The dynamic generalisation evaluation research taxonomy

Author 1, author 2, author 3, ...

Abstract

The ability to generalise well is one of the primary desiderata of natural language processing (NLP). Yet how 'good generalisation' should be defined and what that entails in practice is not well understood. As a consequence, newly proposed models are not usually systematically tested for their ability to generalise. In this paper, we present a comprehensive taxonomy that can be used to characterise generalisation research in NLP along five different axes: their main *motivation*, the *type* of generalisation they aim to attack, the type of *data shift* they are considering, the *locus* of this shift, and the *data shift* they are considering and the *source* by which this data shift is obtained. We explain the axes of our taxonomy by providing ample examples from the literature and then use it to survey N previous papers that present generalisation tests. Then, we use those results to more generally assess where we are when it comes to evaluating generalisation in NLP, identify areas that are over- or underrepresented, and make recommendations for what questions should still be addressed in the future. Along with this paper, we release a webpage where the results of our survey can be dynamically viewed and updated as new NLP generalisation studies come out.

1 Introduction

Good generalisation, roughly described as the ability to successfully transfer representations, knowledge, and strategies from past experience to new experiences, is one of the primary desiderata for models of natural language processing (NLP) (Elangovan et al., 2021; Lake et al., 2017; Linzen, 2020; Plank, 2016; Schmidhuber, 1990; Wong and Wang, 2007; Yogatama et al., 2019, i.a.), as well as in the wider field of machine learning (e.g. Kirk et al., 2021; Shen et al., 2021). There is, however, little agreement about what kind of generalisation behaviour modern-age NLP models should exhibit, and under what kind of conditions that should be evaluated. Broadly speaking, generalisation is evaluated by assessing how well a model performs on a test dataset, given the relationship of this dataset with the data that this model was trained on. For decades, it was common that the only constraint put on this relationship, was that the train and test data were different. Generalisation was evaluated by training and testing models on different, but similarly sampled data—or, more precisely, independent and identically distributed (*i.i.d.*) data. Typically, such training and test data are generated by randomly splitting a corpus into a training and a test partition. In the past 20 years, we have seen great strides on such random train-test splits, in a range of different applications. Since the first release of the Penn Treebank (Marcus et al., 1993), F1 scores went from values in the high 80s at the end of the previous millennium (Collins, 1996; Magerman, 1995) and the first ten years of the current one (e.g. Petrov and Klein, 2007; Sangati and Zuidema, 2011) to scores up to 96 in the most recent past (Mrini et al., 2020; Yang and Deng, 2020). On the same corpus, performance for language modelling went from perplexity scores well above 100 (Kneser and Ney, 1995; Rosenfeld, 1996) to a score of 20.5 in 2020 (Brown et al., 2020). Progress in many areas of NLP has become even faster in the very last years. Scores for the popular evaluation set GLUE went from scores between 60-70 at its release (Wang et al., 2018), to scores exceeding 90 less than a year after (most famously, Devlin et al., 2019), with performances on a wide range of tasks reaching near-perfect accuracies (e.g. Devlin et al., 2019; Liu et al., 2019b; Wang et al., 2019, 2018). More recently, strongly scaled-up models (e.g. Chowdhery et al., 2022) show astounding performances on almost all existing i.i.d. natural language understanding benchmarks.

With this impressive progress, however, also came the realisation that, for a neural network, having a high or human-ceiling scores on an i.i.d test set does not necessarily imply that this model in fact robustly generalises to a wide range of different scenarios. In the recent past, we witnessed a surge of different generalisation studies that point out generalisation failures in neural models (Blodgett et al., 2016; Khishigsuren et al., 2022; Kim and Linzen, 2020; Lake and Baroni, 2018; McCoy et al., 2019; Plank, 2016; Razeghi et al., 2022; Sinha et al., 2021, to give just a few examples). Some show that when models perform well on i.i.d. test splits, they might rely on simple heuristics that do not robustly generalise in a wide range of non-i.i.d. scenarios (McCoy et al., 2019), that models over-rely on stereotypes (Parrish et al., 2022; Srivastava et al., 2022), or bank on memorisation rather than generalisation (Razeghi et al., 2022). Others, instead, discuss cases in which performances drop when the evaluation data differs from the training data in terms of genre, domain or topic (e.g. Malinin et al., 2021; Michel and Neubig, 2018; Plank, 2016), or for different subpopulations (e.g. Blodgett et al., 2016; Dixon et al., 2018). Yet others focus on models' inability to generalise compositionally (Dankers et al., 2022; Kim and Linzen, 2020; Lake and Baroni, 2018; Li et al., 2021b), structurally (Sinha et al., 2021; Weber et al., 2021; Wei et al., 2021), or to longer sequences (Dubois et al., 2020; Raunak et al., 2019). More recently, Srivastava et al. (2022) show that despite their impressive performances on expansive test suites, state-of-the-art models do not generalise well to slightly different task formulations of the same problem.

These are just a few examples in a long list of studies that aim to investigate the generalisation abilities of NLP models, focussing in particular on models and training regimes that score well on traditional train-test splits. Taken together, this body of work brings into question the kind of generalisation capabilities recent breakthroughs actually reflect, and how generalisation should be tested for, if not with i.i.d. splits. At the same time, these works differ amply in the definitions they give of generalisation, the assumptions they make about when and how models should generalise, and even the evaluation settings they use. They encompass a wide range of generalisation-related research questions, and they also use a wide range of different methodologies and experimental setups. Consequently, it is not easy to understand how their results relate to each other, what types of generalisation are being addressed and which are neglected, what types of generalisation we should prioritise in the field of natural language processing, and how we can adequately assess generalisation in the first place.

With this work, we aim to provide structure to the field of generalisation evaluation as well as analyse the work that has been done so far. By carefully surveying existing work on generalisation evaluation, we identify 5 main axes of variation along which those studies differ. We incorporate those five axes in a taxonomy, that can be used to better understand the heterogenous landscape of generalisation testing, with as ultimate goal to help researchers better structure and understand generalisation evaluation research in the future. The different axes in our taxonomy target the following five questions:

- What is the high-level *motivation* for testing generalisation (Section 2)?
- What is the *type* of generalisation the test is addressing? (Section 3)?
- What kind of *data shift* occurs between training and testing? (Section 4)?
- What is the *source* of the data shift considered (Section 5)?
- What is the *locus* of the data shift? (Section 6)?

We describe the meaning of these axes and the possible values that generalisation studies can take on these axes, providing representative examples for each. Next, in Section 7, we use our axis-based taxonomy to survey N studies. We present these results in comprehensive figures, which we use to describe

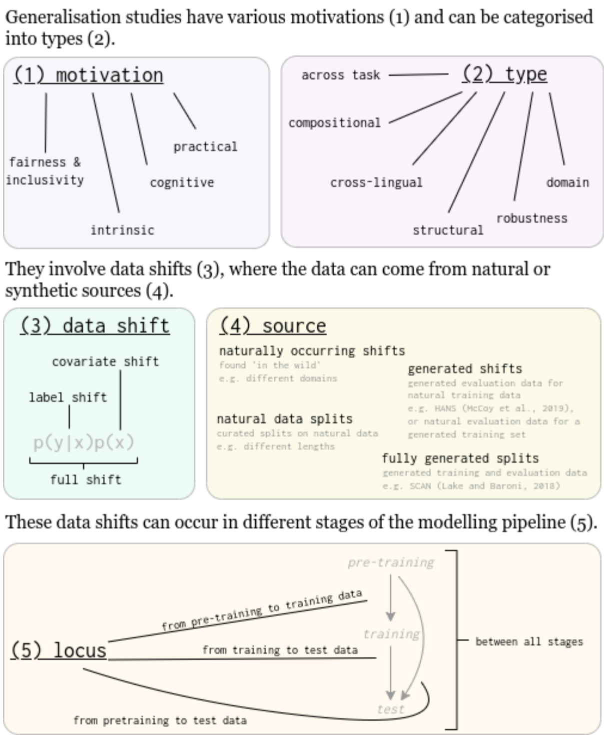

Figure 1: A graphical representation of the NLP generalisation taxonomy we present in this paper. The taxonomy conists of five different (nominal) axes, that describe the high-level *motivation* of the work (§ 2); the *type* of generalisation the test is addressing (§ 3); what kind of *data shift* occurs between training and testing (§ 4), what the *source* is of the data shift considered in the test (§ 5) and what the *locus* of the data shift is (§ 6)

the current landscape of generalisation testing in NLP, and identify areas where more work is needed. DH: Potentially, add a few findings from our survey, which we are still in the process of finishing.

In summary, our contributions are the following:

i) We present an axis-based *generalisation taxonomy* that can be used to characterise generalisation studies in NLP;

ii) We survey generalisation studies in NLP, along the five main axes of variation in this taxonomy;

iii) With these survey results, we discuss the status of generalisation research in NLP;

(iv) and we provide suggestions to steer the field towards more sound and exhaustive generalisation tests.

Along with this paper, we also present a website `insertlink`, where our survey results can be visualised dynamically, and where we encourage readers to add (new) generalisation studies that are not yet included.

2 High-level motivations for evaluating generalisation

The first axis we consider in our taxonomy is the high-level motivation that is given to study generalisation. Broadly speaking, there are four closely intertwined motivations to do so, which we call the *practical*, the *cognitive*, the *intrinsic*, and the *fairness* motivation.[1] We discuss each of them below.

