Construction Grammar Provides Unique Insight into Neural Language Models

Leonie Weissweiler[*◇], **Taiqi He**[†], **Naoki Otani**[†],
David R. Mortensen[†], **Lori Levin**[†], **Hinrich Schütze**[*◇]
[*]Center for Information and Language Processing, LMU Munich
[◇]Munich Center of Machine Learning
[†]Language Technologies Institute, Carnegie Mellon University
weissweiler@cis.lmu.de
{taiqih,notani,dmortens,lsl}@cs.cmu.edu

Abstract

Construction Grammar (CxG) has recently been used as the basis for probing studies that have investigated the performance of large pretrained language models (PLMs) with respect to the structure and meaning of constructions. In this position paper, we make suggestions for the continuation and augmentation of this line of research. We look at probing methodology that was not designed with CxG in mind, as well as probing methodology that was designed for specific constructions. We analyse selected previous work in detail, and provide our view of the most important challenges and research questions that this promising new field faces.

1 Introduction

In this paper, we will analyse existing literature investigating how well constructions and constructional information are represented in pretrained language models (PLMs). We provide context to support the argument that this is one of the most important challenges facing Language Models (LMs) today, and provide a summary of the current open research questions and how they might be tackled.

Our paper is organised as follows: In Section 2, we explain why LMs must understand constructions to be good models of language and perform effectively on downstream tasks. In Section 3, we analyse the existing literature on non-CxG-focused probing to determine its limitations in analysing constructional knowledge. In Section 4, we summarise the existing probing work that is specific to CxG and analyse its data, methodology, and findings. In Section 5, we argue that the development of an appropriate probing methodology for constructions remains an open and important research question (§5.1), and highlight the need for data collection and annotation for facilitating this area of research (§5.2). Finally, in Section 5.4, we suggest next steps that LMs might take if CxG probing reveals fundamental problems.

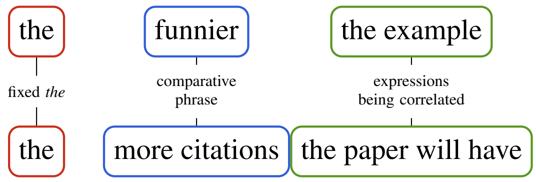

Figure 1: An example illustrating the complexity of a construction. It is an instance of the English Comparative Correlative (CC), with its syntactic features highlighted above the text and paraphrases illustrating its meaning below.

1.1 Construction Grammar

Although there are many varieties of CxG, they share the assumption that the basic building block of language structure is a pair of form and meaning. The form can be anything from a simple morpheme to the types of feature structures seen in Sign-Based Construction Grammar (SBCG) (Boas and Sag, 2012), which can be constellations of inflectional features, morphemes, categories like parts of speech, and syntactic mechanisms. Constructions with many detailed parts in SBCG include comparative constructions in sentences such as *The desk is ten inches taller than the shelf* (Hasegawa et al., 2010) and the causal excess construction as in *It was so big that it fell over* (Kay and Sag, 2012). Most importantly, the form or syntax of a sentence is not reduced to an idealized binary-branching tree or a set of hierarchically arranged pairs of head and dependants. For the purposes of this paper, we take the meaning of a construction to be a combination of Frame Semantics (Petruck and de Melo, 2014) and comparative concepts in semantics and information packaging from language typology (Croft, 2022). Because CxG does not have a clear line separating the lexicon and the grammar, the same kinds of meanings that can be associated with words can be associated with more complex structures. Table 1.1, adapted from Goldberg (2013) illustrates constructions at different

Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023), pages 85 - 95

Construction Name	Construction Template	Examples
Word		Banana
Word (partially filled)	pre-N, V-ing	Pretransition, Working
Idiom (filled)		Give the devil his due
Idiom (partially filled)	Jog <someone's> memory	She jogged his memory
Idiom (minimally filled)	The X-er the Y-er	The more I think about it, the less I know
Ditransitive construction (unfilled)	Subj V Obj1 Obj2	He baked her a muffin
Passive (unfilled)	Subj aux VPpp (PP by)	The armadillo was hit by a car

Table 1: Standard examples of constructions at various levels, adapted from Goldberg (2013)

levels of complexity that contain different numbers of fixed lexemes and open slots.

