Evaluating Large Language Models along Dimensions of Language Variation: A Systematik Invesdigatiom uv Cross-lingual Generalization

Niyati Bafna, Kenton Murray, and David Yarowsky
Johns Hopkins University, Center for Language and Speech Processing
{nbafna1,kenton,yarowsky}@jhu.edu

Abstract

While large language models exhibit certain cross-lingual generalization capabilities, they suffer from performance degradation (PD) on unseen closely-related languages (CRLs) and dialects relative to their high-resource language neighbour (HRLN). However, we currently lack a fundamental understanding of what kinds of linguistic distances contribute to PD, and to what extent. Furthermore, studies of cross-lingual generalization are confounded by unknown quantities of CRL language traces in the training data, and by the frequent lack of availability of evaluation data in lower-resource related languages and dialects. To address these issues, we model phonological, morphological, and lexical distance as Bayesian noise processes to synthesize artificial languages that are controllably distant from the HRLN. We analyse PD as a function of underlying noise parameters, offering insights on model robustness to isolated and composed linguistic phenomena, and the impact of task and HRL characteristics on PD. We calculate parameter posteriors on real CRL-HRLN pair data and show that they follow computed trends of artificial languages, demonstrating the viability of our noisers. Our framework offers a cheap solution for estimating task performance on an unseen CRL given HRLN performance using its posteriors, as well as for diagnosing observed PD on a CRL in terms of its linguistic distances from its HRLN, and opens doors to principled methods of mitigating performance degradation.[1]

1 Introduction

Advances in the capabilities of large language models (LLMs) have resulted in a paradigm shift in natural language processing, with LLMs being used for and evaluated over a variety of classification and generation tasks (Xue et al., 2021; Bang et al., 2023a; Hendy et al., 2023). However, even multilingual models such as `bloomz7b1`, `mT0` (Muennighoff et al., 2023) and Aya (Üstün et al., 2024) only extend model capabilities to 100 of the world's highest-resourced languages. The vast majority of the world's 3800 written languages have drastically less data available (Joshi et al., 2020), although many have a related high-resource neighbour (Asai et al., 2023). This underscores the need for cross-lingual generalization in LLM capabilities from high-resource languages on which they have been trained to related low-resource languages (LRLs), variants, and dialects, i.e. a theoretical language continuum centered at the high-resource language.

hin: … उन्होंने टैक्सी, चीयर्स, और द ट्रेसी उल्मैन
mai: … ओ टैक्सी, चीयर्स आ द ट्रेसी उलमैन
…he-nom Taxi, Cheers, and The Tracy Ullman

hin: <u>शो</u> जैसे <u>शो</u> में काम किया था।
mai: <u>कार्यक्रम</u> जेहन <u>नाटक</u> में काज कएलनि।
Show like shows in work did

p^{hin}: …he worked on shows such as Taxi, Cheers, and The Tracy Ullman Show. ✅
p^{mai}: …Maithili phrase: O Taxi, Cheers to Tracy Ullman, a play written by Ken Kailani was performed in the theatre. ❌

Figure 1: Phonological/orthographic, morphological, and function and content word variation, and <u>lexical choice</u> difference, between `hin` and `mai`; p^*: `bloomz7b1` MT output.

Previous literature has reported evidence of multilingual and cross-lingual zero-shot capabilities in LLMs for a number of tasks, also finding, unsurprisingly, that model performance suffers in such settings (Jiao et al., 2023; Cahyawijaya et al., 2024) (see Figure 1). While it's reasonable that the farther a closely-related language (CRL) is to its high-resource language neighbour (HRLN), the greater the performance degradation (PD) in a zero-shot setting, we lack a principled understanding of how much different dimensions of linguistic distance (phonological, morphological, and lexical) affect

[1] `https://github.com/niyatibafna/llm-eval-crosslingual-generalization`

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 18742–18762
November 12-16, 2024 ©2024 Association for Computational Linguistics

PD. Given that we can find a systematic relationship between each such dimension and PD, and compute the associated distance between a CRL-HRLN pair, this insight would allow us to (a) diagnose observed PD on a CRL, (b) estimate PD for a CRL without task data, as well as (c) suggest targeted interventions aimed at mitigation of PD.

In this work, we model phonological/orthographic, morphological, and lexical distance as cross-linguistic "noise", generated by Bayesian processes applied on a source language, thus positing a parametrization of the HRL dialect continuum. We generate artificial languages with varying extents of each noise type, and study LLM zero-shot cross-lingual generalization for three NLU-focused tasks. We discuss the effects of task, noise type, and language family on PD. Crucially, our noise generation processes have tractable posteriors cheaply computable from bilingual lexicons/bitext. This allows us to place real CRLs within the parametrized dialect space of a HRL. We show that PD on real CRLs given their posteriors follows expected trends observed over artificial languages, demonstrating that our noise processes capture useful information about the factors of linguistic distance as they contribute to PD.

Our use of artificial languages allows us to systematically populate the dialect space of an HRL; further, the noise generation process produces task datasets for each hypothetical language. This solves three problems: firstly, we often do not have task data for real closely-related languages that are unseen in our LLM; secondly, we may not have enough CRLs per HRL, especially CRLs of varying distance along each dimension of interest, to be able to establish and study systematic trends for that language family. Further, we are not guaranteed that a given CRL or its task data is entirely unseen from the training data, confounding a study of LLM zero-shot generalization. Our main contributions are as follows:

- We study the dimensions of linguistic distance that make an input closely-related language difficult relative to its high-resource language neighbour for an LLM in zero-shot settings, quantitatively and qualitatively describing model robustness to each dimension, and discuss the relevance of the task under consideration and the typology and resource-level of the language.
- We introduce a parametrization of the dialect space of a language along three linguistic axes that allows for the generation of artificial languages given a set of parameters, as well as for cheaply computing the parameters of a real language pair. We demonstrate its utility for predicting and analysing LLM PD on unseen languages using real CRL-HRLN pairs. Our framework also opens pathways to mitigating PD on low-resource languages, e.g., by reducing damaging distances using linguistic or other tools.

2 Modelling linguistic variation

We model **p**honological/orthographic, **m**orphological, and lexical (**c**ontent and **f**unction word) variation as parametrized probabilistic "noisers" applied to a source language to generate related languages. We denote a noiser as ϕ^n_v, parametrized $\theta_n = v$, where $n \in \{p, m, c, f\}$ indicates the noise type. For every language, task, and ϕ^n, we are interested in the function $\psi^n_* : \theta_n \to PD$, where

$$PD = \frac{(s_\theta - s_{\text{rand}}) - (b - s_{\text{rand}})}{b - s_{\text{rand}}} \tag{1}$$

Here, s_θ is the performance on the noised source, b is the score on the clean source, and s_{rand} is the random baseline.[2] This notation extends to composite noisers, e.g. $\psi^{m,c}_{0.5,*}$ computes PD as a function of θ_c, given $\theta_m = 0.5$. See examples of the outputs of our noisers in Table 1 and § D.1.

2.1 Noiser details

ϕ^p: Phonological/Orthographic This model mimics sound change in closely related languages, and is based on the following ideas from theories of sound change (Joseph et al., 2003): (i) Sound change is applied to a phoneme given some phonological left and right context e.g. (`d |a_,_EOW)`$\to$`t`). (ii) Sound change, given context, is regular: it applies consistently in all words of the language. (iii) Consonant sound change largely occurs between phonologically similar phonemes (e.g. difference in voicing: `f`$\to$`v`). This is not relevant for vowels, which change fluidly.

We use manually constructed character$\to$IPA maps to obtain a set of potential underlying phonemes for script characters. For any given occurrence of a character, we make a random guess for its corresponding phoneme if there are several.[3]

[2] 0 for `X`$\to$`eng`, 33.33 for `XNLI`, 50 for `XSC`; i.e. if `XNLI` score drops to 33.33%, we say that it shows 100% PD.

[3] Since our goal is to inject random noise into the input roughly guided by the underlying phonology of the text, we can tolerate the imprecision introduced by this process.

We model phonological context as the left and right character of the source character (including word boundaries); thus, a `(phoneme, context)` pair is simply a character 3-gram. Each `(phoneme, context)` is affected with probability θ_p. In order to find a phonologically plausible target set for each IPA character, we construct a list of IPA character sets covering all phonemes used by the languages in this study, such that the phonemes in each set differ from each other in roughly one (or at most two) phonological features, and a phoneme can plausibly change via sound shift to another phoneme in any of the sets it belongs to. (See Appendix A.) Our list is inspired by Index Diachronica. We can now find a plausible replacement for a given character by mapping it into IPA, sampling a replacement IPA character, and mapping the IPA back into the relevant script. The change to a character given context applies globally throughout the text.

ϕ^m**: Morphological** Our noiser models concatenative suffixation guided by the following intuitive premises. (i) Affixal change is global (iii) The replacement suffix must be plausible for the language family in terms of its phonology and script, and the original suffix, e.g. if one of them starts with a vowel, the other one is also likely to have an initial vowel. We approximate a set of linguistic affixes by collecting the k[4] most common string suffixes of content words in the language corpus. Each collected suffix is noised with probability θ_m, by passing it through the phonological noiser as described above, with a high dial ($\theta_p = 0.5$); this ensures the plausibility of the noised target suffix. Finally, we construct a vocabulary map by swapping out all occurrences of an affected source suffix with its generated target in all source words; the vocabulary map applies globally for every occurrence of the word in the text.

$\phi^{f,c}$**: Lexical** We model function word change and non-cognate content word change separately, guided by the following premises: (i) The replacement non-cognate equivalent for a content word must be plausible in the relevant script, may not resemble the original word at all, and must not be a word in the source vocabulary. Note that we only model complete lexical change and not lexical choice differences: i.e., when languages have different usage patterns or show semantic shift for the same words. (ii) The length of the replacement word may loosely depend on the length of the original word (for example, words with rare semantics may be longer in both dialects). (ii) Function words in related languages are probably distant cognates, very similar in length.

We identify function words in the input using a list of words appearing with relevant UPOS tags in the Universal Dependencies corpus (Nivre et al., 2016) for each language. Note that since functional words are relatively few and highly frequent, collecting them even over small corpora will yield almost perfect coverage for a given language. Any word not in this list is treated as a content word.