Practical: in what settings can the model be used? One frequently posed motivation to study generalisation is of a highly practical nature. It concerns in what kind of scenarios a trained model can be reasonably used. Questions with a primarily practical motivation often relate to how well models generalise to different domains or differently collected data. For instance, Michel and Neubig (2018) consider how well machine translation models trained on canonical text can generalise to noisy data from an internet platform; Lazaridou et al. (2021) investigate language model generalisation to different time periods; and Talman and Chatzikyriakidis (2019) investigate how well natural language inference (NLI) models generalise from one NLI dataset to another. Other questions that are frequently addressed from a practical perspective concern biases in the data, and pertain to whether models robustly generalise to different demographics and subpopulations. For instance, Koh et al. (2021) present a study using the CivilComments-Wilds dataset, in which models are evaluated in terms of their worst-group accuracy, instead of their average across all demographic groups.

Cognitive: does the model generalise like a human? A second high-level motivation that is driving generalisation research is cognitively oriented, and consists of two separate underlying categories. The first category is related to model behaviour: human generalisation is a useful reference point for the evaluation of model generalisation in NLP, because human generalisation is considered to be very powerful (e.g. Lake et al., 2017). Humans are known to learn quickly, from fewer data than models (Linzen, 2020), and they easily (compositionally) recombine things they already know to understand new concepts they have never seen before. These feats are arguably also important for models, and are, therefore, good targets for generalisation testing. In terms of concrete aims, there is thus a strong overlap between cognitively-inspired and practical motivations: assuming human generalisation is strong, a model that generalises like a human should score well also on practically motivated tests. In our axes-based taxonomy, the difference between *cognitive* and *practical* resides mostly in the types of scenarios that are considered in tests: are they scenarios that are artificially created to get a higher-level, isolated impression of generalisation, or scenarios that might also occur in practice?[2]

The second, truly cognitively inspired category, contains work that evaluates generalisation to learn more about cognition and language (Baroni, 2021; Hupkes, 2020). In studies that are motivated as such, the question of how a particular model generalises is primarily investigated to derive new hypotheses about how human generalisation might work. For instance, Lakretz et al. (2021b) perform a detailed study of how LSTM models generalise to specific kinds of nested syntactic constructions, which they then use to inform a human experiment on the same syntactic constructions.

Intrinsic: does the model capture the task correctly? A third motivation to evaluate generalisation in models of NLP, that cuts through the two previously mentioned motivations, appertains to the question "*did a model actually learn the task we intended it to learn, as we intended it to learn it?*". The assumption underpinning this type of research is that if a model is truly implementing the task it is trained to do, it should be able to execute this task also in settings that differ from the exact setting the model was trained in. The most important dimension in which studies that are motivated by this question differ, is

[1]As we will see in what follows, the same questions can often be asked with different underlying motivations, and sometimes it might be difficult to tease apart what the exact motivation is. A few studies genuinely stem from two or more motivations; we will mark them accordingly in our survey. For most cases, however, while a generalisation test may inform research along all four directions, it is often possible to identify its main guiding motive.

[2]Furthermore, it is important to keep in mind that there is no one-to-one overlap between the type of generalisation that is relevant for humans and for models. There are several cases in which models generalise better than humans – consider, for instance, calculators, which since long outperform humans when it comes to performing arithmetic operations, and would be useless if they did not – or cases in which it would be useful if they were better than humans, such as generalisation to new languages, which humans above a certain age typically do not excel at.

when they consider a model to have appropriately understood a task. For instance, researchers studying compositional generalisation (see § 3.1) assume that a correct understanding of language implies that the assumed compositional structure of language is modelled. Under that assumption, a model should not have trouble to generalise to new inputs that are generated using the same compositional system. Others instead assume that true language understanding implies being able to use language across a wide variety of tasks (see Section 3.3). Yet others argue that if a model truly captures the relationship between two sentences in NLI tasks (e.g. Bowman et al., 2015a; Marelli et al., 2014; Williams et al., 2018), it should be able to do so across different data sets, even if those were sampled in a slightly different way. In studies that consider generalisation from this perspective, generalisation failures are taken as a proof that the model – in fact – did not properly implement the task it was evaluated on to begin with (but instead showed behaviour that made us think it did, for instance by relying on some spurious patterns or non-generalisable heuristics).

Fairness and inclusivity: does the model generalise in a fair and responsible way? A last yet very important motivation for generalisation research is the desire to have models that are fair, responsible and unbiased. One category of studies driven by these concepts, often ethical in nature, asks questions about how well models generalise to diverse demographics, potentially including minority or marginalised groups (e.g. Bender et al., 2021; Blodgett et al., 2016), or investigates to what extent they perpetuate (undesirable) biases from the data they are trained on (e.g. Dixon et al., 2018; Hutchinson et al., 2020; Sheng et al., 2019). Another important line of research related to both fairness and inclusivity, instead focusses on efficiency, both in terms of the amount of data that is required for a model to converge to a solution, as well as the amount of compute that is required to do so. The relationship of efficiency with generalisation stems from the idea that models that generalise well should learn more quickly, and require fewer data. Efficiency can thus be seen as a correlate of generalisation, and has a strong relation with inclusivity and responsibility: models that can generalise from small amounts of data are more inclusively applicable – for instance for low-resource languages for which little data is available – and models that require less compute to train are more accessible for groups with smaller computational resources, and they have a lower environmental impact (see, e.g. Strubell et al., 2019).

3 What type of generalisation is the test addressing?

A second important consideration when it comes to characterising generalisation tests, is what *type* of generalisation a test aims to evaluate. We identify and describe five broad generalisation types that are frequently considered in the literature. Some of those are rooted in knowledge about human generalisation, such as tests that target compositional (§ 3.1) or structural generalisation (§ 3.2). Others, instead, are motivated by more practical concerns, such as work focussing on generalisation across tasks (§ 3.3), languages (§ 3.4), generalisation across different domains (§ 3.5) or the sensitivity of models to the exact data they are trained on (§ 3.6).

3.1 Compositional generalisation

The first prominent type of generalisation that can be found in the literature is *compositional generalisation*, which is often argued to underpin human's ability to quickly generalise to new data, tasks and domains (Fodor and Pylyshyn, 1988; Lake et al., 2017; Schmidhuber, 1990). Because of this strong connection with humans and human language, work about compositional generalisation often has a primarily cognitive motivation, although practical concerns such as sample efficiency and quick adaptation and good generalisation in low-resource scenarios are frequently mentioned as arguments to consider compositional generalisation (Chaabouni et al., 2021; Linzen, 2020, to give just a few examples). While compositional generalisation has a strong intuitive appeal and clear mathematical definition (Montague,

Figure 2: DH: Infographic that illustrates how train and test differ for different generalisation types.

1970), it is not easy to pin down empirically. For an elaborate account of the different arguments that come into play when defining and evaluating compositionality for a neural network, we refer to Hupkes et al. (2020). Here, we follow Schmidhuber (1990) in defining compositionality as the ability to systematically recombine previously learned elements to map new inputs made up from these elements to their correct output. Because compositionality involves mapping forms (e.g. phrases, sentences, larger pieces of discourse) to their meaning, it is usually evaluated in sequence-to-sequence domains such as sequence classification (e.g. Bowman et al., 2015b; Hupkes et al., 2018; Veldhoen et al., 2016), machine translation (e.g. Dankers et al., 2022; Liu et al., 2021; Raunak et al., 2019), semantic parsing (e.g. Finegan-Dollak et al., 2018; Keysers et al., 2019; Kim and Linzen, 2020; Shaw et al., 2021) or other kinds of generation tasks (e.g. Hupkes et al., 2020; Lake and Baroni, 2018). As far as we know, there have been no explicit systematic attempts to evaluate compositionality in language models (LMs), or in an in-context learning setup.[3] If and how compositionality can be adequately evaluated in such a setup, where the domains of form and meaning are conflated in one space, is a question that is yet to be answered.[4]

In constructing datasplits that require compositional generalisation, researchers often focus on cases that require recombinations of elements (e.g. words, phrases) that did not occur in the training set. Creating or finding such test examples requires a detailed understanding of the underlying structure of the in- and output data, which makes evaluating compositionality in fully natural corpora – rife with ambiguities, exceptions and other kinds of phenomena difficult to capture in fully compositional accounts – a challenging enterprise (see for instance Dankers et al., 2022). The vast majority of tests focussing on compositionality in neural models therefore considers synthetic or generated natural language, where the compositional structure is clear and the underlying structure of the domain fully defined (e.g. Bahdanau et al., 2018; Bastings et al., 2018; Hupkes et al., 2020; Keysers et al., 2019; Lake and Baroni, 2018; Li et al., 2021b; Loula et al., 2018; Mul and Zuidema, 2019; Qiu et al., 2021). More recently, however, more studies have come out that have considered compositionality also in fully natural setups,

[3]There are, however, several studies that focus on *structural* generalisation in such models. Contrary to compositional generalisation, structural generalisation does not focus on the ability of models to correctly interpret new inputs, or assign meanings to them, but only on whether they can generalise to their correct form. We will discuss structural generalisation in the next subsection.

[4]An interesting example to consider in this context is for instance the 5-example qualitative test done by Brown et al. (2020), where they test if GPT-3 can use novel words correctly in a sentence.

using automatic parses of the underlying domain (e.g. Finegan-Dollak et al., 2018; Liu et al., 2021; Shaw et al., 2021). However, with the exception of studies that focus on compositional generalisation to *longer* sequences (Raunak et al., 2019) and the study presented by Dankers et al. (2022), none of these works considers models trained on fully natural training corpora that were not systematically adapted in any way.