In this paper, we ask whether PLMs model constructions as gestalts in both form and meaning. For example, we want to know whether a PLM represents a construction like the Comparative Correlative (*The more papers we write, the more fun we have*) as more than the sum of its individual phrases and dependencies. We also want to know whether the PLM encodes knowledge of the open slots in the construction and what can fill them. In terms of meaning, we want to find out whether the sentence's position in embedding space indicates that it has something to do with the correlation between the increase in writing more papers and having more fun. We would like to know whether PLMs represent the meaning of a correlative sentence as close to the meaning of other constructions in English and other languages that have different forms but similar meanings (e.g., *When we write more papers, we have more fun*).

1.2 Language Modelling

This paper is partially concerned with the fundamental questions of language modelling: what is its objective, and what is required of a full language model? We see the objective of language modelling very pragmatically: we aim to build a system that can predict the words in a sentence as well as possible, and therefore our aim in this paper is to point out where this requires knowledge of constructions. We do not take the objective of language modelling to mean that LMs should necessarily achieve their goal the same way that humans do. Therefore, we do not argue that language models need to "think" in terms of constructions because humans do. Rather, we consider constructions an inherent property of human language, which makes it necessary for language models to understand them.

2 Motivation

There has recently been growing interest in developing probing approaches for PLMs based on CxG. We see these approaches as coming from two different motivational standpoints, summarised below.

2.1 Constructions are Essential for Language Modelling

According to CxG, meaning is encoded in abstract constellations of linguistic units of different sizes. This means that LMs, which the field of NLP is trying to develop to achieve human language competency, must also be able to assign meaning to these units to be full LMs. Their ability to assign meaning to words, or more specifically to subword units which are sometimes closer to morphemes than to words, has been shown at length (Wiedemann et al., 2019; Reif et al., 2019; Schwartz et al., 2022). The question therefore remains: are PLMs able to retrieve and use meanings associated with patterns involving multiple tokens? We do not take this to only mean contiguous, fixed expressions, but much more importantly, non-contiguous patterns with slots that have varying constraints placed on them. To imitate and match human language behaviour, models of human language need to learn how to recognise these patterns, retrieve their meaning, apply this meaning to the context, and use them when producing language. Simply put, there is no way around learning constructions if LMs are to advance. In addition, we believe that it is an independently interesting question whether existing PLMs pick up on these abstract patterns using the current architectures and training setups, and if not, which change in architecture would be necessary to facilitate this.

2.2 Importance in Downstream Tasks

Regardless of more fundamental questions about the long-term goals of LMs, we also firmly believe that probing for CxG is relevant for analysing

Lang	Reference Translation	DeepL Translation
German	Sie nieste den Schaum von ihrem Cappuccino runter.	Sie nieste den Schaum von ihrem Cappuccino.
Italian	Lei ha starnutito via la schiuma dal suo cappuccino.	Starnutì la schiuma del suo cappuccino.
Turkish	Cappuccino'sunun köpüğünü hapşırdı.	Hapşırarak cappuccino'sunun köpüğünü uçurdu.

Table 2: Translations of 'She sneezed the foam off her cappuccino.' given by DeepL[1]. Translated back to English by humans, they all mean "She sneezed her cappuccino's foam.", which does not correctly convey the resultative meaning component, i.e., that the foam is removed from the cappuccino by the sneeze (as opposed to put there).

the challenges that face applied NLP, as evaluated on downstream tasks, at this point in time. Discussion is increasingly focusing on diagnosing the specific scenarios that are challenging for current models. Srivastava et al. (2022) propose test suites that are designed to challenge LMs, and many of them are designed by looking for 'patterns' with a non-obvious, non-literal meaning that is more than the sum of the involved words. One example of such a failure can be found in Table 2, where we provide the DeepL[1] translations for the famous instance of the caused-motion construction (Goldberg, 1995, CMC;): 'She sneezed the foam off her cappuccino', where the unusual factor is that *sneeze* does not usually take a patient argument or cause a motion. For translation, this means that it either has to use the corresponding CMC in the target language, which might be quite different in form from the English CMC, or paraphrase in a way that conveys all meaning facets. For the languages we tested, DeepL did not achieve this: the resulting sentence sounds more like the foam was sneezed onto the cappuccino, or is ambiguous between this and the correct translation. Interestingly, for Russian, the motion is conveyed in the translation, but not the fact that it is caused by a sneeze.