For content words, we sample the length of the replacement word from a $\mathrm{Poisson}(\lambda{=}l)$ where l is the length of the source word, and use a character 3-gram model trained on the language task corpus to generate plausible non-words of the required length. For function words, we generate a replacement by applying a high degree of phonological noise to the functional word ($\theta_p = 0.5$). All replacements for content and function words are global.

We study lexical change as a combination of ϕ^c and ϕ^f. Since content word change is the more dynamic of the two, likely to show variation depending on language distance, whereas function word change is likely to be high even for related dialects, and show less variation for differently distant languages, we primarily study the PD dynamics of $\phi^{f,c}_{\theta_f,*}$. We experiment with varying θ_c, given $\theta_f \in \{0,0.5,0.8\}$ ($\phi^{f,c}_{\theta_f,*}$), and with varying θ_f given $\theta_c = 0$ ($\phi^{f,c}_{*,0}$).

Composite We compose noisers by independently applying phonological, morphological, and lexical noise in this order (allowing "overwrites"). While this is a simplification, it is well-motivated; lexical noise is often the most dynamic and continuous of the three while phonological and affixal change are much more gradual and/or fixed given a time period.

2.2 Posterior computation

We now demonstrate the utility of our noisers and associated ψ^n in understanding PD on real linguistic variation. We assume that CRLs are "generated" by applying a composition of noisers on the source language. Now, if we can find the underlying θ_n, we can estimate $PD = \psi^n_*(\theta_n = v)$, and therefore task performance.

Given a bilingual lexicon in the source and target, we use word alignments to estimate the Bernoulli

[4] empirically chosen per language, e.g. $k = 150$ for `hi`.

parameter $\theta \in \{\theta_p, \theta_m, \theta_c, \theta_f\}$. In our noisers, all changes to the concerned units (trigrams, suffixes, words) are global. In reality we may not observe a global change between source and target unit; language change may be noisy, we may have one-off phenomena, and we may have noisy word alignments. We compute θ in the following way:

$$E[\theta] = \frac{\sum_u I_u}{T}, \qquad E\left[\frac{\sum_u I_u}{T}\right] = \sum_u \frac{E[I_u]}{T}$$

where I_u is a binary random variable indicating whether unit u was affected, and T is the total number of units. We can now estimate $E[I_u] = \frac{C_u}{T_u}$ for each u i.e. the fraction of times that u was affected. Note that it remains to be decided how we will categorize a given change in a non-identical source-target pair.

Phonological If source-target normalized edit distance (NED) is high,[5] we attribute changes in the target word to phonological change. We find the minimal list of edits from source to target; if we observe a character change with the same left-right context, we count it towards θ_p.

Morphological If a content target word has a different suffix (identified as in § 2.1) but the same stem but (i.e. it is not lexical change) , we count it towards θ_m.

Lexical We count any change in a function word towards θ_f . For content words, if the source-target NED is low (i.e. not phonological/morphological change) and the target word is not present in the source vocabulary, we count it towards θ_c.

While real languages exhibit the above kinds of noise simultaneously, i.e. they are the result of composite noising, our model of noise composition permits us to compute the posteriors of individual noisers independently of each other. Note that allowing overwrites by successive noisers does not affect this property: although lexical change may "overwrite" a suffix change, it does not change the fraction of suffixes/trigrams affected (i.e. the MLE estimate of θ_m), since the noisers are independent of each other.[6]

[5]We use language-specific empirically determined thresholds for NED-based decisions, e.g. 0.5 for de in this case

[6]We compute θ_m only over words that have the same stem in source and target; any word pair with different stems is ignored. Since lexical noise is applied uniformly over words and independently of morphological noise, we expect that while it will "disqualify" a set of word pairs for the θ_m posterior computation, the remaining set will give us the same estimate (in expectation) of θ_m. An analogous argument applies for θ_p.

3 Experimental Setup

Model and Tasks We obtain initial zero-shot results on a number of tasks for `bloomz7b1`and `mt0XXL` (Muennighoff et al., 2023), and select three tasks to work with: `X`→`eng` machine translation on `FloRes200` (Team et al., 2022),[7] `XStoryCloze` (`XSC`; Lin et al., 2021b), and `XNLI` (Conneau et al., 2018), as covering a large enough mutual set of languages as well as two tasks paradigms of interest, namely, multiple-choice questions and sequence-to-sequence. We found that the performance of both models on multilingual ARC, HellaSwag and MMLU (Dac Lai et al., 2023) is close to or worse than chance for many languages; this makes these tasks unsuitable for studying model PD.

Our experiments are conducted on `bloomz7b1`, using the `mlmm-eval` evaluation framework (Dac Lai et al., 2023). See § B and § C.1 for all evaluated tasks and further experimental details.

Languages We work with Hindi, Indonesian, Arabic, German, French, Spanish, and English. This set of languages was curated with language presence in `bloomz7b1`[8] and availability of task datasets in mind. We include three macrolanguages (`hi`, `id`, `ar`) with dozens of real closely related low-resource languages and dialects. In order to validate our computed trends with real language data, we require languages and dialects related in varying extents to the respective HRLN, unseen from `bloomz7b1`, with task dataset availability. We study trends for `X`→`eng` for the following CRLs: Awadhi-`awa`, Bhojpuri-`bho`, Magahi-`mag`, Maithili-`mai`, and Chhattisgarhi-`hne` (Hindi), Danish-`dan`, Icelandic-`isl`, and Swedish-`swe` (German), Malay-`zsm`-(Indonesian), Occitan-`oci` (French), Galician-`glg` (Spanish), and Iraqi-`acm`, Yemeni-`acq`, Tunisian-`aeb`, Levantine-`ajp`, North Levantine-`apc`, Najdi-`ars`, Moroccan-`ary`, and Egyptian-`arz` (Arabic). This list includes language pairs with a range of degrees of relatedness; e.g. `zsm` and `ind` are much closer than `dan` and `deu` (Dryer, 2013).

4 Results and Discussion

See ψ^n for noiser, task, and language combinations in Figure 2 (single run per noiser parametrization).

[7]We loosely refer to `X`→`eng` as an NLU task; since the LLM is fluent in English, its performance primarily depends on comprehension of the input (Nguyen et al., 2024).

[8]German is "low-resource" for `bloomz7b1`, constituting only 0.21% of the training corpus (Muennighoff et al., 2023).

Noiser	Strategies	Example I/O
ϕ^f_*	(a) Infers sentence meaning from content words (b) Partially correct (c) Incorrectly connects content words* (d) Breaks: Function word was part of a construction (e) Hallucination† (f) No translation/off-target†	s: Pasangan **ini dapat** memilih **untuk** membuat rencana adopsi **bagi** bayi **mereka**. s': Pasangan **eni tawat** memilih **antuk** membuat rencana adopsi **vige** bayi **marequ**. p: The couple may choose to make an adoption plan for their baby. p': The couple decided to adopt a baby. Ref: These couples may choose to make an adoption plan for their baby.
$\phi^{f,c}_{\theta_f,*}$	(a) Guesses correct word from context* (b) Keeps the original word, code-switched, if surrounding context is clear. (c) Keeps the word, garbles sentence (d) Breaks: wrong guess. (e) Ignores the word and translates the rest	s: **Der** Satellit wurde **von** einer Rakete **ins** Weltall **befördert**. s': **Tyh** Satellit wurde **vän** einer Rakete **wange** Weltall **veraumoden**. p: The satellite was sent into space by a rocket. p': The satellite was sent into orbit by a rocket. Ref: The satellite was sent into space by a rocket.
ϕ^p_*	(a) Guesses word meaning from context and spelling clues* (b) Makes a wrong guess. (c) Breaks: function word changes. (d) Breaks: many changes in proximity.	s: **Cualquier** persona **que** esté **programando** un **viaje** a un país **que** podría **tildarse** como zona de guerra debería **recibir** un **entrenamiento** profesional. s': **Cualqeyer** persona **cue** esté **programendo** un **viajo** a un país **cue** podría **tyldurse** como zona de guerra debería **recibor** un **yntrenamiento** profesional. p: Any person planning a trip to a country that could be considered a war zone should receive professional training. p': Any person planning a trip to a country that could be considered a war zone should receive professional training. Ref: Anyone planning a visit to a country that could be considered a war zone should get professional training.
ϕ^m_*	(a) Model faces no issues (b) Breaks: too much corruption*	s:यहाँ सूर्योदय **देखने** की कुछ जगहों पर ईस्टर की पूरी रात **जागने** की परंपरा है। s': यहाँ सूर्योदय **देखनइ** की कुछ जगहों पर ईस्टर की पूरा राट **जागनइ** की परंपरा है। p: There are some places where the Easter night is celebrated by staying up all night. p': In some places, Easter is celebrated with a full moon. Ref: There's a tradition to pass the Easter night awake at some exposed point to see the sunrise.

Table 1: Output type classification for each noise type. * marks the case that the example belongs to. †: applicable to all noisers, only listed once. Example languages from top to bottom: `id, de, es, hi`.

Tasks We find that **the rate of mean PD given a noise type is the same across tasks**. **This indicates that model performance for one task for a CRL relative to its HRLN can be used to extrapolate its performance on other tasks; i.e. PD is largely a function of language distance.**

While we see linear trends for mean PD for all tasks and noise types, and individual languages trends are also linear for `X→eng`, this is less true for individual language trends for `XSC` and `XNLI` (e.g. 3b, 3c, 4b, for `arb,hin`). This is a result of sampling variance in our noising process: ϕ^n_v may produce a range of artificial languages varying in the specific set of units that are noised. The relationship between PD and θ_n is mediated by task sensitivity to the comprehension of specific words (phones/morphs) as opposed to general comprehension of the input: we compute std. deviation of PD for multiple artificial languages generated from the same θ_n for `hi` and `ar`, and find much lower SD for `X→eng` than the other tasks. Using PD means over multiple artificial languages per θ_n removes the instability of the trend at the individual language level and is key to computing reliable trends for a language. See § D.3 for task-wise std. dev. and stabilized trends for `hi` and `ar`.