3.2 Structural generalisation

Another category of cognitively-inspired generalisation instead focuses on the extent to which models can generalise to correct grammatical forms, rather than if they can *understand* them. Contrary to compositional generalisation, testing structural generalisation does not require two domains: rather than considering whether a model can compositionally assign a correct interpretation to inputs, structural generalisation considers only whether models can generate correct (grammatically or structurally) correct forms. Because of this, structural generalisation is most straightforwardly evaluated in form-only models (i.e. language models). Furthermore, since evaluating structural generalisation requires understanding only the input domain, it is much easier evaluated in completely natural setups, and we will therefore focus only on work that considers structural generalisation in natural language. Structural generalisation studies typically focus on two broad categories: syntactic generalisation, and morphological generalisation. We discuss both of them below.

Syntactic generalisation One category of structural generalisation focuses specifically on *syntactic generalisation*, by testing whether models can generalise to novel syntactic structures, or novel elements in known syntactic structures. For instance, Jumelet et al. (2021) and Weber et al. (2021) consider how models generalise to specific licensing environments for negative polarity items when those are filtered out of the training data. For the recently popular large language models, doing such studies is not computationally feasible, given their training cost. Unfortunately, even generating specific test splits given knowledge of what is in the training data is often not possible for such models, given that their training data is not in the open domain. This makes it impossible to study the relationship between the evaluation and training data, and, consequently, it is hard to assess to what extent the incidental examples reported by the respective release papers are reflective of generalisation. Some interesting exceptions were presented by Wei et al. (2021) and Razeghi et al. (2022). Wei et al. (2021), in particular, investigate how test performance of models on tests reflecting syntactic rule learning in a pre-trained model (BERT, in their case) is affected by a term's training data frequency, by varying those frequency in the training corpus. Razeghi et al. (2022), instead, focus on a larger model trained on more data, and while they do not systematically vary the training corpus, they do an elaborate analysis of how test performance in their trained model (GPT-Z) is affected by absolute and relative frequencies of specific terms in the model's training data.

Note that the vast majority of other studies focussing on the syntactic abilities of (masked) language models (e.g. Giulianelli et al., 2018; Jumelet and Hupkes, 2018; Linzen et al., 2016; Warstadt et al., 2019, 2020), focus more on what kind of abilities models represent, rather than whether those abilies are *generalisations* from something. These works do not (explicitly) consider the relationship between the data they test on and the data that a model was trained on. We will not further discuss these studies, but in our survey (Section 7), we will include a few papers in which there is an implicit yet clear assumption that the test data substantially differs from the training data, for instance because it includes sentences created with semantically nonsensical words (Gulordava et al., 2018), or unusually deep levels of recursion (Lakretz et al., 2021a,b) that are not likely to naturally occur in corpora.

Morphological generalisation A second direction included in the category of structural generalisation focuses on a domain that has been a popular testing ground for questions about human generalisation: morphological inflection. Papers focussing on morphological inflection (e.g. Corkery et al., 2019; Dankers et al., 2021; Kirov and Cotterell, 2018; Liu and Hulden, 2022; Malouf, 2017; McCurdy et al., 2020) are typically rooted in strong cognitive motivations. While most of this work considers i.i.d. train/test splits (e.g. several previous SIGMORPHON shared tasks, Cotterell et al., 2018, 2017, 2016), recent ones have focused on how morphological transducer models generalise across languages (McCarthy et al., 2019; Pimentel et al., 2021a; Vylomova et al., 2020). Further, a few recent works (Calderone et al., 2021; Li and Wilson, 2021; Liu and Hulden, 2022; Pimentel et al., 2021b; Szolnok et al., 2021; Wilson and Li, 2021) attempt to evaluate these transducers' generalisation within each language, taking inspiration from *wug* tests which are used in psycholinguistics to probe morphological generalisation to novel words in humans (Berko, 1958; Marcus et al., 1995). In principle, such studies could also be conducted with large language models but the lack of access to the training data is, again, a complication for determining whether the novel words are truly unseen.

3.3 Generalisation across tasks

A third and completely different direction of generalisation research considers the ability of a single model to adapt to multiple NLP problems. We refer to this type of generalisation with the term *task* generalisation. Along with the nature of models used in NLP, also the nature of tests considering task generalisation has quite substantially changed in the past ten years, which we will discuss in the present section.

Multitask learning Initially, across-task generalisation was strongly connected to transfer and multitask learning (Collobert and Weston, 2008). In multitask learning, a model is either trained on a set of tasks and evaluated on those same tasks, or pretrained on some tasks and then adapted to others. As this setup favours approaches that benefit from positive transfer across tasks, it implicitly studies forms of cross-task generalisation.[5] Examples of benchmarks that were originally meant to address this kind of cross-task transfer – although they are not used as such anymore – are multitask benchmarks like DecaNLP (McCann et al., 2018), GLUE (Wang et al., 2018) and the latter's successor SuperGLUE (Wang et al., 2019). In recent times, a common approach has been to formulate all tasks as sequence-to-sequence problems, as explored in DecaNLP (McCann et al., 2018), and by T5 (Raffel et al., 2020), exT5 (Aribandi et al., 2022) and UnifiedSKG (Xie et al., 2022), among others.

The pretrain-finetune paradigm In the context of multitask learning, across-task generalisation was an extremely challenging topic. The relatively recently introduced *pretrain-finetune paradigm*, however, has not only substantially changed that, but has also shifted thoughts on how to evaluate generalisation. Rather than evaluating how learning one task can benefit from another, this paradigm instead gives a central role to the question of how well a model that has acquired some general knowledge about language during pretraining can be used to generalise to different kinds of tasks in a finetuning stage, introducing a second round of training involving task specific parameters (e.g. Devlin et al., 2019; Liu et al., 2019b). Interestingly, in this setup the tasks themselves are typically evaluated with random train/test splits in the finetuning stage, and thus do not necessarily consider generalisation at the level of individual tasks.

In-context and zero-shot learning More recently, the focus shifted even further, to a scenario in which is considered how well pretrained language models generalise to different tasks *without* any ad-

[5]Noteably, as illustrated by the work of Weber et al. (2021), the definition of *task* can be taken liberally in this context, ranging from traditional notions of tasks, to considering subparts of something seen as a single task as separate tasks.

ditional parameters.[6] In the most extreme case, this implies evaluating a language model directly on a range of tasks without any further training. To do so, tasks are reformulated tasks as text-completion problems, such that language models can be *prompted* directly with a question representing a specific task (*zero-shot learning*), potentially preceeded by a few examples (*few-shot learning*) (Radford et al., 2019). Datasets to do so are typically created by adapting conventional multitask datasets, where prompting templates are (often manually) designed for each task (e.g. Mishra et al., 2022; Wang et al., 2022; Weller et al., 2020). Similarly to work focussing on structural generalisation in large language models, studies that investigate the relationship between the training and test data are rare, and there are many open questions in that domain. Where Brown et al. (2020) report that data leakage from training had a small impact on their results, other recent work suggests that the impressive capabilities of large language models on zero- or few-shot learning tasks can largely be attributed to the presence of similar or identical examples in the training corpus (Han and Tsvetkov, 2022; Razeghi et al., 2022). Moreover, models have been reported to be sensitive to exact task formulation (Jiang et al., 2020; Schick and Schütze, 2021) and even the order of the examples given in the few-shot setting (Lu et al., 2022), to some extent contradicting the intutive idea of task understanding (and thus generalisation).

In-context finetuning A different range of studies that considers task evaluation in the prompting setup for which the relationship with generalisation is more clear, is the class of studies that finetunes a pretrained model with prompts from one set of tasks and then evaluates them on another set of tasks (e.g. Sanh et al., 2022; Wei et al., 2022; Zhong et al., 2021). While also in this case the pretraining corpus is uncontrolled, at least the relationship between the finetuning train and test data can be clearly monitored, and the performances on the test data with and without finetuning easily compared. Nevertheless, there are few studies that actually do so.

3.4 Cross-lingual generalisation

A fourth type of generalisation, which has recently gained in popularity with the strong improvements on English models, is generalisation across *languages*. Cross-lingual generalisation is highly relevant from a practical perspective: while the data for a selected amount of languages (English in particular) is plentiful, for many others, resources are much more scarce or virtually non-existent. Strong generalisation across languages would be beneficial for increasing the coverage of the amount of languages that we have adequate models for, and as such contributes to the democratisation and inclusiveness of NLP.

Cross-lingual finetuning There are several ways in which cross-lingual generalisation can be evaluated. Most existing cross-lingual studies focus on the scenario where labelled data is available in a single language (typically English), and the model is evaluated in multiple languages. A common approach to address this is to finetune a multilingual language model on the English training data, and zero-shot transfer to the rest of the languages (e.g. Papadimitriou et al., 2021; Pires et al., 2019; Wu and Dredze, 2019).[7] For instance, Pires et al. (2019) show that Multilingual BERT (Devlin et al., 2019) finetuned on English generalises well even to languages with different scripts, but exhibits some systematic deficiencies that affect specific language pairs. Papadimitriou et al. (2021), instead, investigate how grammatical features generalise across languages for the same Multilingual BERT model. There is a large amount of benchmarks available to investigate cross-lingual generalisation to different tasks, which we will discuss below.

[6] If the pretraining corpus is seen as a large collection of different uncontrolled task, this scenario is more similar to the original multitask learning scenario than the pretrain-finetune paradigm.