Targeted adversarial test suites like this translation example can be a useful resource to evaluate how well LMs perform on constructions, but more crucially, CxG theory and probing methods will inform the design of better and more systematic test suites, which in turn will be used to improve LMs (§5.4).

2.3 Diversity in Linguistics for NLP

Discussions about PLMs as models of human language processing have recently gained popularity. One forum for such discussions is the Neural Nets for Cognition Discussion Group at CogSci2022[2]. The work is still very tentative, and most people agree that LMs are not ready to be used as models of human language processing. However, the discussion about whether LMs are ready to be used as cognitive models is dominated by results of probing studies based on Generative Grammar (GG), or more specifically Transformational Grammar. This means that GG is being used as the gold standard against which the cognitive plausibility of LMs is evaluated. Studies using GG assume a direct relationship between the models' performance on probing tasks and their linguistic competency. Increased performance on GG probing tasks is seen as a sign it is becoming more reasonable to use LMs as cognitive models. Another linguistic reason for theoretical diversity is that if we could show that LMs conform better to CxG rather than GG, this might open up interesting discussions if they ever start being used as cognitive models.

3 Established Probing Methods Are Only Applicable to Some Aspects of CxG

Established probing methods have focused on different aspects of the syntactic and semantic knowledge of PLMs. In this section, we summarise the major approaches that were not designed specifically with constructions in mind. We show that although each of these methodologies deals with some aspect of CxG, and might even fully investigate some simpler constructions, none of them fully covers constructional knowledge as defined in Section 1.1.

3.1 Probing Using Contextual Embeddings

Various probing studies (Garcia et al., 2021; Chronis and Erk, 2020; Karidi et al., 2021; Yaghoobzadeh et al., 2019; *inter alia*) have focused on analysing contextual embeddings at different layers of PLMs, either of one word or multiple words, or both. The common thread in their methodology is that they compare the embeddings of the same word in different contexts, or of different words in the same context. From a constructional point of view, this requires finding two

[1] https://www.deepl.com/translator
[2] http://neural-nets-for-cognition.net

constructions with similar surface forms. By comparing the embeddings over many sentences, they are able to investigate if a certain word "knows" in which construction it is, which provides evidence for the constructional knowledge of a model.

While this is a useful starting point for probing, it is also limited. Sentences with similar constructions have to be identified, which is not always possible. More importantly, this methodology currently does not tell us anything about if the model has identified the extent of the construction correctly, or if the model has correctly learned how each slot can be filled.

3.2 Probing for Relationships Between Words

Some probing studies investigate whether a PLM recognises a word pair associated with a meaningful relationship of some kind (Rogers et al. (2020)). Most prominently, probing based on Universal Dependencies (UD; de Marneffe et al. (2021)) by Hewitt and Manning (2019) attempts to find out whether there is a high attention weight between words that are in a dependency relation where one word is the head and the other word is the dependent. They found different attention heads at different layers that seem to represent specific dependency relations such as a direct object attending to its verb, a preposition attending to its object, determiners attending to nouns, possessive pronouns attending to head nouns, and passive auxiliary verbs attending to head verbs.

The methodology as it was used by Hewitt and Manning (2019) looked at the one token that each token attended to the most. This made sense for the Hewitt and Manning (2019) study because they were probing for UD structures, which consist of binary relationships of heads and dependents in a hierarchical structure.

However, the methodology would have to be extended if we want to find out whether a whole construction with many construction elements is represented in the model in something other than a hierarchical set of binary relations. Most varieties of CxG recognise constructions with more than two daughters and constructions such as *thirty miles an hour* (Fillmore et al., 2012) in which no element is the head (headless constructions). As a research question, it is still unclear what patterns of attention we would consider as evidence that a model encodes a construction that may have headless and non-binary branches. An appropriate probing methodology has not yet been developed.

3.3 Probing with Minimal Pairs

Some works in probing based on Generative Grammar have relied on finding minimal pairs of sentences that are identical except for one specific feature that, if changed, will make the sentence ungrammatical (Wei et al., 2021). For example, in *The teacher who met the students is/*are smart*, a language model that encodes hierarchical structure would predict *is* rather than *are* after *students*, whereas a language model that was fooled by adjacency might predict *are* because it is next to *students*. The sentences can be safely compared, because only one feature, in this case, the verb being assigned the same number as the subject, is changed, and no other information can intervene or distort the probe. Other studies use a more complicated paradigm of minimal pairs involving filler-gap constructions, contrasting *I know what the lion attacked (gap) in the desert* and *I know that the lion attacked the gazelle (no gap) in the desert*.