These findings back the intuition that while translation depends on local understanding of input, suffering predictably with increasing noise, the model relies only on certain words rather than the entire sentence for classification tasks, and is therefore more sensitive to whether those are corrupted rather than the general extent of noise, although of course these two are correlated. `XNLI` in particular is also highly sensitive to whether its three label words are noised, strongly cautioning any zero-shot evaluation to be mindful of its treatment of label words. **This suggests that `X→eng` is a more robust test of NLU in a LRL for a model, and less susceptible to fluke performances.**

Languages We see that `ar` and `id` suffer most from ϕ^m (e.g. 6a), perhaps due to their rich morphology (Lopo and Tanone, 2024), and that `de` particularly suffers from ϕ^c (e.g. 4a), possibly because word compounding results in a higher extent of lost information per noised word. See Figure 3 for mean PD over all parametrizations of a given noiser per language for `X→eng`. In general, we find that lower-resource languages in `bloomz7b1` such as `de`, `ar`, `id`, and `hi` have higher mean PD as compared to HRLs like `fr` and `es`; **more exposure to a language makes the model more adept at unseen related languages.**

Noise types The slope of ψ^n signals how damaging noise type n is (higher is worse). We con-

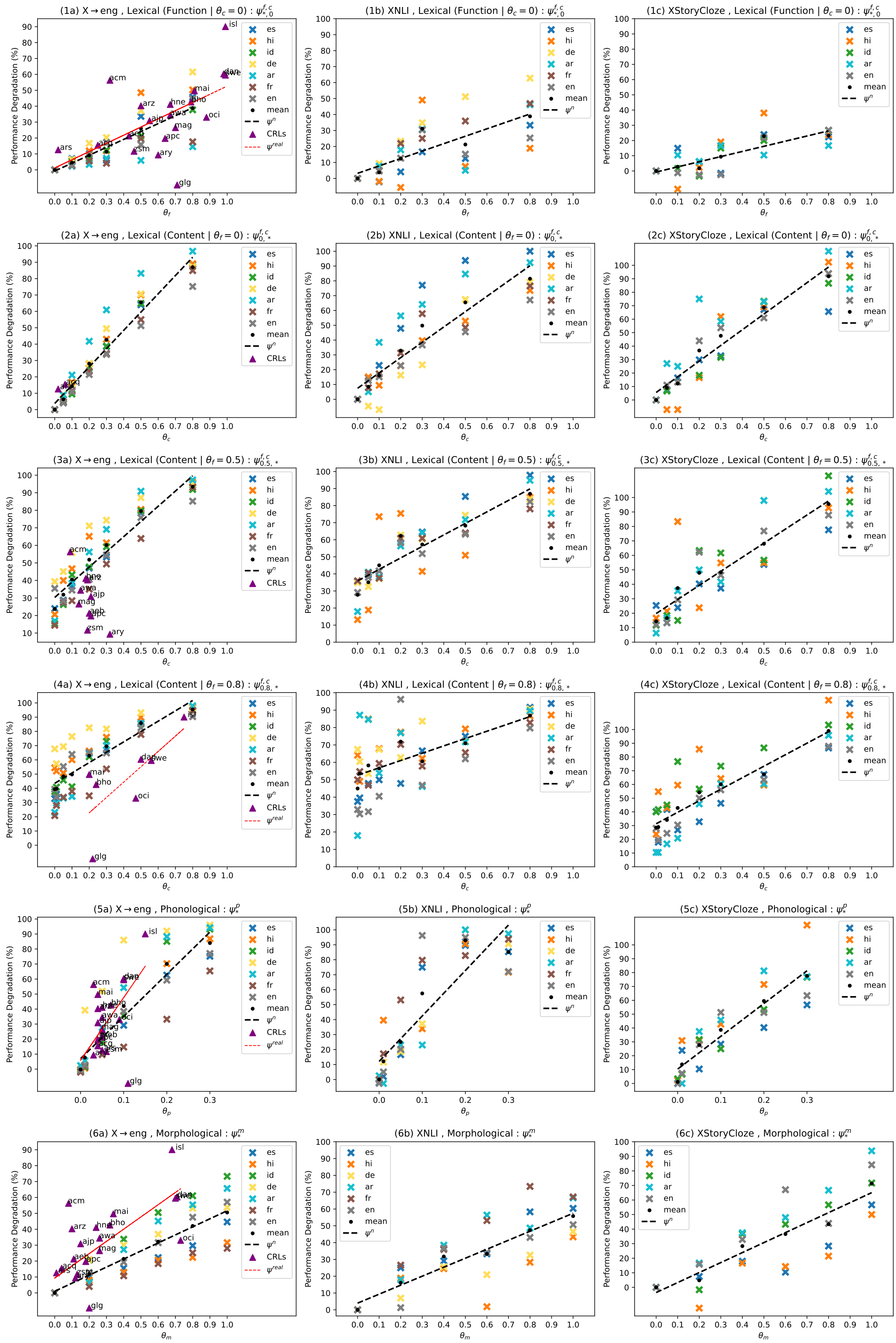

Figure 2: PD% for each language, task, and noiser. ψ^n depicts mean language PD trends. We show $(\theta_n, PD\%)$ points for real CRL-HRLN pairs using computed posteriors for X→eng. See § 3 for corresponding HRLNs per CRL. ψ^{real} depicts trends for real CRLs, shown only when θ_{real} has a wide-enough range.

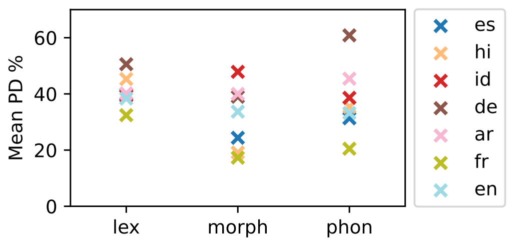

Figure 3: Mean PD over all parametrizations per noiser for X$\rightarrow$eng

textualize these trends over θ using the posteriors computed over real language pairs, which provide a sense of the natural range of θ for related languages per noiser. Note that absolute PD values for a given θ_n, and therefore absolute slopes, are not comparable across noise types, since θ_n differs in meaning depending on the noiser; however, these can be compared directly for different lexical noisers.

We find that $\phi^{f,c}_{*,0}$ shows lower PD rate as compared to $\phi^{f,c}_{0,*}$: naturally, **content loss is more damaging than function word loss**. However, note that real θ_f values are high even for very closely related language pairs (e.g. hne-hin; see 1a), and correspond to significant PD values. On the other hand, θ_c may be low (< 0.2) for closely related languages, but is more costly. Note that $\psi^{f,c}_{\theta_f,*}$ for $\theta_f \in \{0, 0.5, 0.8\}$ have similar slopes but increasing y-intercepts based on θ_f. Given that function words form a closed and relatively small set for a given language, and may be easier to deal with than open class, possibly rare, content words, **this suggests that we can cheaply tackle a non-trivial portion of PD by simply handling "easier" function word correspondences**.

We observe that ψ^m_* displays a low slope; corrupting 100% of our set of linguistic suffixes results in a mean $50 - 70\%$ PD. **This indicates that the model is largely capable of capturing important information from word stems.** Note that for distant related cousins like de-dan, θ_m can be high and correspond to significant PD.

Finally, ψ^p_* indicates sharp PD; this is natural since ϕ^p affects chargrams with possibly widespread effect in the corpus. Once again, while our chosen LRLs cover a range of natural values for θ_p, even very closely-related languages display θ_p values corresponding to significant PD (5a), **suggesting that the model is vulnerable to natural levels of phonological/orthographic variation**.

PD over noise composition While overall PD for a language with composite noising is a presumably a function of PD for each contained noise type, the nature of this function remains to be understood. We study $\phi^{f,c,m}_{0.5,*,0.5}$, composing lexical and morphological noise (see Figure 4 for X$\rightarrow$eng),and observe that the resulting PD is well-explained simply by $\psi^{f,c}_{0.5,*}$; indicating that overall PD may be a simple max (as opposed to incremental) in this case.

We can show that within our framework of composition, the effect of a constant amount of morphological/phonological noise decreases as lexical noise grows, due to an increasing "overwrite" probability in composition.[9] This matches our intuition about linguistic variation well: **as languages grow lexically distant, lexical change becomes the dominating factor in PD**. This is because growing lexical change transforms more words to non-cognates, rendering the underlying phonological or morphological patterns affecting cognates decreasingly relevant. This idea offers one explanation of the observed PD of isl, i.e. that the PD effect is dominated by $\phi^{f,c}_{0.8,*}$. We leave a detailed study of PD for composite noisers as a function of individual noiser PD to future work. We believe that it is likely to depend on the noiser combination (e.g. $\phi^{p,m}$ vs. $\phi^{p,f,c}$), as well as the comparative initial PD for the isolated noisers: whether we are composing equal or imbalanced levels of noise (or resulting PD) from isolated noisers may influence the nature of composition.

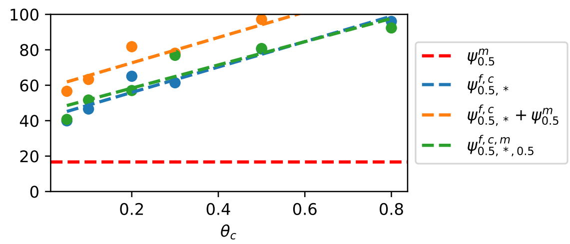

Figure 4: Composing $\phi^{f,c}$ and ϕ^m: studying $\psi^{f,c,m}$ given θ_m for Hindi for X$\rightarrow$eng. $\psi^{f,c} + \psi^m$ shows the theoretical additive trend.

Posteriors and trends for real CRLs We calculate posteriors for real CRLs as described in § 2.2. This procedure requires bilingual lexicons: we obtain these from Google Translate when available, and alternatively use statistical word alignment with FastAlign (Dyer et al., 2013) on FloRes bitext.

[9]See Appendix E for a formal explanation, as well as for $\psi^{f,c,m}_{0.5,*,0.5}$ for XNLI and XSC.