[7] Other approaches instead use machine translation to translate the test set into English and directly use an English model, or to translate the training data into another language and fineune a multilingual model on the augmented data. As this setup does not focus on generalisation per se, but rather depends on the quality of the translation model, we will not further discuss it.

Multilingual learning A second way in which cross-lingual generalisation can be evaluated, is by considering whether models trained on multiple languages at the same time (multilingual models) perform better than models trained on only one language. In multitask learning, approaches that are simultaneously trained on multiple tasks can be seen as an implicit evaluation of generalisation across tasks. Similarly, multilingual models trained on multiple languages can be seen as implicitly evaluating generalisation across languages. There is a large number of papers that investigates and proposes multilingual models, mostly in the domains of language modelling and machine translation (e.g. Aharoni et al., 2019; Al-Shedivat and Parikh, 2019; Costa-jussà et al., 2022; Fan et al., 2021; Freedman et al.; Zhang et al., 2020). Most of these papers have as main aim to introduce improved models, and they are not motivated by generalisation questions. Some, however, do include explicit generalisation experiments in their setup. For instance, Zhou et al. (2018) investigate how generalisation depends on the amount of data added for different languages; Aharoni et al. (2019) investigate how zero-shot generalisation changes depending on the amount of different languages that a model is trained on.

Multilingual benchmarks As pointed out before, while the field focussing on multilingual modelling is vast and is associated with many interesting generalisation questions, papers in this area do not often focus explicitly on generalisation. We would, therefore, like to end this subsection by discussing the most important benchmarks available to evaluate generalisation in this context. Benchmarks or datasets used to evaluate cross-lingual generalisation are created in a variety of different ways. Several benchmarks are created by translating monolingual benchmarks into different languages, usually through a professional translation service (Artetxe et al., 2020; Conneau et al., 2018; Ebrahimi et al., 2022; FitzGerald et al., 2022; Lewis et al., 2020; Li et al., 2021a; Lin et al., 2021; Longpre et al., 2021; Mostafazadeh et al., 2016; Ponti et al., 2020; Williams et al., 2018; Xu et al., 2020; Yang et al., 2019; Zhang et al., 2019). Other multilingual benchmarks, instead, have been built by separately annotating each language via its native speakers (e.g. Adelani et al., 2021; Asai et al., 2021; Clark et al., 2020; Muller et al., 2021). Another way to construct multilingual benchmarks is to leverage existing resources that cover multiple languages. For instance, several multilingual summarisation datasets have been created by extracting article-summary pairs from online newspapers or how-to guides (e.g. Hasan et al., 2021; Ladhak et al., 2020; Nguyen and Daumé III, 2019; Scialom et al., 2020; Varab and Schluter, 2021). Also Wikipedia has been used as a resource to derive multilingual benchmarks (Botha et al., 2020; Liu et al., 2019a; Pan et al., 2017; Rahimi et al., 2019). Similarly, various linguistic resources have been used to derive multilingual benchmarks: for instance, the Universal Dependencies treebank (Nivre et al., 2020) has been used to evaluate cross-lingual part-of-speech tagging, and Raganato et al. (2020) used multilingual WordNet and Wiktionary to build XL-WiC, an extension of WiC (Pilehvar and Camacho-Collados, 2019) which reformulates word sense disambiguation in 12 languages as a binary classification task. Finally, there are also several aggregated benchmarks that include selected sets of benchmarks previously proposed by others, similar to GLUE and SuperGLUE in English (Hu et al., 2020; Liang et al., 2020; Ruder et al., 2021; Wang et al., 2022), which allow to evaluate cross-task and cross-language generalisation simultaneously.

3.5 Domain generalisation

For the types of generalisation we have discussed so far, datasets were often quite deliberately split to target specific kinds of generalisation behaviour. The next category we consider, instead, considers a type of generalisation that occurs more naturally, and is very important in practical scenarios: generalisation to different domains. For instance, a sentiment analysis model might be trained to classify the sentiment of reviews for some products, and then needs to generalise to newly developed products, that were not in its training data (Ryu et al., 2018; Tan et al., 2019); a model trained on data collected from one demographic needs to generalise to the entire population (Blodgett et al., 2016); and a machine translation model trained on canonical text needs to generalise to noisy data from an internet platform (Blodgett et al., 2017; Michel and Neubig, 2018) or to data from a different domain (Malinin et al., 2021).

While there is not a precise definition of what constitutes a domain, different domains broadly refer to collections of texts exhibiting different topical and/or stylistic properties, such as different genres or formality levels. For instance, MultiNLI (Williams et al., 2018) collected training corpora from five different sources, and included both an in-domain evaluation set with corpora from those five sources, and an out-of-domain evaluation set with corpora from five different sources. Blodgett et al. (2016) consider how language tools trained on data collected from white African-American speakers generalises to text from non-white ones. Fried et al. (2019) compare how neural and non-neural constituency parsers generalise out-of-domain, whereas Artetxe et al. (2021) compare how sparse and dense language models generalise in-domain and out-of-domain. Kamath et al. (2020) study the problem of selective question answering under domain shift, where the test distribution includes both in-domain and out-of-domain questions and the model must abstain from answering when not confident. Connected to that, there is a substantial body of work in out-of-domain detection (Hendrycks et al., 2020; Lane et al., 2007; Ryu et al., 2017, 2018; Tan et al., 2019).

Domain generalisation has often been studied in connection with domain adaptation, where an existing general model needs to be adapted to a new domain (Daumé III, 2007). This has been a very active research area in machine translation (Axelrod et al., 2011; Bertoldi and Federico, 2009; Chu et al., 2017; Chu and Wang, 2018; Freitag and Al-Onaizan, 2016; Hu et al., 2019; Joty et al., 2015; Koehn and Schroeder, 2007; Luong and Manning, 2015; Wang et al., 2017a,b), with several standard datasets (Malinin et al., 2021; Michel and Neubig, 2018) and dedicated tracks in popular shared tasks like WMT (Bojar et al., 2019; Specia et al., 2020). In addition to machine translation, domain adaptation has also been studied in other tasks like part-of-speech tagging (Blitzer et al., 2006), sentiment analysis (Blitzer et al., 2007) and language model pre-training (Gururangan et al., 2020), among others.

Finally, domain generalisation is closely related to temporal generalisation, where the training data encompasses a specific time period and the model needs to generalise to a different time period, either in the future or in the past. This problem has been studied in the context of language modelling (Lazaridou et al., 2021), named entity recognition in social media (Derczynski et al., 2016; Fromreide et al., 2014; Rijhwani and Preotiuc-Pietro, 2020), named entity disambiguation (Agarwal et al., 2018), document classification (He et al., 2018; Huang and Paul, 2018, 2019) and sentiment analysis (Lukes and Søgaard, 2018), among others.

3.6 Robustness evaluation

One last category of generalisation research studies shifts that stem from the data collection process. Different from the previous categories, such shifts are generally unintended and can be hard to spot. As such, existing research focuses on characterising such phenomena and understanding their impact. Oftentimes, studies intend to show that models do not generalise in the way we would expect them to, because the training data was in some very subtle manner not representative of the true target distribution. Such studies start from the idea that generalising solutions should abstract away over specific, often spurious correlations that may occur in the training data, and instead learn the underlying generalising solution associated with the task (e.g. Gururangan et al., 2018; McCoy et al., 2019; Talman and Chatzikyriakidis, 2019). In other words, such studies thus investigate how robustly models generalise, independently from the exact data that they are trained on. We refer to this type of training with the term *robustness evaluation*. Robustness evaluation is very important from a practical perspective. If a model has a strong sensitivity to spurious patterns in the training data, this can result in overestimating the performance of models – either generally or for specific use cases – with potentially harmful consequences, for instance when a model does not generalise well to particular population demographics.

Annotation artefacts Overestimation may occur when there are *annotation artefacts* in the training data. Datasets collected through crowdsourcing depend strongly on how the annotation procedure was set up, which often results in subtle artefacts. For instance, annotators may naturally tend to minimise their cognitive effort, resorting to patterns that models learn to exploit. Popular NLI datasets like SNLI (Bowman et al., 2015a) and MultiNLI (Williams et al., 2018) have been found to be particularly susceptible to such artefacts. For instance, Gururangan et al. (2018) and Poliak et al. (2018) showed that a hypothesis-only baseline performs better than chance by exploiting spurious patterns in word choice and grammatical features (e.g. negation being indicative of the *contradiction* class). Similarly, McCoy et al. (2019) showed that NLI models rely on syntactic heuristics, and Talman and Chatzikyriakidis (2019) demonstrated that NLI models do not generalise well across different datasets. Besides NLI, other tasks like question answering have also been reported to suffer from annotation artifacts (Jia and Liang, 2017; Kaushik and Lipton, 2018). Finally, Lewis et al. (2021) showed that open-domain question answering datasets have a high-overlap between train and test instances, revealing that memorisation plays a bigger role in these benchmarks than previously assumed.

Subpopulation bias More harmful consequences of overestimation are visible especially in the case where certain demographics are under- or over-represented in the training data and this results in models generalising poorly to specific demographic groups. For instance, Dixon et al. (2018) show that toxicity classifiers suffer from unintended bias, caused by certain identity terms being disproportionately represented in the training data (e.g. *"I am a gay man"* being assigned high toxicity scores because of *"gay"* being often used in toxic comments). Similarly, Park et al. (2018) show that abusive language detection models exhibit gender bias, which is caused by the training data being imbalanced. Finally, Blodgett et al. (2016) show that dependency parsing and language identification tools perform poorly on text from non-white African-American speakers. Robustness evaluation can thus be relevant not only from the perspective of intrinsic task understanding – somewhat akin to how cross-validation is used – but it is also particularly important from a practical and fairness perspective.