These probing methodologies have led to productive lines of research and have been applied to complex constructions such as the Comparative Correlative Construction (Weissweiler et al., 2022). However, they depend on finding two minimally different constructions, which differ only in one way (e.g., singular/plural or gap/no gap), but close minimal pairs are simply not available for every construction.

4 CxG-specific Probing

We have argued that the most commonly used and straightforward probing methods are not sufficient for fully investigating constructional knowledge in PLMs. However, there have been several papers which have created new probing methodologies specifically for constructions. In this section, we will analyse them in terms of

- Which constructions were investigated? Does the paper investigate specific constructions or does it use a pre-compiled list of constructions or restrain itself to a subset?
- For the specific instances of their construction or constructions, what data are they using? Is it synthetic or collected from a corpus? If from a corpus, how was it collected?
- What are the key probing ideas?

Paper	Language	Source	Construction	Example
Tayyar Madabushi et al. (2020)	English	From automatically constructed list by Dunn (2017)	Personal Pronoun + didn't + V + how	We didn't know how or why.
Li et al. (2022)	English	Argument Structure Constructions according to Bencini and Goldberg (2000)	caused-motion	Bob cut the bread into the pan.
Tseng et al. (2022)	Chinese	From constructions list by (Zhan, 2017)	a + 到 + 爆, etc.	好吃到爆了！ *It's so delicious!*
Weissweiler et al. (2022)	English	McCawley (1988)	Comparative Correlative	The bigger, the better.

Table 3: Overview of constructions investigated in CxG-specific probing literature, with examples.

- Does the paper only investigate probing of (unchanged) pretrained models or is finetuning also considered?

For ease of reference, we provide an overview of the constructions investigated by each of the papers in Table 3.

4.1 CxGBERT

Tayyar Madabushi et al. (2020) investigate how well BERT (Devlin et al., 2019) can classify whether two sentences contain instances of the same construction. Their list of constructions is extracted with a modified version of Dunn (2017)'s algorithm: they induce a CxG in an unsupervised fashion over a corpus, using statistical association measures. Their list of constructions is taken directly from Dunn (2017), and they find their instances by searching for those constructions' occurrences in WikiText data. This makes the constructions possibly problematic, since they have not been verified by a linguist, which could make the conclusions drawn later from the results about BERT's handling of constructions hard to generalise from.

The key probing question of this paper is: Do two sentences contain the same construction? This does not necessarily need to be the most salient or overarching construction of the sentence, so many sentences will contain more than one instance of a construction. Crucially, the paper does not follow a direct probing approach, but rather finetunes or even trains BERT on targeted construction data, to then measure the impact on CoLA. They find that on average, models trained on sentences that were sorted into documents based on their constructions do not reliably perform better than those trained on original, unsorted data. However, they additionally test BERT Base with no additional pre-training on the task of predicting whether two sentences contain instances of the same construction, measuring accuracies of about 85% after 500 training examples for the probe. These results vary wildly depending on the frequency of the construction, which might relate back to the questionable quality of the automatically identified list of constructions.

4.2 Neural Reality of Argument Structure Constructions

Li et al. (2022) probe for LMs' handling of four argument structure constructions: ditransitive, resultative, caused-motion, and removal. Specifically, they attempt to adapt the findings of Bencini and Goldberg (2000), who used a sentence sorting task to determine whether human participants perceive the argument structure or the verb as the main factor in the overall sentence meaning. The paper aims to recreate this experiment for MiniBERTa (Warstadt et al., 2020) and RoBERTa (Liu et al., 2019), by generating sentences artificially and using agglomerative clustering on the sentence embeddings. They find that, similarly to the human data, which is sorted by the English proficiency of the participants, PLMs increasingly prefer sorting by construction as their training data size increases. Crucially, the sentences constructed for testing had no lexical overlap, such that this sorting preference must be due to an underlying recognition of a shared pattern between sentences with the same argument structure. They then conduct a second experiment, in which they insert random verbs, which are incompatible with one of the constructions, and then measure the Euclidean distance between this verb's contextual embedding and that of a verb that is prototypical for the corresponding construction. The probing idea here is that if construction information is picked up by the model, the contextual embedding of the verb should acquire some constructional meaning, which would bring it closer to the corresponding prototypical verb meaning than to the others. They indeed find that this effect is significant, for both high and low frequency verbs.