We verify that computed posteriors over possibly noisy alignments are similar to those computed on clean lexicons by comparing posteriors obtained from noisy and manually cleaned lexicons for `mai` and `hne`: we find that this is largely the case for θ_f, θ_p, and θ_m, but that θ_c is prone to being over-estimated from noisy alignments.

We plot (θ, PD) points for X→eng in Figure 2. We bucket the θ_f posterior and show (θ_c, PD) on the relevant $\psi^{f,c}_{\theta_f,*}$ plot. Note that we can use posteriors for a CRL-HRLN pair to generate artificial languages that are equally distant from the HRLN as the LRL; we provide examples in § F.2 to illustrate the plausibility of our noisers and associated posteriors. **We observe that PD vs. θ_n for real languages generally follow similar trends as ψ^n, indicating that our constructed ϕ^n offer useful parametrizations of linguistic distance as it contributes to PD.**

Notable outliers are are `oci`, `zsm`, and `acm` for $\phi^{f,c}$. Further, `glg` actually performs with +4 BLEU over `es` (§ F.1), which is a clear red flag. Such anomalies, where observed PD is much lower than expected PD, could indicate unreported amounts of the language in the training data or, in the case of `glg`, possibly test set leakage.

Note that since real languages contain a composition of all noise types, we expect total PD to be higher than that predicted by any individual ψ^n. However, this is not true, notably observed for ψ^c_* and ψ^f_* (3a, 4a). This is attributable to code-switching and traces of the unseen language in training data. For artificial languages, the cost of a completely unknown word is high (as compared to a partially known, suffix-corrupted word); however, it's likely that the model actually knows some percentage of words identified as unknown by our posterior computation in the real unseen languages. The unknown word may be present in another language than the HRLN (e.g. `fr-oci changement-cambiar`; `cambiar` is a Spanish equivalent), or it may be non-identical but very close to an HRLN synonym (`certain-qualques` - French synonym `quelques`), or it may simply be known because the model has seen data in the "unseen" language. This would have the effect of reducing the absolute PD while maintaining the trend. **The observed delta between the trends gives us an idea of the benefits of multilinguality and language contamination in training data by providing the counterfactual.**

See § F.1 for more details on the effect of noisy alignments on posteriors, and the computed posteriors, associated BLEU scores, and PD for each θ_n and CRL.

Realistic quality of artificial languages We provide examples of our generated artificial languages in Tables 1, 6, 7, 8, and 12 for various HRLNs. The languages appear phonologically and orthographically plausible given the language family of the HRLN, with occasional transgressions. For example, an artificial language may not respect rules for diacritic vs character usage for vowels in Devanagari, or may over-use rarely observed characters in a script. Currently, we also do not have special treatment for named entities, which should ideally remain unnoised.

	#	Noise type	Error class	#
Correct	27			
Partial	17	Morph	Wrong tense / person	6
		Morph+Func	Missing words	6
		Content	Missing words	2
		Morph+Func	Ungrammatical	1
		Morph	Mistranslation	1
		Phon	Missing words	1
Garbled	5	Morph+Func	Incoherent	5
Breaks	19	Morph+Func	Garbage	8
		Morph+Func	Keywords lost	6
		Morph+Func	Repeats input	2
		Content	Garbage	2
		Content	No translation	1
Hallucination	2	Func	Added concepts	2

Table 2: Case study in MT error type classification for `awa-eng`. "Noise type" refers to the type of divergence from Hindi that causes the translation issue.

Error Modes See Table 1 for a qualitative classification of model error modes for each noiser, obtained via a manual examination of outputs over representative θ_n. See an expanded version of this table in § D.1. We also perform a small case study in error type characterization over 70 sentences for `awa-eng` X→eng, shown in Table 2. These qualitative analyses indicate that the model is able to withstand a good extent of phonological, morphological, and function word change (exhibited in its 38.5% near-perfect translations for `awa-eng` MT), but fails in different ways when multiple morphological and function word changes are in close proximity. Entire content word shifts as opposed to lexical choice variation were rare between Awadhi and Hindi in our sample but cause breakage when they occur. We note that these error type and diagnosis distributions will differ based on the language pair under consideration, the nature of the divergence between the CRL and HRLN and their typologies, as well as the LLM proficiency in both.

In general, while PD over a dataset varies smoothly as a function of θ_n, we observe that **success/failure modes over individual inputs are not easily predictable: the model displays both surprising fragility as well as robustness in different cases.**

5 Related Work

Multilingual evaluation of LLMs Recent studies show that LLMs demonstrate certain multilingual capabilities accompanied with performance degradation for LRLs for machine translation (Jiao et al., 2023; Hendy et al., 2023; Robinson et al., 2023) as well as other tasks like POS, NER, and summarization (Lai et al., 2023; Bang et al., 2023b; Asai et al., 2023). Kantharuban et al. (2023) attempt to identify economic, social, and linguistic correlates of MT performance in LLMs for dialects; they find positive correlations for dataset size and lexical similarity among other factors. It is difficult to draw principled insights from such studies about what the bottlenecks for cross-lingual transfer are, since the tested languages may simultaneously vary in their relatedness to high-resource languages, and presence in the pretraining data.

Linguistic distance as a factor in performance Recent work explores providing "missing" linguistic knowledge of LRLs (lexical, morphosyntactic) in LLMs by providing dictionaries, bitext, and grammar books via in-context learning for LRLs (Tanzer et al., 2024; Zhang et al., 2024b,a). Other works look at cleverly choosing shots for the context by exploring the prompt space, choosing exemplars that are "close" to the output using lexical distance (Zhu et al., 2023; Zhang et al., 2024a; Cahyawijaya et al., 2024). However, this search space of what can be provided is large, and we lack an understanding of which linguistic distances LLMs need "help" with: these ideas motivate a study such as ours.

Robustness Earlier studies have looked at robustness of machine translation systems to orthographic variants, typos, and other kinds of noise (Belinkov and Bisk, 2018; Heigold et al., 2018). Moradi and Samwald (2021) perform a similar study of BERT-like models for sentiment analysis, QA, and NER, among other tasks, with the intent of stress-testing LMs against natural user-generated noise such as synonym replacement, common misspellings, and verb tense errors. Wang et al. (2023) discuss the robustness of ChatGPT against adversarial and out-of-distribution input datasets such as ANLI and DDXPlus. Havrilla and Iyer (2024) investigate character-level static and dynamic noise for chain-of-throught prompting processes. As far as we know, ours is the first work to stress test LLMs under noise models of linguistic distance.

6 Conclusion

We study the robustness of an LLM to 4 types of linguistically-motivated (phonological, morphological and lexical) Bayesian noise models on 7 languages and 3 tasks, generating artificially languages controllably distant from a given HRL and computing trends in performance degradation. This allows us to quantitatively and qualitatively characterize the impact of each included factor of linguistic variation on task performance in isolation. Our noisers are amenable to cheap posterior computation; we show that PD for real unseen languages follow expected trends given their computed posteriors, validating our noiser construction. Our work offers a framework for the principled linguistic analysis of cross-lingual generalization and opens avenues in mitigating LLM performance degradation in low-resource settings.

Limitations

Noiser choice Our work is limited by the three linguistic phenomena we study. Notably, we do not study syntactic change, since it is not naturally modeled by our framework of smoothly increasing distances in a hypothetical continuum. This is for mainly two reasons: firstly, there are simply far fewer possible syntactic changes in total (core syntax can be described within 10-15 features); secondly, systematic syntactic change is much rarer in related languages (very few of those features actually change within language families).

It is certainly possible to extend this study to other noisers modeling relevant phenomena in the context of language continua. One example is the phenomenon of semantic shift, whereby words with the same form shift in meaning in related languages, resulting in different lexical choice for the languages (although not lexical change); lexical usage patterns in general may also be of interest. We give an example of this in Figure 1. This can be modeled within our framework as a noiser that moves a word to its synonym with some probability; we leave such ideas to future work.

Noiser design Our noisers incorporate several simplifications from a linguistic standpoint. Each noiser can be further nuanced to increase the plausibility of the resulting synthesized languages; some examples of possible detailing include (a) ϕ^p: using language-family-specific sound change models that weight commonly observed sound changes in that family higher than others (b) ϕ^m: using morphological tools to more accurately identify linguistic suffixes, (c) ϕ^m: modeling other kinds of morphology, e.g. non-concatenative, templatic, prefixal. This is particularly relevant to languages such as Arabic. (d) ϕ^c: introducing weighting by (log) frequency such that commoner words are more likely to be affected by the noiser. Note that some of these changes may introduce complications for posterior computation. We leave it to future work that is interested in particular noisers for particular language families to look into fine-graining noiser design in a given context.

Comprehensiveness: Languages, Tasks, and Models Our insights on PD characterization are limited to the 3 tasks and 7 languages we study, in a zero-context context for `bloomz7b1`. Each of these dimensions can naturally be expanded: it is possible that the observed PD dynamics are different for different models (individual trends for a noiser will certainly differ depending on model, language, and task), or for a few-shot context. We focus on three NLU-oriented tasks for our study; our conclusions about cross-lingual transfer may change for different task paradigms (Ahuja et al., 2022). Further, we are also able to provide our results on real language posteriors only on `X`$\rightarrow$`eng`; we are constrained by task dataset availability for truly low-resource languages. We make our code available and encourage a similar analysis to ours for any new combination of language, model, task, noiser, and experimental setting.

Noiser composition dynamics Our work focuses mainly on PD dynamics for individual noise types to isolate the effect of each linguistic phenomenon, and touches only briefly on the PD dynamics for composed noisers, although our noise processes and posteriors offer natural extensions for noise composition. While we demonstrate the complexity of observed PD dynamics on a single language and single noise composition setup for 3 tasks, we leave a detailed investigation of the same, which should include a large enough selection of noiser combinations for different language typologies, tasks, and parametrizations per noiser, to future work.

Ethics Statement

Our work is motivated by the need to increase language inclusivity in the large language model space; however, this assumes that speakers of these communities desire the incorporation of their languages into such tools, which may not be the case (Bird, 2020). Further, we also acknowledge that striving for zero-shot generalization to CRLs based on LLM capabilities in HRLNs undermines the need to represent CRL-specific culture and perspective of the world in LLMs (Hershcovich et al., 2022).