4 Shift type: what kind of shift is considered?

As we have seen in the previous section, tests to evaluate generalisation differ in terms of their *motivation* and the *type* of generalisation that they target. What they instead share, is that they all focus on cases in which there is a form of *data shift* between the data a model was (pre)trained on and the data the model was evaluated on. In the third axis of our taxonomy, we consider more explicitly how the shifts between datasets used in a generalisation experiment can be characterised. To be able to more formally describe those shifts, we define the *data distributions* involved in generalisation tests as follows:

$$p(\mathbf{x}_{\mathrm{tst}}, \mathbf{y}_{\mathrm{tst}}) \qquad \texttt{test} \tag{1}$$

$$p(\mathbf{x}_{\mathrm{tr}}, \mathbf{y}_{\mathrm{tr}}) \qquad \texttt{training / finetuning} \tag{2}$$

$$p(\mathbf{x}_{\mathrm{ptr}}, \mathbf{y}_{\mathrm{ptr}}) \qquad \texttt{pretraining} \tag{3}$$

In generalisation research, there are three main ways in which the (pre)training and test data can differ from each other. We formalise these differences as shifts between data distributions[8], which can be expressed as the products of the probability of the input data $p(\mathbf{x})$ and the conditional probability of

[8]For clarity, we leave pretraining distributions aside and focus on train-test shifts, as this is the most intuitive setting. However, the shifts described in this section can be used to describe the relation between any two data distributions involved in the training process.

the output labels given the input $p(\mathbf{y}|\mathbf{x})$:

$$p(\mathbf{x}_{\mathrm{tr}}, \mathbf{y}_{\mathrm{tr}}) = p(\mathbf{x}_{\mathrm{tr}})\, p(\mathbf{y}_{\mathrm{tr}}|\mathbf{x}_{\mathrm{tr}}) \tag{4}$$

$$p(\mathbf{x}_{\mathrm{tst}}, \mathbf{y}_{\mathrm{tst}}) = p(\mathbf{x}_{\mathrm{tst}})\, p(\mathbf{y}_{\mathrm{tst}}|\mathbf{x}_{\mathrm{tst}}) \tag{5}$$

The four terms on the right hand side of Eq. 4 and 5 define four main types of relations between two data distributions. One of those types constitutes the case in which both $p(\mathbf{x}_{\mathrm{tr}}) = p(\mathbf{x}_{\mathrm{tst}})$, and $p(\mathbf{y}_{\mathrm{tr}}|\mathbf{x}_{\mathrm{tr}}) = p(\mathbf{y}_{\mathrm{tst}}|\mathbf{x}_{\mathrm{tst}})$. In this case, there is no shift in data distributions, which matches the i.i.d. evaluation setup that is traditionally used in machine learning. As discussed earlier, this type of evaluation, also referred to as *within-distribution generalisation*, has frequently been reported not to be indicative of good performance for the more complex forms of generalisation that we often desire from our models. We will therefore not further discuss it here, but instead focus on the other three cases, commonly referred to as *out-of-distribution* (o.o.d.) evaluation.[9]

Covariate shift The most commonly considered shift in o.o.d. generalisation research is the case in which $p(\mathbf{x}_{\mathrm{tst}}) \neq p(\mathbf{x}_{\mathrm{tr}})$. In this scenario, often referred as *covariate shift* (Moreno-Torres et al., 2012; Storkey, 2009), the conditional probability of the labels given the input describing the *task* does not change, but the distribution of the data $p(\mathbf{x})$ that it is applied to does. With this type of shift, one thus evaluates if a model has learned this underlying task distribution while only being exposed to $p(\mathbf{x}_{\mathrm{tr}}, \mathbf{y}_{\mathrm{tr}})$.

Virtually all research in NLP considering evaluation generalisation at the model or training procedure level focuses on covariate shift. For example, challenge test sets such as HANS (McCoy et al., 2019), PAWS (Yang et al., 2019), or COGS (Kim and Linzen, 2020) contains a test set with of deliberately unusual, out-of-distribution examples, selected or generated to violate invalid heuristics in assigning labels to data samples. Less deliberate cases of covariate shift, on the other hand, are evaluated in out-of-domain evaluation datasets, such as the sentiment analysis datasets presented by Tan et al. (2019) and Ryu et al. (2018). In their case, the process by which the sentiment of a sentence is to be computed is assumed not to change, but the data that this process needs to be applied to does. Of the three o.o.d. shifts we discuss in this section, covariate shift is also the only shift that can be solved without performing additional training or pre- or postprocessing. As we will see in the next paragraphs, a common approach to address other, more complex shifts, is to turn them into covariate shifts.

Label shift A second potential shift concerns the case in which there is no difference between the input distributions, $p(\mathbf{x}_{\mathrm{tst}}) \neq p(\mathbf{x}_{\mathrm{tr}})$, but instead in the conditional distribution of the labels/output: $p(\mathbf{y}_{\mathrm{tst}}|\mathbf{x}_{\mathrm{tst}}) \neq p(\mathbf{y}_{\mathrm{tr}}|\mathbf{x}_{\mathrm{tr}})$ (). Label shift can happen within the same task when there is a change of domain – e.g. the phrase *it doesn't run* can lead to different sentiment labels depending on whether it appears in a review for software or one for mascara; when there are inter-annotator disagreements; or when there is a temporal shift in the data (see § 3.5). Another common case of label shift is a change in task (as in § 3.3), where the meaning of the labels themselves changes as well. For example, the same sentence needs to be analysed for sentiment in some cases, and judged for toxicity in others. In even

[9] While ideally, all research considering generalisation would explicitly consider the relationship between the data distributions they use in their experiments, there are several examples of studies that claim to be about generalisation in which it is instead *assumed* that there is a shift between train and test data, but this is not actually verified. In some cases, the assumed shift is not explicitly checked because it is considered plausible given general (linguistic) knowledge about language. Consider, for instance, how Gulordava et al. (2018) and Lakretz et al. (2021b), as discussed earlier in Section 3.2, regard sentences with semantically non-sensical words and unusually deep levels of recursion as out-of-distribution with respect to the training data. In other cases, the relationship between training and testing data is not investigated because the researchers do not have access to the training data. Some of the tasks presented in the BigBench benchmark (Srivastava et al., 2022), for instance, contain several tasks that might measure generalisation, but the training datasets of the models investigated are not in the public domain. In other cases, the training data is available to the authors of the paper, but simply no extensive analysis is presented (e.g. Brown et al., 2020; Chowdhery et al., 2022). In our survey, we also consider this body of work, which we mark *assumed shift*.

	$P(\mathbf{x})$	$P(\mathbf{y}\|\mathbf{x})$
No shift		
Covariate shift	✓	
Label shift		✓
Full shift	✓	✓

Table 1: Types of data distribution shifts. DH: replace with figure

more extreme cases, the labels themselves might be changing, for example when shifting from language modelling to POS-tagging. These situations constitute a shift in $p(\mathbf{y}_{\text{tst}}|\mathbf{x}_{\text{tst}}) \neq p(\mathbf{y}_{\text{tr}}|\mathbf{x}_{\text{tr}})$, while the input distribution $p(\mathbf{x})$ stays exactly the same. At the model or training level, label shift is an obstacle that needs to be overcome, rather than directly evaluated: if the same example has a different label in training and test data, this is not something that can be solved with generalisation.

There are two main ways in which label shift is typically addressed, and turned into a generalisation problem. The first is by adding an additional finetuning, or continual learning phase. In that scenario, there is a label shift between the pretraining and finetuning training data, but not between the finetuning training and testing data. The level at which generalisation is (somewhat implicitly) evaluated in that case, is at the pretraining level: does my pretraining model adapt well to different conditional label distributions when further trained? The second way to address label shift is to augment the input data with domain or task indicators. We saw before that the phrase *it doesn't run* can be both positive and negative, depending what it describes. Without further information, it is impossible for a model to infer the correct meaning. However, if we add an indicator that specifies the domain (review for mascara, review for software), the problem is converted into a covariate shift (or potentially even no shift), which then can be solved by correct generalisation. Something similar happens in the in-context-learning setup: by adding a *prompt* that describes what needs to be done with the input, label shifts representing different tasks are turned into a shift that can be solved without further finetuning. That new shift – which might be a covariate shift or no shift at all, depending on the data that the model was trained on – can then be evaluated at the model instance or potentially training level.

Full shift The most extreme case of shift is the case in which both $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$ change simultaneously: $p(\mathbf{x}_{\text{tst}}) \neq p(\mathbf{x}_{\text{tr}}), p(\mathbf{y}_{\text{tst}}|\mathbf{x}_{\text{tst}}) \neq p(\mathbf{y}_{\text{tr}}|\mathbf{x}_{\text{tr}})$. We may encounter such a situation when switching languages in sequence-to-sequence or classification tasks (as described in § 3.4); when changing modality, as from linguistic to visual processing (Lu et al., 2021); or when switching data types completely from language to gameplay (Ciolino et al., 2020), robotics (Jang et al., 2021), and other non-linguistic (Papadimitriou and Jurafsky, 2020) or non-textual data (Kao and Lee, 2021). Like label shift, these *full shifts* are not evaluated at the model instance or training level, but need to be considered at the pretraining level, or turned into a different type of shift that can be addressed at the model instance or training level.