4.3 CxLM

Tseng et al. (2022) study LM predictions for the slots of various degrees of openness for a corpus of Chinese constructions. Their original data comes from a knowledge database of Mandarin Chinese constructions (Zhan, 2017), which they filter so that only constructions with a fixed repetitive element remain, which are easier to find automatically in a corpus. They filter this list down further to constructions which are rated as commonly occurring by annotators, and retrieve instances from a POS-tagged Taiwanese bulletin board corpus. They binarise the openness of a given slot in a construction and mark each word in a construction as either constant or variable. The key probing idea is then to examine the conditional probabilities that a model outputs for each type of slot, with the expectation that the prediction of variable slot words will be more difficult than that of constant ones, providing that the model has acquired some constructional knowledge. They find that this effect is significant for two different Chinese BERT-based models, as negative log-likelihoods are indeed significantly higher when predicting variable slots compared to constant ones. Interestingly, the negative log-likelihood resulting from masking the entire construction lies in the middle of the two extremes. They further evaluate a BERT-based model which is finetuned on just predicting the variable slots of the dataset they compiled and find, unsurprisingly, that this improves accuracy greatly.

4.4 Probing for the English Comparative Correlative

Weissweiler et al. (2022) investigate large PLM performance on the English Comparative Correlative (CC). There are two key probing ideas, corresponding to the investigation of the syntactic vs. the semantic component of CC. They probe for PLM understanding of CC's syntax by attempting to create minimal pairs, which consist of sentences with instances of the CC and very similar sentences which do not contain an instance of the CC. They collect minimal pairs from data by searching for sentences that fit the general pattern and manually annotate them as positive and negative instances, and additionally construct artificial minimal pairs that turn a CC sentence into a non-CC sentence by reordering words. They find that a probing classifier can distinguish between the two classes easily, using mean-pooled contextual PLM embeddings. They also probe the models' understanding of the meaning of CC, for which they choose a usage-based approach, constructing NLU-style test sentences in which an instance of the construction is given and has then to be applied in a context. They find no above-chance performance for any of the models investigated in this task.

4.5 Summary

In this section, we summarise the findings of previous work on CxG-based LM probing and analyse them in terms of the constructions that are investigated, the data that is used and the probing approaches that are applied.

4.5.1 Constructions Used

So far, Tseng et al.'s (2022) study is only the work that chose a set of constructions from a list precompiled by linguists. They constrain their selection to contain only constructions that are easy to search for in a corpus, and the resource they use only contains constructions with irregular syntax, but it is nevertheless to be considered a positive point that they are able to reach a diversity of constructions investigated. In contrast, both Li et al. (2022) and Weissweiler et al. (2022) pick one or a few constructions manually, both of which are instances of 'typical' constructions frequently discussed in the linguistic literature. This makes the work more interesting to linguists and the validity of the constructions is beyond doubt. But the downside is selection bias: the constructions that are frequently discussed are likely to have strong associated meanings and do not constitute a representative sample of constructions, from a constructions-all-the-way-down standpoint (Goldberg, 2006). Lastly, Tayyar Madabushi et al. (2020) rely on artificial data collected by Dunn (2017). We consider this method to be unreliable, but it has the resulting dataset has the advantage of variety and large scale.

4.5.2 Data Used

The two main approaches to collecting data are: (i) *patterns*: finding instances of the constructions using patterns of words / part-of-speech (POS) tags and (ii) *generation* of synthetic data. Tseng et al. (2022), Weissweiler et al. (2022) and Tayyar Madabushi et al. (2020) use patterns while Li et al. (2022) and a part of Weissweiler et al. (2022) generate data based on formal grammars. Patterns have the advantage of natural data and are less prone to accidental unwanted correlations. But there is a risk of errors in the data collection process, even after the set of constructions has to be constrained to even allow for automatic classification, and the data may have been post-corrected by manual annotation, which is time-intensive. On the other hand, generation bears challenges for making the sentences as natural as possible, which can eliminate confounding factors like lexical overlap.

