; Bruening, 2018). As a theory of these terms themselves is beyond the scope of this thesis, I will here focus on the general formal trends that underlie these distinctions—the formal dimension of the lexicality spectrum.

The basic idea underlying all these terms is this: some linguistic units have "bigger" forms than others. Of course, this is implied by the compositional nature of linguistic structure: a phrase may be composed of several words, a word may be composed of several morphemes, and each morpheme can contain a variable number of phonemes. Yet even comparing morphemes to morphemes, some are formally bigger than others. Here are some ways in which this can manifest:

Boundness One aspect of formal size is the notion of BOUNDNESS. While Haspelmath has argued for a sharp cross-linguistic definition of boundness **??** as "unable to occur in isolation", I share **?**'s scepticism of the utility of this as a cross-linguistic criteria and share his feeling that this is better understood as a gradient notion. There are many languages where no morpheme can occur in isolation—surely our comparative concepts should apply to them! Further, the notion of "isolation" is itself problematic, as language always occurs in a discursive context. Here, I sketch boundness as a continuum of cluster properties. In Chapter 3, I operationalize some relevant aspects of this continuum, but here I will simply focus on what the formal trends *are*.

At one end of the spectrum of boundness are free morphemes. In many languages, these morphemes can form whole utterances by themeselves in the right context (Consider "Cat." as a response to "What is your favourite animal?"). In some languages, even the freeest morphemes may not be able to stand alone, requiring some obligatory bound marking (e.g. case or tense marking), but the free morpheme behaves in some way like the "root" of the word. This often takes the form of the free morpheme occurring at the periphery of the word (usually the beginning).[10] Further, they typically occur immediately coincident to that host morpheme. If morphemes occur between a bound morpheme and its root, those morphemes are typically intermediate in terms of these formal properties–that is, the most bound morphemes occur furthest from the root morpheme.

Boundness is a similar notion to VALENCY, the notion that certain linguistic units require a certain number of arguments. For example, nouns typically have a valency of zero. Most verbs, on the other hand, have a valency of one or more, requiring a subject and one or more objects. Syntactic valency is different

[10] In cases where the free morpheme cannot stand alone, a critical aspect of the argument for the *freeness* of this morpheme is typically semantic. However, I am here focusing exclusively on the *formal* properties of boundness.

from semantically valency. For example the verb *to rain* arguably requires no semantic arguments in English, but it still requires a subject syntactically (*It rains*). Putting things in mathematical terms, we can view such words as functions, which require certain arguments to form a complete expression. Similarly, bound morphemes are often formalized as predicates which take in a root morpheme to produce a combined expression. However, prototypical free but valent morphemes (e.g. verbs) are distinguished from bound morphemes by their degree of syntagmatic integration with their arguments. Either the arguments are themselves free morphemes, able to move around depending on the construction or take their own bound morphemes, or else they are expressed through bound morphemes on the verb itself.

Allomorphy More bound forms are also phonetically more variable. They may be subject to special phonological processes that do not apply to other morphemes in the language (such as the English plural -s being realized variably as [s], [z], or [ɪz] depending on the phonological context). The more of these processes apply, the more phonologically bound the morpheme is. **?** argue that lexically- or morphologically-conditioned variants of morphemes (*allomorphs*) are a formal sign of greater bondedness.

Length Perhaps the most obvious dimension of formal size can be seen in the number of phonemes in a morpheme—which can vary dramatically. The length parameter, at the short end, is intimately tied up with the other dimensions of formal size. Allomorphy may reduce the number of shared phonemes between morpheme variants. Morphemes can also get a length shorter than one phoneme in some instances. *Portmanteau* morphemes share multiple (unrelated) features in a single marker, meaning that the phonological material dedicated to any one of them can be the equivalent of less than a single segment. Tightly bound morphemes can become shorter than a segment by becoming suprasegmental

or process morphemes, e.g. by changing tone or root morpheme vowel quality (*Ablaut*). We can therefore think of all dimensions of formal size outlined here as related to the concept of length.