5 Data sources: how are the train and test data produced?

In the previous section, we considered what kind of shifts may occur in generalisation tests. Another relevant dimension, concerns how that shift was produced or found, or, in other words, what is the *source* of the differences occurring between the pretraining, training and test data distributions. Do shifts naturally occur between existing corpora, or are they the result of deliberate *splitting* of a corpus? Is the *test set* generated or selected with a particular kind of shift in mind, or is *all data* involved generated? In the fourth axis of our taxonomy, we consider how the pretraining, training and test data distributions – and the shifts between them – are produced. We distinguish four different sources of shifts: i) *naturally*

Figure 3: DH: Figure that illustrates different sources of splits/shifts.

occurring shifts, describing scenarios in which a generalisation test considers shifts occurring naturally between different corpora; ii) *splits of natural corpora*, in which both the training and pretraining data are fully natural, but they are partitioned along a specific dimension; iii) *generated shifts*, where the training data is natural, but the test data is designed with a specific shift in mind; and iv) *fully generated datasets*, where all data involved in generated.

To formalise the description of these different sources of shift, we consider the unobserved *base distribution* which describes all data considered in an evaluation test:

$$p(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}}, \boldsymbol{\tau}) \qquad \qquad \texttt{base} \tag{6}$$

The variable $\boldsymbol{\tau}$ represents a *data property of interest*, with respect to which a specific generalisation ability is tested. This can be an observable property of the data (e.g. the length of an input sentence), an unobservable property (e.g. the timestamp that defines when a data point was produced), or even a property relative to the model (architecture) under investigation (e.g. τ could represent how quickly a data point was learned in relation to overall model convergence). The base distribution over $\mathbf{x}$, $\mathbf{y}$ and $\boldsymbol{\tau}$ can be used to define different partition schemes, which can be adopted in generalisation experiments. Formally, such a partitioning scheme is a rule $f : \mathcal{T} \rightarrow \{\texttt{pretrain, train, test}\}$ that discriminates data points according to a property $\boldsymbol{\tau} \in \mathcal{T}$. To investigate how a partitioning scheme impacts model behaviour, the pretraining, training and test distributions can be defined as:

$$p(\mathbf{x}_{\text{ptr}}, \mathbf{y}_{\text{ptr}}) = p(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}} \,|\, f(\boldsymbol{\tau}) = \texttt{pretrain}) \tag{7}$$

$$p(\mathbf{x}_{\text{tr}}, \mathbf{y}_{\text{tr}}) = p(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}} \,|\, f(\boldsymbol{\tau}) = \texttt{train}) \tag{8}$$

$$p(\mathbf{x}_{\text{tst}}, \mathbf{y}_{\text{tst}}) = p(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}} \,|\, f(\boldsymbol{\tau}) = \texttt{test}) \tag{9}$$

Using these data descriptions, we can now discuss four different sources of shifts.

Naturally occurring shifts The first option we consider is the case in which shifts naturally occur between different corpora. Such shifts correspond to the case in which the variable $\boldsymbol{\tau}$ refers to properties of the data that naturally differ between collected datasets. What characterises this type of shift source, is that both the data partitions of interest are naturally occurring corpora, to which no systematic operations are applied: for the purposes of a generalisation test, experimenters have no direct control over the partitioning scheme $f(\boldsymbol{\tau})$. Examples of naturally occurring shifts emerge from splits containing data from different annotators, sources or domains (e.g. Artetxe et al., 2021; Talman and Chatzikyriakidis, 2019), data sampled from different populations (e.g Dixon et al., 2018; Talat et al., 2018) or data from different points in time (e.g. Lazaridou et al., 2021). This category also includes separately collected corpora targeting the same task, such as MNLI (Williams et al., 2018) and WNLI (Wang et al., 2018).

Splits of natural corpora A slightly less natural setup is the one in which a natural corpus is considered, but it is split along very specific dimensions. The primary difference with the previous category is that the variable $\boldsymbol{\tau}$ refers to data properties along which data would not naturally be split, such as the length or complexity of a sample, and thus that experimenters have control over the partitioning scheme $f(\boldsymbol{\tau})$. Raunak et al. (2020), for instance, split naturally occurring machine translation corpora such that longer sentences occur in the test data, and Weber et al. (2021) split a language modelling corpus such that the training data does not contain specific types of NPI licensers. Other examples of natural data splits could be splits that maximise compound divergence to investigate compositionality (Keysers et al., 2019).[10]

Generated shifts The third category on our source of shift axis concerns the case in which one data partition (usually the *training* set) is a fully natural corpus, but the other partition is designed with specific properties in mind, to address a generalisation aspect of interest. Not only do the experimenters control the partitioning scheme, but they can also influence the underlying base distributions (Eq. 6) by arbitrarily constructing one of the partitions. Data in the constructed partition may avoid simple syntactic patterns (), violate heuristics about gender (), or include unusually long sequences (). As an example of this shift source, Dankers et al. (2022) investigate compositionality in MT models trained on fully natural corpora by constructing test data that addresses compositional generalisation given the specific properties of the training corpus. For NLI, McCoy et al. (2019) design a test set that cannot be solved with specific heuristics. Another category of studies that fit into this type are those with *adversarial* test sets, generated either by humans (Kiela et al., 2021) or automatically using a specific model (). In examples above, all of the constructed data occurs in the test data; note that the opposite – where instead the *training data* is synthetic or generated and the test data natural – is also possible, yet less common.[11]

Fully generated or selected splits The last source of shift are splits that use only generated, or even fully synthetic data. Generating data is often the most precise way of measuring the inductive bias of a model or whether a particular structure is transferred successfully, as experimenters have direct control over both the base distribution and the partitioning scheme. Sometimes the data involved is entirely synthetic (e.g. Hupkes et al., 2020; Lake and Baroni, 2018), other times it is templated natural language, or a narrow selection of an actual natural language corpus (e.g Keysers et al., 2019; Kim and Linzen, 2020). Generated splits can vary in a number of different dimensions. Sometimes, $\boldsymbol{\tau}$ is a simple observable data property. For instance, Hupkes et al. (2020) split their corpus based on the presence of particular function pairs $\mathcal{P}$, implicitly setting $\boldsymbol{\tau} = \mathcal{P} \in x$. In some cases, $\boldsymbol{\tau}$ may also be defined relative to the $\boldsymbol{\tau}$ of other examples, and can only be computed globally, such as in the case of *maximum compound divergence* splitting (Keysers et al., 2019).

Figure 4: DH: Figure that illustrates different shift loci.

6 Locus of shift: between which data distributions does the shift occur?

In the previous sections, we have discussed high-level motivations for studying generalisation in neural NLP models, types of generalisation that have been frequently evaluated in the literature, kinds of data distribution shifts, and possible sources of data shift. These four axes demonstrate the depth and breadth of generalisation evaluation research, and they also clearly illustrate that generalisation is evaluated in a wide range of different experimental setups. What we have not discussed yet, is between which data distributions those shifts can occur (the *locus* of the shift), and how that impacts which part of the modelling pipeline is evaluated.

Given the three data distributions that we have considered in § 4, there are four possible loci of shifts: shifts only between the *training and the test data*, shifts only between the *pretraining and the training data*, shifts only between the *pretraining and the test data*, and shifts between *all data distributions*. The locus of shift determines what component of the modelling pipeline is assessed by a generalisation test, and thus impacts what kind of generalisation questions can be asked. We describe the loci of shift as well as how they interact with the different components in the modelling pipeline with the aid of three *modelling distributions*. These modelling distributions correspond to the different stages in contemporary machine learning pipelines – testing a model, training, and potentially pretraining it:

$$p(\mathcal{Y}_{\text{tst}} \mid \mathcal{X}_{\text{tst}}, \boldsymbol{\theta}^*) \qquad \texttt{model} \tag{10}$$

$$p(\boldsymbol{\theta}^* \mid \mathcal{X}_{\text{tr}}, \mathcal{Y}_{\text{tr}}, \boldsymbol{\phi}_{tr}, \hat{\boldsymbol{\theta}}) \qquad \texttt{training/finetuning} \tag{11}$$

$$p(\hat{\boldsymbol{\theta}} \mid \mathcal{X}_{\text{ptr}}, \mathcal{Y}_{\text{ptr}}, \boldsymbol{\phi}_{pr}, \boldsymbol{\theta}_0) \qquad \texttt{pretraining} \tag{12}$$

where $\boldsymbol{\phi}$ broadly denotes training and pretraining hyperparameters, $\boldsymbol{\theta}$ refers to model parameters, and $\mathcal{X}, \mathcal{Y}$ indicate sets of inputs $(\mathbf{x})$ and their corresponding output $(\mathbf{y})$. Each pair $(\mathbf{x}, \mathbf{y})$ is assumed to be sampled from one of the data distributions described in the previous sections (Equation 1-3). In short, Equation 10 defines a model instance, which specifies the probability distribution over the target test labels $\mathcal{Y}_{\text{tst}}$, given the model's parameters $\boldsymbol{\theta}^*$ and a set of test inputs $\mathcal{X}_{\text{tst}}$. Equation 11, instead, defines a

[10]Keysers et al. (2019) themselves do not apply this split to fully natural data

[11]For instance Papadimitriou and Jurafsky (2020) investigate whether pretraining on *music* can help learning natural language in a second stage.

training procedure, specifying a probability distribution over model parameters $\boldsymbol{\theta}^* \in \mathbb{R}^d$ given a training dataset $\mathcal{X}_{\mathrm{tr}}$, $\mathcal{Y}_{\mathrm{tr}}$, a set of training hyperparameters $\boldsymbol{\phi}_{tr}$, and a (potentially pretrained) model initialisation $\hat{\boldsymbol{\theta}}$. Lastly, Equation 12 defines a pretraining procedure, specifying a conditional probability over the set of parameters $\hat{\boldsymbol{\theta}}$, given a pretraining dataset, a set of pretraining hyperparameters $\boldsymbol{\phi}_{pr}$, and a model initialisation.[12] Where a shift occurs, impacts which of these modelling distributions can be evaluated. We discuss the different potential loci of shifts below.