4.5.3 Probing Approaches Used

Regarding the probing approaches, all previous work has had its own idea. Weissweiler et al. (2022) and Li et al. (2022) both operate on the level of sentence embeddings, classifying and clustering them respectively. Tayyar Madabushi et al. (2020) could maybe be classified with them, as it employs the Next Sentence Prediction objective (Devlin et al., 2019), which operates at the sentence level. On the other hand, another part of Weissweiler et al. (2022), as well as Tseng et al. (2022), works at the level of individual predictions for masked tokens.

The greatest difference between these works is in their concept of evidence for constructional information learned by a model, and what this information even consists of. Tayyar Madabushi et al. (2020) frame this information as 'do these two sentences contain the same construction', Li et al. (2022) as 'is clustering by the construction preferred over clustering by the verb', Weissweiler et al. (2022) as 'can a small classifier distinguish this construction from similar-looking sentences' and 'can information given in form of a construction be applied in context', and Tseng et al. (2022) as 'are open slots more difficult to predict than closed ones'. There is little overlap to be found between these approaches, so it is difficult to draw any conclusion from more than one paper at a time.

4.5.4 Overall Findings

We nonetheless make an attempt at summarising the findings so far about large PLMs' handling of constructional information. Regarding the structure, all findings seem to be consistent with the idea that models have picked up on the syntactic structure of constructions and recognised similarities between different instances of the same construction. This appears to hold true even when tested in different rigorous setups that exclude bias from overlapping vocabulary or accidentally similar sentence structure. This has mostly been found for English, as Tseng et al. (2022) are the only ones investigating it for a non-English language, and it remains to be seen if it holds true for lower-resources languages. Considering the acquisition of the meaning of constructions, only Weissweiler et al. (2022) have investigated this, and found no evidence that models have formed any understanding of it, but were not able to provide conclusive evidence to the contrary.

5 Research Questions

In this section, we lay out our view of the problems that are facing the emerging field of CxG-based probing and the reasons behind these challenges, and propose avenues for potential future work and improvement.

5.1 How Can We Develop Probing Methods that are a Better Fit for CxG?

Going forward, we see two directions. One is what has already been happening: keep finding new ways to get around the inherent difficulty of probing for constructions, which leads us to mostly non-conclusive and not entirely reliable evidence. The better, and more difficult way forward, is to adopt a fundamentally different methodology that would establish a standard of evidence/generalisability comparable to GG-based probing.

5.2 Data

Another reason why so little work has been done in this important field is likely the lack of data. We view the lack of data as divided into three parts: the lack of lists of constructions, the lack of meaning descriptions or even a unified meaning formalism for them, and the lack of annotated instances in corpora. We explain different opportunities for the community to obtain this data going forward below.

5.2.1 Exploiting Non-constructicon Data

Many resources are available, as already stated above, that have collected or created data with specific constructions, with the aim of making certain tasks more challenging to the models in a specific way. We can analyse those datasets and the results on them from a CxG point of view, and this can add to our pool of knowledge about what models struggle with regarding constructions. They will probably not contain any meaning descriptions, but some, like in Srivastava et al. (2022), are grouped naturally by construction, and contain instances in data, which may however be artificial.

5.2.2 Making Constructicons Available

Recently, there has been substantial work by linguists to develop constructicons for different languages (Lyngfelt et al., 2018; Ziem et al., forthcoming). Some of these constructicons are readily available online, e.g., the Brazilian Portuguese one, but many are either not available or have an interface that makes them difficult to access, e.g., because it is in the constructicon's language. Although to our knowledge, none of these constructicons contain annotated instances in text, and their meaning representations will be very difficult to unify, they are an important resource at least for lists of constructions that can be investigated by probing methods. They are especially valuable because of their linguistic diversity (English, German, Japanese, Swedish, Russian, Brazilian Portuguese), the lack of which is a major flaw in the current literature, as we stated above in §4.5.4.

5.2.3 Universal Constructicon

As a more ambitious project than simply making these constructicons available online, we firmly believe that the field would benefit greatly from an attempt to unify their representations and make them available as a shared resource. Parallels can be drawn here to UD (de Marneffe et al., 2021), a project which developed a simplified version of dependency syntax that could be universally applied and agreed upon, and then provided funding for the creation of initial resources for a range of languages, which was later greatly added to by community work in the different communities. This was a major factor in the popularisation of dependency syntax within the NLP community, to the point where it is now almost synonymous with syntax itself, due in no small part to its convenience for computational research.