Acknowledgments

We would like to thank Kaiser Sun and Vilém Zouhar for proof-reading this paper. We also thank the reviewers for their helpful feedback and suggestions.

References

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi task learning for zero shot performance prediction of multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*.

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Xiangru Tang, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. Promptsource: An integrated development environment and repository for natural language prompts. *Preprint*, arXiv:2202.01279.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023a. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *Preprint*, arxiv:2302.04023.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023b. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation. *Preprint*, arxiv:1711.02173.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. *arXiv preprint arXiv:2403.16512*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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*. Association for Computational Linguistics.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

Matthew S Dryer. 2013. The world atlas of language structures online. *Max Planck Institute for Evolutionary Anthropology*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 644–648.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830.

Alex Havrilla and Maia Iyer. 2024. Understanding the Effect of Noise in LLM Training Data with Algorithmic Chains of Thought. *Preprint*, arxiv:2402.04004.

Georg Heigold, Stalin Varanasi, Günter Neumann, and Josef van Genabith. 2018. How Robust Are Character-Based Word Embeddings in Tagging and MT Against Wrod Scramlbing or Randdm Nouse? In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 68–80, Boston, MA. Association for Machine Translation in the Americas.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *Preprint*, arxiv:2302.09210.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. *Preprint*, arxiv:2301.08745.

Brian D Joseph, Richard D Janda, and Barbara S Vance. 2003. *The handbook of historical linguistics*. Wiley Online Library.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the Dialect Gap and its Correlates Across Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021a. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

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 T. Diab, Veselin Stoyanov, and Xian Li. 2021b. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Joanito Agili Lopo and Radius Tanone. 2024. Constructing and expanding low-resource and underrepresented parallel datasets for indonesian local languages. *arXiv preprint arXiv:2404.01009*.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516, Bangkok, Thailand. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

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.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A Benchmark for Learning to Translate a New Language from One Grammar Book. *Preprint*, arxiv:2309.16575.

NLLB Team, 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, Loic 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. *Preprint*, arXiv:2207.04672.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. *Preprint*, arXiv:2106.12066.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024a. Teaching Large Language Models an Unseen Language on the Fly. *Preprint*, arxiv:2402.19167.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024b. Hire a Linguist!: Learning Endangered Languages with In-Context Linguistic Descriptions. *Preprint*, arxiv:2402.18025.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. *Preprint*, arxiv:2304.04675.

A Details of Phonological Noiser

See Figure 5 for the list of IPA character sets that we used in our phonological noiser described in § 2. An IPA character to be noised can be transformed with uniform probability to another IPA character in any set that it belongs to.

B Baseline results for tasks

See baseline results for bloomz7b1 and mt0XXL for the languages we considered in Table 3 and Table 4 respectively, for multilingual ARC, HellaSwag, MMLU (Dac Lai et al., 2023), X→eng (Team et al., 2022), XSC (Lin et al., 2021b), XNLI (Conneau et al., 2018), XCopa (Roemmele et al., 2011; Ponti et al., 2020), XWinoGrad (Tikhonov and Ryabinin, 2021; Muennighoff et al., 2022), TruthfulQA (Lin et al., 2021a). We see that bloomz7b1 is generally better for XSC and XNLI and work with it for the rest of our experiments. Russian and German are not included in both models but have traces in the training data as described in Muennighoff et al. (2023); we choose to include German in our experiments as a low-resource language in bloomz7b1.

C Further Experimental Details

C.1 Prompt Details and Variations

We tried various prompts for our chosen tasks, and we note that the model performance is highly sensitive to the prompt; this has been observed in several previous studies (Shin et al., 2020; Gao et al., 2021; Schick and Schütze, 2021). We choose a single prompting framework per task with a reasonable baseline performance in line with previous evaluations of bloomz7b1 (Muennighoff et al., 2023). We work in the zero-shot setting for our experiments. This is in keeping with our goal to study zero-shot generalization to unseen languages. While we note some uniform gains from including a few shots $(5 - 10)$ in the high-resource language, we do not study this dimension in our work.

We tried a few different prompting styles inspired by templates from Promptsource (Bach et al., 2022) as well as the defaults in the MLMM evaluation framework (Dac Lai et al., 2023) and noted considerable variation between the worst and best performing prompts (up to 15 points for XNLI and 20 points for XSC). Note that for XNLI and XSC, we see large baseline performance gains when the options are mentioned in the prompt. For XNLI, we also note that Prompt 3 (default) in fact requires the loglikelihood of the entire input sequence to be compared with the corresponding labels replacing [MASK], whereas the other two setups simply compare loglikehoods of the label options. See Table 5.

We also note that for XNLI, model performance is sensitive to the choice of word in the target language for the entailment, neutral, and contradiction labels. Interestingly, using "No" for the Spanish contradiction label results in bloomz7b1 loglikelihood always being highest for contradiction, possibly because it is a shared token with English, yielding near-random performance on xnli_es (33%)

For the translation tasks, we use Prompt 2 for the baselines, but Prompt 1 for the noised languages; we note that this does better than Prompt 2 for the latter.

The above choices give rise to considerable variation in baseline performances; we work with a single setup for our experiments.

Finally, we make the choice to use English instructions for our prompts, resulting in language-mixed inputs. bloomz7b1 is instruction-tuned in this setup, rather than on translated prompt instructions as in the case of mt0XXL-MT (Muennighoff et al., 2023). We do not experiment with translated prompts to eliminate the additional complexity introduced by the quality of the translation.

C.2 Data details

Each evaluation is conducted over a subset of the test set consisting of 300 samples; this is for time and compute efficiency since we conduct a large number of evaluations over combinations of task, language, noiser, and parametrization. Note that all evaluations for a given language and task are conducted over an identical subset.

All datasets used are publicly available for research use under CC BY-NC 4.0 (mARC, mHellaSwag, mMMLU), CC BY-SA 4.0 (XNLI, XStoryCloze, TruthfulQA, XCopa, FloRes200), or CC BY (XWinograd).

C.3 Compute

We conduct a total of approximately $3 * 6 * 7 * 7 = 882$ evaluation experiments (excluding development) on NVIDIA A100 machines, totalling about 220 GPU hours.

	XStoryCloze	XWinograd	XCopa	mARC	mHellaswag	mMMLU	FloRes	TruthfulQA	XNLI
Hindi	63.67	-	-	21.67	33.67	30	56.44	49.08	51
Russian	57.67	54.33	-	19.67	34.33	26	30.31	52.93	38.33
Arabic	66	-	-	26.33	32	32.33	55.32	48.62	46
Spanish	72.33	-	-	33	42.33	37.33	42.91	51.13	49.67
German	-	-	-	21	26	32	41.25	51.22	47.33
Indonesian	69.33	-	60.33	28	36	37.67	60	54.39	-
English	77.33	83.67	-	-	-	-	99.53	-	60.33
French	-	73.49	-	34.33	33.67	32.33	57.34	46.8	54.67

Table 3: Performance of `bloomz7b1` across different languages and tasks.

	XStoryCloze	XWinograd	XCopa	mARC	mHellaswag	mMMLU	FloRes	TruthfulQA	XNLI
Hindi	57.3	-	-	28.3	34.6	30	52.5	46.4	39
Russian	57.6	65.3	-	28.6	36	32.6	48.1	46.3	37.1
Arabic	56.1	-	-	28.3	33.7	31.3	54.1	50.9	33.7
Spanish	59.3	-	-	26.6	37.7	30	46.1	45.2	38.6
German	-	-	-	25.7	36.3	22.7	54.1	44.6	35.6
Indonesian	58.3	-	62.5	28	39	30.7	57.5	43.6	-
English	58	70.3	-	-	-	-	99.7	-	50.3

Table 4: Performance of `mt0XXL` across different languages and tasks.

Comments	IPA Character Sets
-Bilabials	
Aspiration	p, pʰ
	b, bʱ
	β, bʱ
	ɓ, bʱ
Voicing	b, p
	bʱ, pʰ
	ɓ, p
	β, p
Manner	β, b, ɓ
-Labiodentals	
Place (and manner)	b, β, ɓ, v, w
	p, f, w
Aspiration (and place)	t, ʈ, t̪, t̪ʰ, ʈʰ, θ, tˤ
	d, ɖʱ, d̪, d̪ʱ, ð, dˤ, ðˤ
	s, t͡ʃʰ
	dʒ, d͡ʒʱ
Voicing (and place)	tˤ, dˤ
	d, d̪, ð, ɖ, t, t̪, ʈ, θ, ɾ
	s, z, ʒ, ʑ, ʐ
	t͡ɕ, d͡ʑ
	ks, gz
Manner (and place)	s, dʒ, ts, t͡s, ʃ, ʂ, sˤ, ʦ
	ʃ, ʒ, d͡ʒ, t͡ʃ, t͡ɕ, d͡ʑ, ɕ, ʑ, j, ʐ, ɕtɕ, ɟ
	ks, gz, kʂ

Comments	IPA Character Sets
-Approximants	
	ɽ, r, rɪ, ɾ̩, ʁ
	ɭ, l
	l, j
-Velars	
Aspiration	k, kʰ
	g, gʱ
Voicing	g, k
	gʱ, kʰ
	ɣ, x
	gz, ks
Manner	g, ɣ, gʱ
	k, x, kʰ
-Uvulars, pharyngeals, glottals	q, χ, ʁ, ʕ, ħ, ʔ, h, əh
-Nasals	
	m, n, ɲ, ŋ, ɳ
	əŋ, in, un
-Vowels	i, e, ɛ, æ, ɪ, ø, y, œ, eː, ɛː, øː, œː, ʏ, iː, ʊ, aɪ, yː, aʊ, ə, ʉ, ɨ, ɵ, ɵː, əh, ɑ, ɒ, ʌ, ɤ, o, u, ɔ, ɯ, uː, oː, ɔː, ɯː, uː, oː, ɔː, aː, aʊ, a, ɥ

Figure 5: List of IPA character sets for the phonological noiser.

D Results: Further details

D.1 Noising examples

See Table 6 for more examples of noiser output for certain θ's and languages.

D.2 Error type examples

We provide an expanded version of Table 1, with an example for every mentioned error type for `es`.