From training to test data Probably the most commonly occuring shifts in generalisation experiments are shifts between train and test data. They occur in the classic setup where a model is trained on a certain partition of the base distribution, and then directly evaluated on a shifted (out-of-distribution) test partition. Experiments characterised by this locus of shift are, for example, those testing compositional (see § 3.1) and structural generalisation (§ 3.2), and frequently also domain generalisation (§ 3.5). Virtually only covariate shifts are investigated in this setup, as generalising to label or full shifts without an additional training stage is typically not possible in NLP scenarios (see § 4). When the data shift is placed between training and test data, two parts of the modelling pipeline can be assessed. In some cases, researchers might investigate the generalisation abilities of a particular *model instance* (i.e. a set of parameters $\boldsymbol{\theta}^*$). In those cases, researchers focus on the evaluation of a single model instance – typically made available by others – without considering how exactly it was trained, and how that impacted the model's generalisation behaviour. In other cases, researchers instead evaluate one or more training procedures, by considering if the *training distribution* results in model instances that generalise well[13] – for example to study whether training with bidirectional input yields better generalising model instances than training with uni-directional signal. While also this case requires evaluating model instances, the focus of evaluation is not on one particular model instance, but rather on the procedure that generated model instances.

From pretraining to training data A second potential locus of shift is between the pretraining and training corpus. This locus is common in the pretrain-finetune experimental paradigm, as well as in work considering domain adaptation. In the former case, the shifts considered are often label shifts (different tasks) or full shifts (for instance, when a multilingual model is finetuned to perform a task in one language and then tested on another). The latter case typically focuses on covariate shifts, for instance where a (language) model is trained on one domain, and then finetuned and tested on text from a different genre. The part of the modelling pipeline that is typically (implicitly) tested in the pretrain-training locus, is the *pretraining distribution*: the researcher evaluate whether that procedure, described in Equation 12 results in parameter sets $\hat{\boldsymbol{\theta}}$ that generalise well when further trained.[14] Note that evaluating a shift from pretraining to training data necessarily also requires a second training stage, but that evaluating generalisation at that locus requires an additional shift between the finetune and test data. We consider such cases, in which there are two loci of shifts, below, as a separate category.

From pretraining to test data The third potential locus of shift occurs between pretraining to testing data. Such shifts are less common, but may occur when a pretrained model is not further updated, but evaluated directly – as frequently happens in in-context learning setups ($\mathcal{X}_{\mathrm{tr}}, \mathcal{Y}_{\mathrm{tr}} = \emptyset, \emptyset$) – or when a pretrained model is finetuned on examples that are i.i.d. with respect to the pretraining data and then tested on out-of-distribution instances. The former case, where $\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}$, is similar to the first locus discussed, but distinguishes itself by the nature of the (pre)training procedure, which typically has a

[12] Note that this formalisation generalises to the *training from scratch* paradigm when $\mathcal{X}_{\mathrm{ptr}}, \mathcal{Y}_{\mathrm{ptr}} = \emptyset, \emptyset$, and to the *in-context-learning* setup when $\mathcal{X}_{\mathrm{tr}}, \mathcal{Y}_{\mathrm{tr}} = \emptyset, \emptyset$.

[13] I.e. whether the training distribution places high probability mass on model instances that generalise well.

[14] Or, in other words, whether the pretraining procedure places probability mass on model parameters $\hat{\boldsymbol{\theta}}$ that when further trained (Equation 11) results in model instances that generalise well.

general purpose objective, rather than being trained for a specific task. Furthermore, studies that consider the pretrain-test locus of shift often consider models that are frequently further finetuned by others. Also similarly to the the training-to-test locus, the shifts occurring from pretraining to test data are most often covariate shifts, due to the challenging nature of zero- or few-shot generalisation to more extreme shift types.[15] However, very frequently, specifically in-context-learning based setups do not explicitly consider the relationship between training and test data, but merely assume a shift occurs (e.g. Radford et al., 2019). The stage in the modelling distribution that is typically considered for the pretrain-to-test locus is the pretraining distribution described in Equation 12 – formally: does the pretraining distribution place probability mass on parameters $\hat{\theta}$ that prove successful when *directly tested* for generalisation. – or the model instance that resulted from that procedure.

Between pretraining, training, and test data In the last scenario we consider, data distribution shifts occur both between the pretraining and training data, as well as between the training and test data.[16] In these experiments, there are two loci of shift, and their nature may often not be the same: the shift from pretraining to training can be of any type, while the shift from training to test is typically a less extreme covariate shift (for the same reasons discussed in the previous two paragraphs). Full shifts include cases where a pretrained language model is finetuned on one set of tasks and tested on another; when it is finetuned on one domain and then tested on another (double covariate shift); or when a it is finetuned on a new task and tested on the same task, but with data from a different domain (label + covariate shift). Considering full shift cases allows researchers to evaluate all components of the modelling pipeline at once, analogously to what was described in the previous three paragraphs. To adequately represent those cases, we will mark them as *double shifts* in the shift type axis.

7 A survey of existing generalisation research

In the previous sections, we have presented a taxonomy containing five (sometimes interconnected) axes along which generalisation research can been characterised, providing examples for each of the different positions studies might take on those axes. Now, we use our taxonomy to characterise existing generalisation research[17], with the aim to create a comprehensive map of generalisation research, and to identify gaps. A three-dimensional version of this map is presented in Figure **??**; for the full version, we refer to `https://genbench.github.io`, where also instructions to contribute to the survey can be found. In this section, we discuss the most important findings.

DH: This section will be finished after we finish the survey, and the plots. In terms of content, it will contain the most important findings: which areas are well represented, which areas instead could use some work, are there any other things that stand out?

8 Discussion

DH: In this section we will recap and summarise our work, and also make recommendations for future work. This will also include a description of our the website, and a commitment to add new tests that

[15] As briefly discussed before, more complicated shifts are in this setup are often transformed into covariate shifts, by adding additional content to the input distribution x. For instance, to reiterate our previous example: by adding a domain indicator (software, make-up) to a review, a label shift is turned into a covariate shift.

[16] We do not distinguish cases where the test data is shifted with respect to the pretraining data from cases where it is not, as the latter are very uncommon. It is, however, possible to set up an experiment where the pretraining and test data are drawn from the same distribution, for example to test whether a finetuning procedure results in catastrophic forgetting.

[17] It is very likely that we have missed some studies. If you believe we have missed a study that presents a data set to evaluate generalisation, or that we have misqualified your study on one of the axes, please reach out to the main author of this paper, so that we can include / correct it!

will be sent to us.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Prabal Agarwal, Jannik Strötgen, Luciano del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. diaNED: Time-aware named entity disambiguation for diachronic corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 686–693, Melbourne, Australia. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. 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 1184–1197, Minneapolis, Minnesota. Association for Computational Linguistics.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2022. Ext5: Towards extreme multi-task scaling for transfer learning. In *International Conference on Learning Representations*.

Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giri Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Mona Diab, Zornitsa Kozareva, and Ves Stoyanov. 2021. Efficient large scale language modeling with mixtures of experts.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference* *of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. 2018. Systematic generalization: What is required and can it be learned? In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Marco Baroni. 2021. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *CoRR*, abs/2106.08694.

Jasmijn Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho, and Douwe Kiela. 2018. Jump to better conclusions: SCAN both left and right. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 47–55, Brussels, Belgium. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Jean Berko. 1958. The child's learning of English morphology. *Word*, 14(2-3):150–177.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece. Association for Computational Linguistics.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2017. A dataset and classifier for recognizing social media English. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 56–61, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors. 2019. *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Association for Computational Linguistics, Florence, Italy.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R Bowman, Christopher D Manning, and Christopher Potts. 2015b. Tree-structured composition in neural networks without tree-structured architectures. In *Proceedings of the 2015th International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches*, pages 37–42.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Basilio Calderone, Nabil Hathout, and Olivier Bonami. 2021. Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 274–282, Online. Association for Computational Linguistics.

Rahma Chaabouni, Roberto Dessì, and Eugene Kharitonov. 2021. Can transformers jump around right in natural language? assessing performance transfer from SCAN. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 136–148, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Matthew Ciolino, David Noever, and Josh Kalin. 2020. The Go transformer: Natural language modeling for game play. *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*, pages 23–26.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Michael John Collins. 1996. A new statistical parser based on bigram lexical dependencies. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191, Santa Cruz, California, USA. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. 2019. Are we there yet? encoder-decoder neural networks as cognitive models of English past tense inflection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3868–3877, Florence, Italy. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of* *the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.