As a second step after the creation of a shared online resource to access the existing constructicons, the community could consider developing a shared representation to formalise the surface form of the constructions. A dataset without meaning representation that includes multiple languages would already be a very useful resource. As a next step after that, we could think about aligning constructions across languages that encode a similar meaning. The last and most ambitious step would be unifying and linking the meaning representations, which would ideally be formalised similarly to AMR (Banarescu et al., 2013). This would enable us to develop automatic test suites that can really account for the constructions' meanings and not just their structure.

5.2.4 Annotated Instances in Text

In any stage of the development of 'construction lists' detailed above, it would be necessary to find instances of the constructions in text. Some of the probing literature described above have generated this data artificially, which is time-consuming and also removes two important advantages of precompiled construction lists: objectivity and scale. Therefore, the ideal solution would be to find resources to have data annotated for constructions. This in itself faces many challenges from a constructions-all-the-way-down perspective: annotating even one sentence completely would be very time-consuming and require many discussions about annotation schemata in advance. A more basic way of acquiring data would be to focus on a limited set of constructions, which is selected manually, and to use pre-filtering methods similar to those employed by Tseng et al. (2022) and Weissweiler et al. (2022), to acquire simply an Inside-Outside-Beginning marking in sentences that might be instances of a construction. On the downside, this is far less linguistically rigorous and also less timeless than Universal Dependencies, which guarantees that any annotated sentence has been fully annotated and will probably not need to be revised. Nevertheless, a compromise will need to be found if annotated data is to be created at all.

5.3 CxG and Transformer Architecture

As more work is done on CxG-based probing, the field will hopefully soon be able to approach the questions that we see as crucial. Current probing techniques have not yet shown that PLMs are able to adequately handle the meaning of constructions. Assuming that more comprehensive probing techniques will show conclusively that this is not the case, is it due to a lack of data? Or is there a fundamental incompatibility of current architectures and the concept of associating a pattern with a meaning? In 5.3.1 and 5.3.2, we elaborate on why the latter might be the case.

5.3.1 Non-compositional Meaning

It is possible that constructions are intrinsically difficult for LMs because they include non-compositional meaning that is not attached to a token. It is tempting to compare them to simpler multiword expressions, which also have meaning that spans several words and that is only instantiated when they appear together. They also pose a challenge to LMs because of this, as their concept of sentence meaning is often too compositional (Liu and Neubig, 2022). The key difference is in our view, that for very complex constructions, it is not clear where in the model we can search or probe for the additional meaning.

The meaning is not attached to the words instantiating the construction, but rather to the abstract pattern itself (Croft, 2001), which we can recognise, connect mentally to previous instances and store meaning for. Once we have retrieved this meaning, it is potentially applied to the whole sentence, and can therefore have consequences for the contextual meaning of words which were never even involved in it. In a transformer-based LM, this additional meaning component cannot be stored in the static embeddings and contextualised through the attention layers, because unlike for MWEs, many constructions have very open slots, so that it is impossible to say that their meaning should somehow be stored with the meaning of the words that may instantiate them. The only place to store constructional information, therefore, remains the model weights, which are much harder to investigate or alter than the model's input, and further probing might reveal that they are unable to store it at all.

5.3.2 The Language Modelling Objective

Another possibility for fundamental difficulties arises from the nature of the training objective. PLMs are typically trained either on a masked or causal language modelling objective (Devlin et al., 2019; Radford et al., 2019). It makes sense that this incentivises them to learn word meaning in context, which they will need to predict certain words, and also relationships between words, such as simple morphological dependencies. However, information about the meaning of a construction might not often be learned in a language modelling setting, simply because it will not be needed to make the correct prediction. The meaning of a construction might not be necessary information to predict one of its component words correctly when it is masked, although its structure certainly will. In contrast, finetuning on a downstream task that requires assessment of sentence meaning, such as sentence classification, might enable us to better access the constructional meaning contained in PLMs, because the finetuning objective has required explicit use of this meaning. On the other hand, this might also be thought of as a distortion of the lens, as grammatical knowledge is not typically evaluated on finetuned models, because the findings might not generalise well.