2.4.2 The functional-semantic dimension

Information and Frequency At the semantic core of the lexicality spectrum is the notion of CONTENTFULNESS.[11] The basic intuition is that some linguistic items contribute more to the overall meaning of an utterance than others. This is a notoriously difficult notion to pin down, as it relates to the deep question of what *meaning* and *content* are in the first place. INFORMATION THEORY both provides a mathematical formalization of this notion and a demonstration of why separating content from form is difficult. Shannon (1948) introduced the notion of *entropy* as a measure of information in bits. The entropy of a random variable X with possible values $\{x_1, x_2, \ldots, x_n\}$ and probability mass function $P(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i).$$

The entropy is related to the notion of "information" in an important respect: it provides a lower bound[12] on the number of bits needed to encode the value of X. That is, it tells you the most efficient encoding/compression of X must be at least $H(X)$ bits long. Shannon's entropy and information-theory more broadly have been widely and successfully applied to linguistic theory (). With respect to "words"/"morphemes", the information-theoretic perspective implies that the information content of a morpheme is its negative log-probability.[13] Under this perspective, more frequent morphemes carry less information. Combining this with the idea from Section 2.4.1 that the formal dimension is all related to length

[11] The terms "semantic weight" and "semantic force" are also common

[12] Technically, an average lower bound over many samples.

[13] This follows from the assumption that morphemes are generated independently, which is not true, however, later work has extended the information-theoretic insight to account for dependencies, with a similar overall conclusion (Piantadosi et al., 2011).

or formal size, we can see the lexicality spectrum as a generalization of Zipf's law of abbreviation, one of the most famous and foundational discoveries in all of computational linguistics (**?**), which states that more frequent words are much shorter than less frequent ones. The information-theoretic corallary of this is that more informative morphemes are longer than less informative ones. Frequency certainly plays a central role in the structure of language and the lexicality spectrum. It has been argued to be a driving force in grammaticalization (Bybee, 1985), and the so-called iconicity of complexity (itself closely related to the notion of a lexicality spectrum) has been argued to be an effect of frequency pressures on efficient communication (**?**). Nevertheless, debates around the nature of the lexical–functional distinction, inflection–derivation distinction, and related phenomena continue unabated, indicating that frequency and thus standard information-theoretic notions of content are insufficient as a full account of the relationship between meaning and function.

Abstractness and Polysemy Another common way of characterizing the functional dimension of the lexicality spectrum is in terms of abstractness. Functional elements are often described as being more abstract than lexical elements; on the other hand, functional elements are also often more polysemous than lexical elements (Haspelmath, 2003). The difference between abstractness and polysemy is not always clear. In Figure 2.1, it is fairly clear that *to* has multiple senses, such as the purpose sense in *I went to the store to buy milk* and the recepient sense in *I gave the book to Mary*. Haspelmath (2003) points out that data is often ambiguous between a *monosemic* position where there is a single vague, abstract meaning that interacts with contexts to serve different functions, and a *polysemic* or even *homonymic* interpretation where there are multiple more specific meanings which share a surface form for motivated or ideosyncratic reasons. Under such an interpretation, the meaning of functional elements is

not vague at all, and perhaps not very abstract. Nevertheless, words like *in* which have spawned whole literatures in linguistics and psychology attempting to characterize their use, and which seem to cover a continuous space of meaning rather than discrete scenarios, are more straightforwardly characterized as abstract than polysemous.[14]

Another problem for abstractness of meaning is that lexical elements can also be highly abstract (e.g. *idea*, *angry*) and some traditionally functional elements can be thought of as being fairly concrete (e.g. plural markers). This problem is sometimes waved away by invoking prototypicality, but such an account leaves some serious questions unanswered (**?**Croft, 2002, p. 225). **?** notes that *concrete* nouns which are frequent do get shorter forms (e.g. *dog*, *car*), indicating that length is more a function of predictability than abstractness. Nevertheless, while these forms are shorter, they do share many formal properties with prototypical concrete lexical items, such as their lack of boundness. So there is still something *different* about these forms compared to functional items, even if abstractness is not the right way to capture it.

Relationality Perhaps one way out of this conundrum is to focus on semantic RELATIONALITY rather than abstractness *per se*. In the simplest terms, a relational meaning is one that inherently implies the existence of at least one other entity (**?**, p. 67). For example, the concept ROUND is relational, as roundness can only be defined with respect to some entity. On the other hand, CIRCLE is non-relational, despite referring to the same properties. In my view, a key property of relationality of meaning is that when composed with entities, the resulting meaning is less abstract than the relational meaning alone. For example *round* is more abstract than *a round rock*. This can help explain some of the more "concrete" grammatical functions: while plural *forms* are not very abstract (e.g.

[14]I refer here to the spatial sense(s) of *in*.

cats is likely to be very concrete), the plurality is itself highly abstract.