D.3 Trend stability for individual languages and tasks

In § 4, we discuss the effect of sampling variance in PD for a given θ, that appears to differ by task

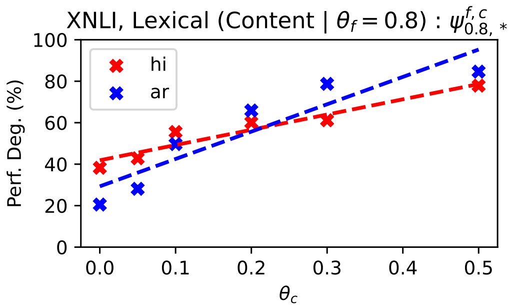

Figure 6: PD for XNLI for `hi` and `ar`, $\phi^{f,c}_{0.8,*}$, averaging over 10 runs for each parametrization; this results in a much stabler trend for PD vs. θ as compared to using a single run as shown in Figure 2.

XNLI	Prompt 1	Suppose that the following is true: premise Can we infer that: hypothesis? Respond with one of the following words: ENTAILMENT_LABEL, CONTRADICTION_LABEL, NEUTRAL_LABEL.
	Prompt 2	Suppose that the following is true: premise Can we infer that: hypothesis? Yes, no, or maybe? Respond in the target language.
	Prompt 3*	premise, QUESTION_WORD? [MASK], hypothesis
XStoryCloze	Prompt 1	What is a possible continuation for the following story ? sentence_1 sentence_2 sentence_3 sentence_4 Choose from the following options: option_1 option_2
	Prompt 2	sentence_1 sentence_2 sentence_3 sentence_4 What is a possible continuation for the story given the following options ? - option_1 - option_2
	Prompt 3	Choose the best continuation of this story: sentence_1 sentence_2 sentence_3 sentence_4
X→eng	Prompt 1	Translate from a dialect of <HRLN> into English
	Prompt 2	Translate from <HRLN> into English
	Prompt 3	Translate into English :

Table 5: Our attempted prompts. *[MASK] is filled with each of the three possible labels, and the model choice is computed using loglikelihood over the entire sequence.

depending on task sensitivity to the specific words that are corrupted as opposed to the general extent of corruption in the input. We choose midrange values of θ_n for $\phi^{f,c}$, ϕ^m, and ϕ^p ($\theta_f = 0.5$, $\theta_c = 0.3$, $\theta_m = 0.5$, and $\theta_p = 0.1$), and generate 10 artificial languages with hi and ar as sources. We report standard deviation in PD for generated languages for each task in Table 9 and Table 10 for hi and ar respectively. We see that std. deviation for X→eng is convincingly lower than for the classification tasks; this is in line with our intuition discussed in § 4. Note that this is std. deviation in percentage PD and not actual scores: e.g., a std. deviation in PD of 10% given a baseline XNLI score of 51 (like for hi) translates to a std. deviation of 1.8 accuracy points.[10] This is low enough for our established trend to be able to provide a good ballpark estimate for the XNLI score for a language for which we have θ.

We also recompute $\psi^{f,c}_{0.8,*}$ for hi and ar for XNLI (4b in Figure 2) using means over 10 runs per θ_c; this combination of language, task, and noiser is motivated by the fact that the associated individual language trends appear most unstable computed over single runs per parametrization. See Figure 6 for the trends; we observe much higher stability in the individual language trend. These findings indicate using means over several generated artificial languages in order to compute reliable trends for a single language, and using associated SD as a confidence measure in the predicted PD.

E PD dynamics on composed noisers

As discussed in § 4, we are interested in how $\psi^{\{x,y,z\}}$ compose to give ψ^{xyz} for two or more noisers, i.e. the nature of the function of PD on individual noisers that gives overall PD on composed noisers.

Effect of growing lexical noise We describe our procedure for composing noisers and its motivation in § 2.1: we apply phonological, morphological,

[10] See § 2 for our calculation of PD.

Noising examples for different languages		
Noiser	Lang	Examples
$\phi^{p}_{0.05}$	id	s: Saat berada di lokasi terpencil dan tanpa jangkauan seluler, telepon satelit mungkin menjadi satu-satunya pilihan Anda. s': Saat berada di lokasi tirpencil dan tanpu jamgkauan seluler, telepon satelit mungkin menjadi satu-satunya pilohan Anda. p: When in remote locations without cell phone coverage, satellite phones may be your only option. p': When you're in the wilderness and without cell phone reception, a satellite phone may be your only option. Ref: In remote locations, without cell phone coverage, a satellite phone may be your only option.
$\phi^{p}_{0.1}$	de	s: Sie haben normalerweise ein besonderes Angebot an Speisen, Getränken und Unterhaltung, um die Gäste bei Laune zu halten und dafür zu sorgen, dass sie bleiben. s': Sie haben nürnalerweise ein bejondehes Ancebot an Speisen, Getränkon und Unterhaltung, um die Gäste bei Laune zu halten und dafür zu sorgen, dacs sie bleiben. p: You usually have a special offer for drinks, food and entertainment, to keep guests at Laune and to make them stay. p': You have a very nice apartment in Speisen, Getränkon and Unterhaltung, to keep the guests at Laune, and to make them stay. Ref: They usually have special food, drink and entertainment offers, to keep guests in a good mood, and keep them at the premise.
$\phi^{m}_{0.6}$	fr	s: Le pays possède une grande variété de communautés végétales en raison de la diversité de ses microclimats, de ses sols et de ses niveaux d'altitude. s': Le pays possèto une grande variédé de communaudéç végétèies en raicon de la diversüté de ses microclimats, de ses sols et de ses niveüu d'altitude. p: The country has a great variety of plant communities due to the diversity of its microclimates, soils, and altitudes. p': The country has a great variety of vegetation due to its microclimates, soils and altitude. Ref: It has a notably wide variety of plant communities, due to its range of microclimates, differing soils and varying levels of altitude.
$\phi^{m}_{0.6}$	es	s: La gran pirámide fue construida en honor al faraón Khufu, y muchas otras de este tipo, tumbas y templos más pequeños se levantaron en honor a sus esposas y familiares. s': La gram pirámide fue construica en honir al faraón Khufu, y muchas otras de este tipo, tumbuc y temples más pequeños se levantarom en honir a sus esposuc y familiaros. p: The great pyramid was built in honor of Pharaoh Khufu, and many other such pyramids, tombs, and temples were built in honor of his wives and family members. p': The pyramid was built to honor the Pharaoh Khufu, and many other such pyramids, tombs, and temples were built to honor his wives and family. Ref: The great pyramid was created to honor the Pharaoh Khufu, and many of the smaller pyramids, tombs, and temples were built to honor Khufu's wives and family members.
$\phi^{f,c}_{0.5,0.3}$	hi	s: हालाँकि हर देश 'स्कैंडिनेवियाई' था, लेकिन डेनमार्क, स्वीडन, नॉर्वे और आइसलैंड के लोगों, राजाओं, रीति-रिवाजों और इतिहास के बीच कई अंतर थे. s': हऔयईंकि अऋ देश 'स्कैंडिनेवियाई' था, लेकिन डेनमार्क, स्वीडन, नॉर्वे औृ आइसलैंड के लोगों, बुक्षे, रीति-रिवाजों औृ इतिहास के बीश कई डरत ठौ. p: Although every country was 'Scandinavian', there were many differences between the people, kings, customs and history of Denmark, Sweden, Norway and Iceland. p': The country 'Scandinavian' was, but the Danes, Swedes, Norwegians and Icelanders, the people, customs and history were very different.
$\phi^{f,c}_{0.5,0.3}$	en	s: Foster care is supposed to provide all the necessities that were lacking in the home they were previously taken from. s': Foster cyal es constaines du provide ayl the necessities did were lacking in the home dee were smenstrainges taken from. p: Foster care is supposed to provide all the necessities that were lacking in the home they were previously taken from. p': Foster care is provided by the government to provide the necessities that were lacking in the home. Ref: Foster care is supposed to provide all the necessities that were lacking in the home they were previously taken from.

Table 6: Examples of noising for different noisers, and model outputs for X→eng on clean and noised source sentences. s: Source, s': Noised source, p: Prediction on source, p': Prediction on noised source, Ref: reference translation.