5.4 Adapting Pretraining for CxG

If we do decide that there is a fundamental problem with the current architecture and/or training regime, the next logical step would be to think about how to alter these so that acquisition of constructional meaning becomes possible. Something similar has already been considered by Tseng et al. (2022), where models are finetuned on data that has been altered to mask entire construction instances at once, and by Tayyar Madabushi et al. (2020), which collects sentences that contain instances of the same construction into 'documents' and pretrains on them. This line of thinking, which can be summarised as data modification with constructional biases, can be further expanded, to give models some help with associating sentences with similar constructions with each other.

A far more radical idea would be to think about injecting something into the architecture that could represent this additional meaning, in the style of a position embedding, or a control token (Martin et al., 2020).

6 Conclusion

We have motivated why probing large PLMs for CxG is a very important topic both for computational linguists interested in the ideal LM and for applied NLP scientists seeking to analyse and improve the current challenges that models are facing. We then summarised and analysed the existing literature on this topic. Finally, we have given our reasons for why CxG probing remains a challenge, and detailed suggestions for further development in this field, within the realms of data, methodology, and fundamental research questions.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan

Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Giulia ML Bencini and Adele E Goldberg. 2000. The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language*, 43(4):640–651.

H. C. Boas and I. A. Sag. 2012. *Sign-Based Construction Grammar*. Center for the Study of Language and Information.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? when it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 227–244, Online. Association for Computational Linguistics.

William Croft. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press on Demand.

William Croft. 2022. *Morphosyntax: Constructions of the World's Languages*. Cambridge University Press.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Dunn. 2017. Computational learning of construction grammars. *Language and cognition*, 9(2):254–292.

C. J. Fillmore, R. Lee-Goldman, and R. Rhodes. 2012. The framenet constructicon. In H. C. Boas and I. A. Sag, editors, *Sign-Based Construction Grammar*. Center for the Study of Language and Information.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Adele Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press, Oxford, UK.

Adele E.. Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Adele E. Goldberg. 2013. 1415 Constructionist Approaches. In *The Oxford Handbook of Construction Grammar*. Oxford University Press.

Yoko Hasegawa, Russell Lee-Goldman, Kyoko Hirose Ohara, Seiko Fujii, and Charles J Fillmore. 2010. On expressing measurement and comparison in english and japanese. *Contrastive studies in construction grammar*, 10.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Taelin Karidi, Yichu Zhou, Nathan Schneider, Omri Abend, and Vivek Srikumar. 2021. Putting words in BERT's mouth: Navigating contextualized vector spaces with pseudowords. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10300–10313, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul Kay and Ivan A. Sag. 2012. Cleaning up the big mess: Discontinuous dependencies and complex determiners. In H. C. Boas and I. A. Sag, editors, *Sign-Based Construction Grammar*. Center for the Study of Language and Information.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Emmy Liu and Graham Neubig. 2022. Are representations built from the ground up? an empirical examination of local composition in language models. *arXiv preprint arXiv:2210.03575*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, and Tiago Timponi Torrent. 2018. *Constructicography: Constructicon development across languages*, volume 22. John Benjamins Publishing Company.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language*

Resources and Evaluation Conference, pages 4689–4698, Marseille, France. European Language Resources Association.

James D McCawley. 1988. The comparative conditional construction in english, german, and chinese. In *Annual Meeting of the Berkeley Linguistics Society*, volume 14, pages 176–187.

Miriam R. L. Petruck and Gerard de Melo, editors. 2014. *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*. Association for Computational Linguistics, Baltimore, MD, USA.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Lane Schwartz, Coleman Haley, and Francis Tyers. 2022. How to encode arbitrarily complex morphology in word embeddings, no corpus needed. In *Proceedings of the first workshop on NLP applications to field linguistics*, pages 64–76, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku, and Shu-Kai Hsieh. 2022. CxLM: A construction and context-aware language model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6361–6369, Marseille, France. European Language Resources Association.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.

Yadollah Yaghoobzadeh, Katharina Kann, T. J. Hazen, Eneko Agirre, and Hinrich Schütze. 2019. Probing for semantic classes: Diagnosing the meaning content of word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5740–5753, Florence, Italy. Association for Computational Linguistics.

Weidong Zhan. 2017. On theoretical issues in building a knowledge database of chinese constructions. *Journal of Chinese Information Processing*, 31(1):230–238.

Alexander Ziem, Alexander Willich, and Sascha Michel. forthcoming. *Constructing constructicons*. John Benjamins Publishing Company.