There are several barriers to associating relationality with the functional dimension of the lexicality spectrum. A first objection is that because verbs and adjectives are both traditionally considered lexical and relational, relationality cannot be the defining property of functional items. In Part II of this thesis, I will argue that relationality is a key dimension for shaping the lexicality spectrum, and that this *includes* adjectives and verbs as in some sense closer to functional elements than nouns. I believe the definition of relationality I have given here helps explain this: prototypical *lexical* relational concepts such as GIVING, ROUND, or RED are, I believe, more concrete and more full of intension than functional relational concepts such as PLURALITY or ANIMACY. A key correlate of this objection is confusing relationality with semantic valency. The number of entities required to specify a meaning (or the number of syntactic arguments) is *not* the sense of relationality I am referring to here. Valency manifests iconically in syntax, but relationality as I define it is more closely related to notions like boundness. For example, RED is less relational than PLURAL, despite the fact that both are monovalent, because RED is more conceptualizable on its own.

A second problem for this view is that meaning itself is relational. This view goes under the CONCEPTUAL ROLE THEORY of meaning (Block, 1998) in philosophy of mind, and also undergirds structuralist and distributionalist views of meaning in linguistics (**??**). Piantadosi and Hill (2022) provide a useful example, pointing out that despite a clear prototypical concrete referent, the concept of POSTAGE STAMP is fundamentally relational, and we can easily imagine *virtual* postage stamps so long as the abstract referent fulfills the role of tracking payment for delivery. Plausibly, this tension could be resolved by countering that the conceptualization of POSTAGE STAMP is not as relational as the inferential or functional extensions it affords—that is, we imagine a concrete, bounded, non-relational object, which has been selected by the complex networks of meaning

in our minds to fulfill a relational role. Nevertheless, I think this is a serious challenge to the entity–relation distinction which should be further investigated, but stands outside the scope of this thesis.

Lastly, a key objection to both relationality and abstractness is that they are hard to specify and often subjectively defined. On this point I agree. For example, the sense that RED is somehow less relational than ROUND, or PLURAL is less relational than GIVING, despite the higher valency of the latter, is intuitive, but difficult to give precise criteria for. This is why a key goal of this thesis is to provide empirical and computational tools for investigating these notions rigorously. While the above discussion is theoretical and subjective, and one can dispute the abstractness I assign to various concepts, the empirical facts in the rest of the thesis remain. The argument here should be seen as a motivation and interpretive lens for the empirical work that follows.

2.4.3 Summary

In this section, I have attempted to sketch the separate formal and functional dimensions of the lexicality spectrum. On the formal side, I have argued that boundness, allomorphy, and length are all correlated dimensions of formal size. On the functional side, I have argued for the importance of relationality alongside information content, and discussed how the former relates to boundness iconically, and the later relates to formal size via economy principles. Together, this provides a high-level overview of the lexicality spectrum, from nouns at one extreme to inflectional affixes at the other. In the rest of this thesis, I will explore specific aspects of this spectrum and distinctions drawn along it in more detail. Therefore properties specific to inflection and derivation, for example, will be discussed in the relevant chapters.

2.5 Chapter Summary

This chapter has provided a theoretical background for the thesis In Section 2.1, I reviewed the problem of comparative concepts in linguistic typology, discussing approaches like semantic maps and retro-definitions, comparing and contrasting them with my approach of *grounding* problematic distinctions in empirical measures. My approach allows for understanding how consistent existing or proposed operationalizations of comparative concepts are in terms of empirical dimensions. In **??**, I provided an overview of a decade of advancements in deep learning models of language, focusing on how these models capture fine-grained and potentially universal semantic distinctions, and how large-scale pretraining is useful for the acquisition of rich linguistic structure. I argued that these models provide a promising avenue for separating meaning from frequency in typology. In **??**, I connected the computational literature in typology to the notion of comparative concepts, which have largely been implicit in this literature. From phonological typology and semantic category systems, I argued that computational models have been able to advance typological theory and understanding as technologies develop that allow the empirical grounding of the underlying space, further motivating my approach.

Finally, in Section 2.4, I provided a high-level theoretical overview of the lexicality spectrum, separating the formal and functional dimensions, and arguing for the role of informativity and relationality in shaping the formal dimension. In the rest of this thesis, I will use deep learning models to operationalize and investigate specific lexicality-related phenomena, showing consistent patterns across distinctions that have been difficult to formalize in the past.