Examples for all error modes		
Noiser	Strategies	Example I/O
ϕ^{f}_{*}	(a) Infers sentence meaning from content words	s: Al parecer, las cabras fueron domesticadas, por primera vez, hace unos 10 000 años, en los montes Zagros, en Irán. s': Al parecer, luc cabras fiaom domesticadas, por primera vez, hace enes 10 000 años, an los montes Zagros, an Irán. p: Apparently, goats were first domesticated about 10,000 years ago in the Zagros Mountains in Iran. p': It seems that the first domesticated goats were bred in the Zagros Mountains of Iran about 10,000 years ago. Ref: Goats seem to have been first domesticated roughly 10,000 years ago in the Zagros Mountains of Iran.
	(b) Partially correct	s: Los esfuerzos para hallar el lugar del accidente deben lidiar con el mal tiempo y el terreno escarpado. s': Los esfuerzos pea hallar al lugar del accidente cebyn lidiar kom al ah tiempo i al terreno escarpado. p: The efforts to find the crash site must contend with bad weather and rugged terrain. p': The efforts were made to find the place of the accident, but the terrain was too rough. Ref: Efforts to search for the crash site are being met by bad weather and harsh terrain.
	(c) Incorrectly connects content words*	s: Las manifestaciones, en ocasiones violentas, fueron provocadas por el hecho de que no se llevan adelante elecciones, en algunos casos desde el año 2011. s': Luc manifestaciones, an ocasiones violentas, fiaom provocadas por al hecho de guu no ze llevan adelante elecciones, an olgones casos ceztu al año 2011. p: The protests, sometimes violent, were sparked by the fact that elections are not held in some cases since 2011. p': In 2011, there were violent protests, sometimes triggered by the failure to hold elections. Ref: The sometimes-violent protests were triggered by failure to hold elections, some due since 2011.
	(d) Breaks: Function word was part of a construction	s: Sin perjuicio de cuán mansos puedan lucir, lo cierto es que los bisones, los uapatíes, los alces, los osos y prácticamente todos los animales grandes pueden se agresivos. s': Sin perjuicio de ceám mansos piedan lucir, li cierto os guu los bisones, los uapatíes, los alces, los osos i prácticamente dodus los animales grandes pieden ze agresivos. p: No matter how docile they may look, bears, bison, moose, elk, bears, and nearly all large animals can be aggressive. p': Without prejudice to the fact that bison, moose, elk, bears, and nearly all large animals can be aggressive, it is true that the bisons, moose, elk, bears, and nearly all large animals can be very docile. Ref: No matter how docile they may look, bison, elk, moose, bears, and nearly all large animals can attack.
	(e) Off-target	s: Se han rescatado varios rehenes y, hasta ahora, se ha confirmado que al menos seis han muerto. s': Ze han rescatado parius rehenes i, hosta ahora, ze he confirmado guu al menos seis han muerto. p: Several hostages have been rescued, and it is confirmed that at least six have died so far. p': Spanish phrase: Ze han rescatado parius rehenes i, hosta ahora, ze he confirmado guu al menos seis han muerto. Ref: Several hostages have been rescued and least six have been confirmed dead so far.
$\phi^{f,c}_{\theta_f,*}$	(a) Guesses correct word from context	s: Todo en el Universo está hecho de materia, **compuesta** por **partículas pequeñas** denominadas átomos. s': Todo en el Universo está hecho de materia, **tespolaci** por **piamplesc obleyón** denominadas átomos. p: Everything in the Universe is made of matter, composed of tiny particles called atoms. p': Everything in the Universe is made of matter, which is made of tiny particles called atoms. Ref: Everything in the Universe is made of matter. All matter is made of tiny particles called atoms.
	(b) Keeps the original word, code-switched, if sur- rounding context is clear	s: Los rasgos que distinguen a una subcultura pueden ser lingüísticos, estéticos, sexuales, geográficos o estar relacionados con la religión o la política, o una mezcla de factores. s': Los rasgos que distinguen a una calincio pueden ser teleamplinempal, estéticos, sexuales, esolaridalla o estar relacionados con la religión o la política, o una mezcla de factores. p: The characteristics that distinguish a subculture can be linguistic, aesthetic, sexual, geographical, religious, or political, or a combination of factors. p': The characteristics that distinguish a calincio can be teleamplinempal, aesthetic, sexual, esolaridalla, or related to religion or politics, or a mixture of factors. Ref: The qualities that determine a subculture as distinct may be linguistic, aesthetic, religious, political, sexual, geographical, or a combination of factors.
	(c) Keeps the word, garbles sentence	s: El satélite en el espacio recibe la llamada y, luego, la refleja de vuelta casi de forma instantánea. s': El devasalv en el espacio recibe la llamada y, vircap, la refleja de vuelta apases de bharítu instantánea. p: The satellite in space receives the call and then reflects it back almost instantly. p': The devasalv in space receives the call and, vircap, reflects it back to the instantaneous bharítu. Ref: The satellite in space gets the call and then reflects it back down, almost instantly.
	(d) Breaks: wrong guess	s: Los entomólogos emplean el término insecto parásito en un sentido formal para referirse a este grupo de artrópodos. s': Los entomólogos ceradida el catacíónit insecto ingaren en un sintaut formal para referirse a este scomp de artrópodos. p: The entomologists use the term insect parasite in a formal sense to refer to this group of arthropods. p': The entomologists use the term insectivore to refer to this group of arthropods. Ref: The term bug is used by entomologists in a formal sense for this group of insects.
	(e) Ignores the word and translates the rest	s: Hershey y Chase insertaron su propio ADN en una bacteria usando fagos, o virus. s': Hershey y Chase insertaron su propio Adn en una resabajectoma usando capandil, o virus. p: Hershey and Chase inserted their own DNA into a bacterium using phages, or viruses. p': Hershey and Chase inserted their own Adn into a somatic cell using capandil, or virus. Ref: Hershey and Chase used phages, or viruses, to implant their own DNA into a bacterium.

Table 7: Examples of each error mode for es. Continued below.

and lexical noise in this order, independently, and allowing overwrites.

Here, we formalize why the influence of phonological and morphological noise decreases as lexical noise grows. Let's consider $\phi^{m,c}$ for simplicity; other combinations with phonological and functional noise are analogous.

Each source word w independently undergoes the process of morphological and content word noising in this order, to give us the translated CRL word w'. At the end, we have three cases: $w' \in \{w, \phi_m(w), \phi_c w\}$. Note that ϕ_c overwrites ϕ_m, so $\phi_c(\phi_m(w)) = \phi_c(w)$.

Recall that a noiser ϕ^n affects w with probability θ_n. We can then see that

$$P(w' = w) = (1 - \theta_m) \cdot (1 - \theta_c)$$

$$P(w' = \phi_m(w)) = \theta_m \cdot (1 - \theta_c)$$

$$P(w' = \phi_c(w)) = \theta_c$$

As $\theta_c \to 1$, $P(w' = w) \to 0$ and $P(w' = \phi_m(w)) \to 0$. This means the value of θ_m stops affecting the resulting language in the presence of high lexical noise. Note that the reverse is not true.

Examples for all error modes		
Noiser	Strategies	Example I/O
ϕ^p_*	(a) Guesses word meaning from context and spelling clues	s: El informe es sumamente crítico con prácticamente cada aspecto de la política vigente del poder ejecutivo en Irak, y apela a un cambio inmediato de dirección. s': Ey informe es sumamenty crítico con prácticamente cada aspecto de la política vigenty del pider eyetutivo ym Irak, e apela a un camvuo inmediato de dirección. p: The report is highly critical of almost every aspect of the present executive policy in Iraq, and urges an immediate change of direction. p': The report is highly critical of almost every aspect of the present policy of the U.S. towards Iraq, and it calls for an immediate change of direction. Ref: The Report is highly critical of almost every aspect of the present policy of the Executive towards Iraq and it urges an immediate change of direction.
	(b) Makes a wrong guess.	s: La investigación en el campo de la IA supone el desarrollo de máquinas a fin de automatizar tareas que requieren un comportamiento inteligente. s': La investigación ym ul campo de la IA sopone ul desarrolyo de máquinas a fin de audymatizor caeas cue reqeyerem un comportamiento inteligente. p: Research in the field of AI involves the development of machines to automate tasks that require intelligent behavior. p': The research in the field of AI involves the development of machines to automate tasks so that machines can exhibit intelligent behavior. Ref: Research in AI involves making machines to automate tasks that require intelligent behavior.
	(d) Breaks: many changes in proximity.	s: No olvide que, básicamente, usted está visitando un lugar que ofició de fosa común y que también es un sitio de un significado prácticamente invaluable para una parte importante de la población del mundo. s': No ylvide que, básicamente, ustat está visitando un lugar cue ofició de fosa común e cue también es un sitio de un signifijado prácticamente imvaluable para una party importanty de la población del mundo. p: Don't forget that, basically, you're visiting a place that served as a mass grave and that it is also a place of essentially invaluable significance to a significant part of the world's population. p': No ylvide that, basically, ustat is visiting a place that was a fosa común and also a place that has a practically invaluable meaning for a party importanty of the population of the world. Ref: Please remember that you are essentially visiting a mass grave site, as well as a site that has an almost incalculable meaning to a significant portion of the world's population.
	(e) Hallucination	s: Es tradición pasar la noche de Pascua en vela en algún sitio expuesto para contemplar la salida del sol. s': Es tradición fasa la noche de Paszua an vyla an algún sutio uxpaesdo fary comtemfla la caleda del sol. p: It is tradition to spend the night of Easter awake at some exposed place to watch the sunrise. p': It is tradition to make the night of Pascuas by lighting a bonfire in the yard. Ref: There's a tradition to pass the Easter night awake at some exposed point to see the sunrise.
ϕ^m_*	(a) Model faces no issues	s: Montevideo se ubica en los subtrópicos, con frecuentes temperaturas superiores a +30° C durante el verano. s': Montevidyo se ubiga en los subtrópicos, con frecuentec temperaturaz superiorec a +30° C durante el verani. p: Montevideo is located in the subtropics, with frequent temperatures above +30°C during the summer. p': Montevideo is in the subtropics, with frequent temperatures above +30°C during the summer. Ref: Montevideo is in the subtropics; in the summer months, temperatures above +30°C are common.
	(b) Breaks: too much corruption*	s: Il est de tradition de passer la nuit de Pâques éveillé à un endroit à découvert pour voir le lever du soleil. s': Il est de traditiin de pasjer la nuèt de Pâques éveillé à un endroèt à découvert pour vâyr le levir du soleel. p: It is traditional to stay up all night on Easter Sunday to see the sunrise. p': Traditionally, it is custom to wake up at dawn on Easter Sunday to see the sunrise at a place of worship. Ref: There's a tradition to pass the Easter night awake at some exposed point to see the sunrise.

Table 8: Continued from Table 7: Examples of each error mode for `es`.

	$\phi^{f,c}_{0.5,0.3}$	$\phi^m_{0.5}$	$\phi^p_{0.1}$	**Task Avg.**
X->eng	4.4	2.6	4.6	3.9
XNLI	18.1	9.7	17.0	14.9
XStoryCloze	16.5	10.7	11.2	12.8
Noiser Avg.	13.0	7.7	10.9	-

Table 9: Std. dev. of PD% over 10 artificial languages generated by a given noiser for each task, for `hi`

	$\phi^{f,c}_{0.5,0.3}$	$\phi^m_{0.5}$	$\phi^p_{0.1}$	**Task Avg.**
X->eng	2.8	2.0	6.9	3.9
XNLI	9.3	10.9	6.5	8.9
XStoryCloze	14.3	14.6	20.3	16.4
Noiser Avg.	8.8	9.2	11.2	-

Table 10: Std. dev. of PD% over 10 artificial languages generated by a given noiser for each task, for `ar`

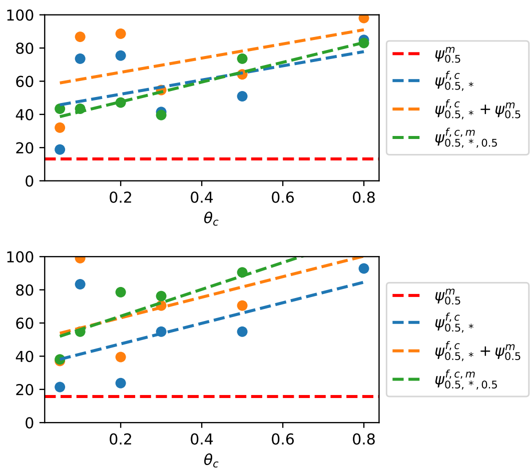

Figure 7: Composing $\phi^{f,c}$ and ϕ^m: studying $\psi^{f,c,m}$ given θ_m for Hindi for `XNLI` (top) and `XSC` (bottom). $\psi^{f,c} + \psi^m$ shows the theoretical additive trend.