Verna Dankers, Anna Langedijk, Kate McCurdy, Adina Williams, and Dieuwke Hupkes. 2021. Generalising to German plural noun classes, from the perspective of a recurrent neural network. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 94–108, Online. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Yann Dubois, Gautier Dagan, Dieuwke Hupkes, and Elia Bruni. 2020. Location Attention for Extrapolation to Longer Sequences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 403–413, Online. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-SQL evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 351–360, Melbourne, Australia. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Richard G Freedman, Rodrigo De Salvo Braz, Hung Bui, and Sriraam Natarajan. Initial empirical evaluation of anytime lifted belief propagation.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation.

Daniel Fried, Nikita Kitaev, and Dan Klein. 2019. Cross-domain generalization of neural constituency parsers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 323–330, Florence, Italy. Association for Computational Linguistics.

Hege Fromreide, Dirk Hovy, and Anders Søgaard. 2014. Crowdsourcing and annotating NER for Twitter #drift. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2544–2547, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Xiaochuang Han and Yulia Tsvetkov. 2022. Orca: Interpreting prompted language models via locating supporting data evidence in the ocean of pretraining data.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Yu He, Jianxin Li, Yangqiu Song, Mutian He, and Hao Peng. 2018. Time-evolving text classification with deep neural networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2241–2247. International Joint Conferences on Artificial Intelligence Organization.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699, Melbourne, Australia. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123, Florence, Italy. Association for Computational Linguistics.

Dieuwke Hupkes, Sara , and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Dieuwke Hupkes. 2020. Hierarchy and interpretability in neural models of language processing.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intellgence Research*, 67:757–795.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. 2021. BC-Z: Zero-shot task generalization with robotic imitation learning. *ArXiv*, abs/2202.02005.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Shafiq Joty, Hassan Sajjad, Nadir Durrani, Kamla Al-Mannai, Ahmed Abdelali, and Stephan Vogel. 2015. How to avoid unwanted pregnancies: Domain adaptation using neural network models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1259–1270, Lisbon, Portugal. Association for Computational Linguistics.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Wei-Tsung Kao and Hung-yi Lee. 2021. Is BERT a cross-disciplinary knowledge learner? a surprising finding of pre-trained models' transferability. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2195–2208, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.

Temuulen Khishigsuren, Gábor Bella, Khuyagbaatar Batsuren, Abed Alhakim Freihat, Nandu Chandran Nair, Amarsanaa Ganbold, Hadi Khalilia, Yamini Chandrashekar, and Fausto Giunchiglia. 2022. Using linguistic typology to enrich multilingual lexicons: the case of lexical gaps in kinship. *CoRR*, abs/2204.05049.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. 2021. A survey of generalisation in deep reinforcement learning. *CoRR*, abs/2111.09794.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic. Association for Computational Linguistics.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 4487–4499.

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

Yair Lakretz, Theo Desbordes, Dieuwke Hupkes, and Stanislas Dehaene. 2021a. Causal transformers perform below chance on recursive nested constructions, unlike humans. *CoRR*, abs/2110.07240.

Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. 2021b. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699. Special Issue in Honour of Jacques Mehler, Cognition's founding editor.

Ian Lane, Tatsuya Kawahara, Tomoko Matsui, and Satoshi Nakamura. 2007. Out-of-domain utterance detection using classification confidences of multiple topics. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):150–161.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European*

Chapter of the Association for Computational Linguistics: Main Volume, pages 1000–1008, Online. Association for Computational Linguistics.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021a. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Jane S.Y. Li and Colin Wilson. 2021. Leveraging paradigmatic information in inflection acceptability prediction: The JHU-SFU submission to SIGMORPHON shared task 0.2. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 289–294, Online. Association for Computational Linguistics.

Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. 2021b. On compositional generalization of neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4767–4780, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019a. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy. Association for Computational Linguistics.

Ling Liu and Mans Hulden. 2022. Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–749, Dublin, Ireland. Association for Computational Linguistics.

Linqing Liu, Patrick S. H. Lewis, Sebastian Riedel, and Pontus Stenetorp. 2021. Challenges in generalization in open domain question answering. *CoRR*, abs/2109.01156.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

João Loula, Marco Baroni, and Brenden Lake. 2018. Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 108–114, Brussels, Belgium. Association for Computational Linguistics.

Kevin Lu, Aditya Grover, P. Abbeel, and Igor Mordatch. 2021. Pretrained transformers as universal computation engines. *ArXiv*, abs/2103.05247.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Jan Lukes and Anders Søgaard. 2018. Sentiment analysis under temporal shift. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–71, Brussels, Belgium. Association for Computational Linguistics.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

David M. Magerman. 1995. Statistical decision-tree models for parsing. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 276–283, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Andrey Malinin, Neil Band, Yarin Gal, Mark J. F. Gales, Alexander Ganshin, German Chesnokov, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panagiotis Tigas, and Boris Yangel. 2021. Shifts: A dataset of real distributional shift across multiple large-scale tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Robert Malouf. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology*, 27(4):431–458.

Gary F Marcus, Ursula Brinkmann, Harald Clahsen, Richard Wiese, and Steven Pinker. 1995. German inflection: The exception that proves the rule. *Cognitive psychology*, 29(3):189–256.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756, Online. Association for Computational Linguistics.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Richard Montague. 1970. Universal grammar. *Theoria*, 36(3):373–398.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapa Nakashole. 2020. Rethinking self-attention: Towards interpretability in neural parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 731–742, Online. Association for Computational Linguistics.

Mathijs Mul and Willem Zuidema. 2019. Siamese recurrent networks learn first-order logic reasoning and exhibit zero-shot compositional generalization. In *CoRR, abs/1906.00180*.

Benjamin Muller, Luca Soldaini, Rik Koncel-Kedziorski, Eric Lind, and Alessandro Moschitti. 2021. Cross-lingual genqa: Open-domain question answering with answer sentence generation.

Khanh Nguyen and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, Online. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben

Ambridge, and Ekaterina Vylomova. 2021a. Sigmorphon 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021b. Sigmorphon 2021 shared task on morphological reinflection part 2: Are we there yet? a shared task on cognitively plausible morphological inflection.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *arXiv preprint arXiv:1608.07836*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Linlu Qiu, Hexiang Hu, Bowen Zhang, Peter Shaw, and Fei Sha. 2021. Systematic generalization on gSCAN: What is nearly solved and what is next? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2180–2188, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Vikas Raunak, Siddharth Dalmia, Vivek Gupta, and Florian Metze. 2020. On long-tailed phenomena in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3088–3095, Online. Association for Computational Linguistics.

Vikas Raunak, Vaibhav Kumar, Florian Metze, and Jaimie Callan. 2019. On compositionality in neural machine translation. In *NeurIPS 2019 Context and Compositionality in Biological and Artificial Neural Systems Workshop*.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot reasoning. *CoRR*, abs/2202.07206.

Shruti Rijhwani and Daniel Preotiuc-Pietro. 2020. Temporally-informed analysis of named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7605–7617, Online. Association for Computational Linguistics.

Roni Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modeling.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Seonghan Ryu, Seokhwan Kim, Junhwi Choi, Hwanjo Yu, and Gary Geunbae Lee. 2017. Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems. *Pattern Recognition Letters*, 88:26–32.

Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. 2018. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 714–718, Brussels, Belgium. Association for Computational Linguistics.

Federico Sangati and Willem Zuidema. 2011. Accurate parsing with compact tree-substitution grammars: Double-DOP. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 84–95, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Jürgen Schmidhuber. 1990. Towards compositional learning in dynamic networks.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 922–938, Online. Association for Computational Linguistics.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *CoRR*, abs/2108.13624.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online. Association for 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. *CoRR*.

Amos Storkey. 2009. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Gábor Szolnok, Botond Barta, Dorina Lakatos, and Judit Ács. 2021. Bme submission for sigmorphon 2021 shared task 0. a three step training approach with data augmentation for morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 268–273, Online. Association for Computational Linguistics.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Online harassment*, pages 29–55. Springer.

Aarne Talman and Stergios Chatzikyriakidis. 2019. Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy. Association for Computational Linguistics.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-domain detection for low-resource text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3566–3572, Hong Kong, China. Association for Computational Linguistics.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sara Veldhoen, Dieuwke Hupkes, and Willem Zuidema. 2016. Diagnostic classifiers: Revealing how neural networks process hierarchical structure. In *Proceedings of the NIPS2016 Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017a. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada. Association for Computational Linguistics.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017b. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Benchmarking generalization via in-context instructions on 1,600+ language tasks.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Lucas Weber, Jaap Jumelet, Elia Bruni, and Dieuwke Hupkes. 2021. Language modelling as a multitask problem. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2049–2060, Online. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

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.

Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E. Peters. 2020. Learning from task descriptions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1361–1375, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Wilson and Jane S.Y. Li. 2021. Were we there already? applying minimal generalization to the sigmorphon-unimorph shared task on cognitively plausible morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 283–291, Online. Association for Computational Linguistics.

Francis CK Wong and William SY Wang. 2007. Generalisation towards combinatorial productivity in language acquisition by simple recurrent networks. In *2007 International Conference on Integration of Knowledge Intensive Multi-Agent Systems*, pages 139–144. IEEE.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

Kaiyu Yang and Jia Deng. 2020. Strongly incremental constituency parsing with graph neural networks. *Advances in Neural Information Processing Systems*, 33:21687–21698.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2856–2878, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhong Zhou, Matthias Sperber, and Alexander Waibel. 2018. Massively parallel cross-lingual learning in low-resource target language translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 232–243, Brussels, Belgium. Association for Computational Linguistics.