We Hebben Een Serieus Translatie: Modeling Intercomprehension as Probabilistic Inference

Thomas Hikaru Clark, Edward Gibson, & Roger Levy
Department of Brain and Cognitive Sciences
MIT
Cambridge, MA 02139, USA
{thclark,egibson,rplevy}@mit.edu

Abstract

Intercomprehension refers to partial intelligibility of an unfamiliar language (L2) by a speaker of a related language (L1). How is this zero-shot cross-language comprehension possible? In this work, we extend past work on algorithmic models of noisy-channel inference to model intercomprehension in a Bayesian framework. The model uses an LM in L1 only for scoring latent hypotheses about the translations of observed L2 utterances, and a general-purpose noise model to infer a mapping between L2 and L1 words based on either form-based similarity or symbolic rules. We then conduct a human behavioral experiment, eliciting inferences for utterances in Dutch, Italian, and Ukrainian from speakers of English, Spanish, and Russian, respectively. Our full model shows a closer alignment to the distribution of human intercomprehension performance than ablations, and also compares favorably to zero-shot prompting of much larger models. These results provide a cognitively plausible computational model of intercomprehension, and highlight the flexible inferences made by comprehenders under wide uncertainty in real-world cross-language scenarios. We share our code publicly.[1]

1 Introduction

Natural languages do not exist in a vacuum. As a result of the process of language evolution, languages exist in relation to each other in a hierarchy of linguistic proximity. Related languages often have some degree of **mutual intelligibility** or **intercomprehension** with each other, meaning that speakers of a language L1 can understand a non-trivial amount of a related language L2, without any formal training or exposure to L2 (Gooskens et al., 2018; Bonvino et al., 2018; Carlo & Garbarino, 2021; Golubović & Gooskens, 2015; Beijering et al., 2008). Given that this process can be asymmetric, rather than mutual, we will use the term **intercomprehension** as a general term for partial understanding of an unfamiliar or unstudied but related language.

From a cognitive science perspective, this ability is a striking phenomenon, reflecting an ability to process and make rich inferences about out-of-distribution observations with large amount of uncertainty (Tenenbaum & Griffiths, 2001; Rule et al., 2020; Ellis et al., 2015; Tenenbaum et al., 2011; Xu & Tenenbaum, 2007; Zhi-Xuan et al., 2020). From a practical perspective, understanding the process by which the human mind performs this type of inference has ramifications for language pedagogy and inter-cultural communication and collaboration (Doye, 2004). In natural language processing, the problem of translating unknown languages is known as *decipherment*, and has applications for archaeological and historical research on untranslated documents in extinct languages (Ravi & Knight, 2011; Luo et al., 2019; Naim et al., 2018; Luo et al., 2021).

[1]Github: https://github.com/thomashikaru/intercomprehension-colm-2026

To provide a motivating example, the Dutch phrase *We hebben een serieus probleem* went viral for being intelligible to English speakers with no prior Dutch exposure [2]. This simple sentence has identical word order to its corresponding English translation, and contains cognates and/or loanwords which have surface-level similarities to their English counterparts. While it may seem obvious that an English speaker would be able to infer the meaning of this sentence, this example poses some challenges for a satisfying cognitive account, especially at an algorithmic level of explanation (Marr, 1982). For example, the Dutch word *een* in fact bears little surface-level similarity to its English counterpart *a*, yet comprehenders appear able to infer its meaning "from context". It remains underspecified how exactly a comprehender performs this inference using context. Similarly, the word *serieus* is also similar to other English words, such as *series*, not just the correct translation of *serious*. Thus, it appears that comprehenders are able to rapidly integrate multiple sources of information to produce an interpretation that aligns with probabilistic inference.

In this paper, we extend the **noisy-channel language processing** framework (Gibson et al., 2013; Levy, 2008; Chen et al., 2023) from psycholinguistics and cognitive science to the problem of intercomprehension. A basic premise of this framework is that humans make use of both a linguistic **prior** (expectations over which messages are likely to be sent) as well as a **likelihood** rooted in an implicit or explicit error model which captures the relative probabilities of different kinds of transformations that could "corrupt" or alter the intended message. Recent work in this paradigm has offered candidate algorithmic models for simulating the fine-grained, incremental behavior of noisy-channel comprehenders under varying assumptions (Clark et al., 2025a;b), yet to our knowledge this framework has not been applied to the more general types of "noise" that comprehenders deal with in the context of intercomprehension.

Past work has modeled mutual intelligibility computationally by using Levenshtein distance (Gooskens & Heeringa, 2004; Beijering et al., 2008). Yet edit distance alone only captures the component of intelligibility that relies upon surface form similarity, and leaves out an important variable: prior expectations about intended messages. It remains unclear if human comprehension of sentences in an unfamiliar language will reflect naive, surface-form-based inferences about its meaning, or integrated rational inference. Furthermore, an open question is which specific algorithmic components of inference most closely predict the pattern of successes and failures of humans at intercomprehension. Nieder & List (2024) propose a model based on linear discriminative learning, but it operates on a dataset of cognates (rather than generating translations of arbitrary sentences). Meanwhile, Luo et al. (2019) propose a decipherment model for ancient lost languages, but it is based on handcrafted patterns of historical sound change (rather than being a cognitive model of a naive zero-shot comprehender), and again is evaluated based on aligning cognates (percentage of cognates identified) rather than on translating full sentences. In this work, our unique contribution is an implemented cognitive model of intercomprehension that operates on arbitrary natural sentences, that integrates both form-based similarity and linguistic priors to form inferences.