This effect is of course a consequence of our composition procedure allowing complete overwrites; however, this matches our intuition about linguistic variation well. Note that it is possible to nuance our noising procedure by allowing a stem overwrite while maintaining a noised suffix; we do not experiment with this idea.

Composite noiser curves See Figure 7 for $\psi^{f,c,m}_{0.5,*,0.5}$ for `XNLI`. We see a similar trend for `XNLI` as we saw in Figure 4 for X→eng, i.e. overall PD simply tracks the maximum individual PD (lexical in this case). However, we see that for XSC, overall PD is closer to the theoretical additive trend and exceeds it for higher θ_c. This difference may be indicate a dependence of the composition function of noisers on task, or the task-specific variance in PD given some θ^n (as discussed in § D.3). We leave a more detailed investigation of noiser composition to future work.

F Posteriors: More details

F.1 Posterior computation details

Source	CRL	θ^c	θ^f	θ^m	θ^p	BLEU	PD (%)
hin	hin	0	0	0	0	56.44	0
	awa	0.15	0.67	0.26	0.05	37.03	34.39
	bho	0.24	0.79	0.32	0.07	32.38	42.63
	hne	0.18	0.67	0.24	0.05	33.24	41.11
	mag	0.14	0.7	0.26	0.05	41.47	26.52
	mai	0.2	0.81	0.34	0.04	28.4	49.68
ind	ind	0	0	0	0	60	0
	zsm	0.19	0.46	0.13	0.06	53.01	11.65
spa	spa	0	0	0	0	42.91	0
	glg	0.22	0.71	0.2	0.11	47.01	-9.55
fra	fra	0	0	0	0	57.34	0
	oci	0.57	0.88	0.73	0.09	38.4	33.03
deu	deu	0	0	0	0	41.25	0
	dan	0.5	0.98	0.71	0.1	16.37	60.32
	isl	0.75	0.99	0.68	0.15	4.11	90.04
	swe	0.56	0.99	0.7	0.1	16.7	59.52
arb	arb	0	0	0	0	55.32	0
	acm	0.09	0.32	0.08	0.03	24.17	56.31
	acq	0.06	0.25	0.04	0.04	46.76	15.47
	aeb	0.2	0.43	0.11	0.05	43.55	21.28
	ajp	0.21	0.55	0.15	0.04	38.25	30.86
	apc	0.21	0.64	0.18	0.04	44.41	19.72
	ars	0.02	0.02	0.01	0.05	48.36	12.58
	ary	0.32	0.6	0.12	0.03	50.16	9.33
	arz	0.19	0.5	0.1	0.04	33.05	40.26

Table 11: Posteriors for related languages, BLEU scores for X→eng, and corresponding PD.

See Table 11 for X→eng BLEU scores on real languages, associated PD, and posteriors for all noisers computed as described in § 2.2. We check that using automatically aligned lexicons, which have naturally poorer quality, does not impact the posteriors too much: we verify 300 accurate entries for the mai-hin and hne-hin silver lexicons, and obtain posteriors within ± 0.05 of the posteriors computed on silver lexicons for all θ_n except for θ_c for hne, which is -0.1. θ_c is most vulnerable to being mis-estimated due to noisy alignments since it only checks for high NED. This is unlike θ_m, which is computed on word pairs with the same stem, and θ_p, which takes into account common phonological context on the source and target. Further, statistical word aligners are more likely to work with on very common function words, and give a roughly accurate estimate of θ_f. We recommend paying attention to the quality of the lexicon for posterior computation of θ_c.

F.2 Examples of pseudo-CRLs

Using the posteriors shown in Table 11 for a CRL relative to its HRLN, we can now generate pseudo-CRLs by composing these noise types using the procedure described in § 2.2 (i.e. we applying ϕ^p, ϕ^m, $\phi^{f,c}$ in this order, independently of each other, to the HRLN). We provide examples of pseudo-CRLs generated in this manner in Table 12, to illustrate noise composition in this manner.

Pseudo-CRLs generated from posterior parameters		
Source	CRL	Examples of I/O with generated pseudo-CRL
hin	mai	s: ब्रह्मांड की सभी वस्तुएँ पदार्थ से बनी हैं सारे पदार्थ सूक्ष्तम कणों से बनें हैं, जिन्हें अणु कहा जाता है s': रह्मांड खी शबु वस्तुएँ पदार्थ शे बनी अःैं सारि पदार्थ सूक्ष्तम कणों शे बनें अःैं, जिन्हें अणु कहा जाता है p: All things in the Universe are made of matter. All matter is made of tiny particles called atoms. p': The universe is made of matter, which is made of tiny particles called atoms. Ref: Everything in the Universe is made of matter. All matter is made of tiny particles called atoms.
hin	hne	s: हमारे ग्रह की नदियों से महासागरों में जाने वाले पानी का 20% हिस्सा अमेज़न से आता है s': हमारे ग्रह की स्टिलेजी शे महासागरों नें झाने वाले पानी का 20% हिस्सा अमेज़न शे आटई पै। p: 20% of the water that pours out of the planet's rivers into the oceans comes from the Amazon. p': Our planet's steel is in the ocean's 20% of the world's water. Ref: A full 20 percent of the water that pours out of the planet's rivers into the oceans comes from the Amazon.
spa	glg	s: La investigación todavía se ubica en su etapa inicial, conforme indicara el Dr. Ehud Ur, docente en la carrera de medicina de la Universidad de Dalhousie, en Halifax, Nueva Escocia, y director del departamento clínico y científico de la Asociación Canadiense de Diabetes. s': La invesdigación todyvío so uboca on ci etapa schiga, conworme indicara el Dr. Ehud Ur, doconti on ya carruu te medicymy te ya Universidad te Dalhousie, on Halifax, Nueva Escocia, e dietcor pori cepartamunto clínico e ciontfico te ya Asociación Canadiense te Diabetes. p: The research is still in its early stages, as Dr. Ehud Ur, a professor in the Department of Medicine at Dalhousie University in Halifax, Nova Scotia, and the clinical and scientific director of the Canadian Diabetes Association, indicated. p': The research is still in an early stage, as indicated by Dr. Ehud Ur, a doctor in the Department of Medicine at Dalhousie University in Halifax, Nova Scotia, and director of the clinical and scientific department of the Canadian Diabetes Association. Ref: Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chair of the clinical and scientific division of the Canadian Diabetes Association cautioned that the research is still in its early days.
spa	glg	s: Durante los años 60, Brzezinski trabajó para John F. Kennedy en el puesto de asesor y, posteriormente, para el gobierno de Lyndon B. Johnson. s': Durante yus hauspe 60, Brzezinski trabujó vara John F. Kennedy on el puesto te aseser e, posteriormente, vara el gicklasigervanu te Lyndon B. Johnson. p: During the 1960s, Brzezinski worked for John F. Kennedy as a counselor and then for the Lyndon B. Johnson administration. p': During the 1960s, Brzezinski worked for John F. Kennedy in the position of advisor and, subsequently, for the administration of Lyndon B. Johnson. Ref: Throughout 1960s, Brzezinski worked for John F. Kennedy as his advisor and then the Lyndon B. Johnson administration.
deu	dan	s: Wie einige andere Experten zeigte er sich skeptisch, ob es möglich sei, Diabetes zu heilen, und wies darauf hin, dass die Befunde für Menschen, die bereits unter Typ-1-Diabetes litten, keine Bedeutung hätten. s': Wie eemöca imtera Experten daufenöttis ir cish skeptisgr, ub uj toteno zei, Diabetes ßu mende, and wiös daryuv rön, tasc tiü Befunde för Menschen, tiü bereits amder Typ-1-Diabetes littum, qeeme Bedeutung rättym. p: Like some other experts, he was skeptical about whether it was possible to cure diabetes, pointing out that the findings had no significance for people who were already suffering from Type 1 diabetes. p': How some among experts are clearly skeptical, whether it means to say diabetes, and what from it, that the findings for people who were already suffering from Type 1 diabetes, would have no significance. Ref: Like some other experts, he expressed skepticism about whether it was possible to cure diabetes, noting that the findings had no relevance to people who already had Type 1 diabetes.
deu	swe	s: Während ein experimenteller Impfstoff in der Lage zu sein scheint, die Ebola-Mortalität zu senken, gibt es bisher keine Medikamente, die als eindeutig zur Behandlung bestehender Infektionen geeignet nachgewiesen wurden s': Während een erschenienkeysto Impfstoff on ter Lage ßu seen vornetivi, tiü Ebola-Mortalität ßu sengöm, auelti uj antallke qeeme Medikamente, tiü ajß eindeutig plan Behandlung bestehentir Infektionen sápmostort nakhgewiösäm böhdem p: While an experimental vaccine appears to be able to reduce Ebola mortality, so far there are no drugs that have been definitively proven to be suitable for the treatment of existing infections. p': While one appeared to be on the verge of a breakthrough in vaccine development, the Ebola mortality rate seemed to decline, yet there were still few medications that clearly outlined effective treatment for existing infections, leaving much to be desired. Ref: While an experimental vaccine appears to be able to reduce Ebola mortality, there are no drugs that have been clearly proven to treat existing infections.

Table 12: Examples of pseudo-CRL generated by setting noise parameters for each noiser equal to the computed posteriors for each source-CRL pair given noise type as shown in Table 11. s: Source, s': Noised source, p: Prediction on source, p': Prediction on noised source, Ref: reference translation.