2 Model: Tempered SMC for Zero-Shot Decipherment

2.1 Generative Model

We set up a generative model, which expresses the process by which sentences from one language L1 may be transformed into sentences from another language L2. We assume that a sentence z in L1 is sampled from a language model, which is simply a probability distribution over strings $p(z)$. In practice, we use simple, autoregressive language models from Goldfish (Chang et al., 2026), which provide monolingual models across 350 languages that are matches for parameter count and training data size. Since the purpose of the LM is only to capture the probability distribution over strings, not to perform reasoning or inference, we use a language model without Chain-of-Thought (Wei et al., 2023), instruction-tuning, or extensive multilingual training data.

[2] https://knowyourmeme.com/memes/we-hebben-een-serieus-probleem

Given the L1 sentence z, the generative model applies transformations at the word level to yield a sentence y in L2. The generative model reflects an L1 comprehender's intuitive theory of language transformations to an unknown but related language, thus the error model should be extremely general rather than reflecting any specific knowledge of the transformations that occur between a given pair of related languages. For each word in z, the generative model randomly samples whether to apply a **string edit** transformation or a **memorized substitution** transformation.

For a given word z_t, the distribution over string edit transformations is defined implicitly by the relationship $\log p(y_t \mid z_t) \propto -\text{editDistance}(y_t, z_t)$. Thus, larger string edits are less probable than smaller string edits, with each additional edit decreasing the probability exponentially. This reflects a naive intuitive theory of string transformations where edits are independent, random processes. We consider two distinct types of edit distance, one which instantiates an **orthographic likelihood** via string-based edit distance, and one which instantiates a **phonetic likelihood** by first converting L2 words to an ARPABET representation using the `g2p_en` word pronunciation module in Python[3], which predicts pronunciations for out-of-vocabulary words using a neural network, and computing edit distances in this space to ARPABET representations of items in an L1 wordlist. The final editDistance() function uses an equally weighted average of orthographic and phonetic edit distance.

The memorized substitution transformation is a component of the generative model which reflects the idea that not all words have a cognate, even in highly related languages. Therefore, certain mappings between words in L1 and L2 simply need to be memorized. The generative model contains a **rule library** $\mathcal{L}$ which stores rewrite rules of the form $a \rightarrow b$. The probability distribution over rule libraries is defined implicitly by the relationship $\log p(\mathcal{L}) \propto \text{charCount}(\mathcal{L})$, enforcing a **minimum description length** prior: rule libraries that are simpler and more concise are more likely *a priori* (Rissanen, 1978; Grünwald, 2007). This prior additionally has the consequence that even if the same rule is used multiple times, it is only counted once in $\mathcal{L}$, implicitly rewarding rules which can explain multiple observations. The full generative model specification can be found in Appendix A.

2.2 SMC Inference

Sequential Monte Carlo (SMC) is an inference algorithm for sampling from distributions that are challenging to sample from exactly, but where decomposing the problem into a sequence of sub-problems is beneficial for inference (Naesseth et al., 2024; Doucet et al., 2001). In simple terms, SMC involves maintaining a set of weighted **particles** which each correspond to a hypothesis about the value of latent random variables, and updating the particle weights sequentially while iterating through subproblems. In many cases, SMC is applied to inference problems that are sequential in time, e.g. inferences about random variables that are conditioned on a temporal sequence of observations. This is the approach employed by Clark et al. (2025a) to model psycholinguistic features of noisy-channel inference: possibly noisy words in a sentence are observed sequentially over time, while SMC updates a set of particles which correspond to a distribution over latent intended messages.

For modeling intercomprehension, instead of modeling sequential observations, we model a sequence of inference targets where linguistic priors take on an increasingly prominent role. Given an observed L2 sentence y, we approximate the distribution over L1 translations of that sentence z. Thus, the target posterior is: $\pi_{\text{post}}(z \mid y) \propto p(y \mid z)p(z)$. We set up a sequence of target distributions by introducing a parameter β which acts as a tempering exponent on the prior, gradually increasing the prior's contribution (Neal, 1998; Del Moral et al., 2006). This reflects the intuition that inference proceeds from solving an easier, more naive problem, i.e. identifying surface-level similarities between L1 and L2, to a more challenging and higher-level problem, i.e. combining likelihood and linguistic prior to identify strong candidate translations: $\pi_\beta \propto p(y \mid z)p(z)^\beta, \quad \beta \in [0, 1]$. π_β now represents a family of targets. For $\beta = 0$, the language model prior does not contribute, and thus we initialize particles solely based on string edit distance, i.e. drawing from $\pi_0(z \mid y)$ by

[3] https://pypi.org/project/g2p-en/, Version 2.1.0

computing edit distances between observations $y_0 \dots y_T$ and words in an L1 wordlist. As β is increased, the target distribution approaches the true posterior; at each step, particle weights are updated. Resampling occurs if the effective sample size (ESS) drops below a threshold. See Appendix B for inference details.

2.3 MCMC Rejuvenation Moves

Rejuvenation is performed after resampling to reintroduce diversity into the particles, without changing particle weights (Gilks & Berzuini, 2001; Del Moral et al., 2006). Several specific MCMC kernels are used to apply rejuvenation. The **word substitution proposal** takes an observed word y_t and simply samples a different word as the latent translation z_t from the distribution over words in a L1 wordlist, with log probability inversely proportional to the edit distance with y_t. The **common-word rule library proposal** takes an observed word y_t and proposes a new library rule that it maps to an L1 function word z_t. The **contextual rule library proposal** takes an observed word y_t in a latent context $z_{<t}$ and proposes a new library rule that maps it to an L1 word z_t which is sampled from the L1 language model $p(\cdot \mid z_{<t})$. During each inference step, all three proposals are applied to each word in the observed sentence y with the following exception: since the **contextual rule library proposal** critically depends on having a coherent sentential context in order to predict a missing word, it is only applied after a burn-in period of 2 iterations.

2.4 Example

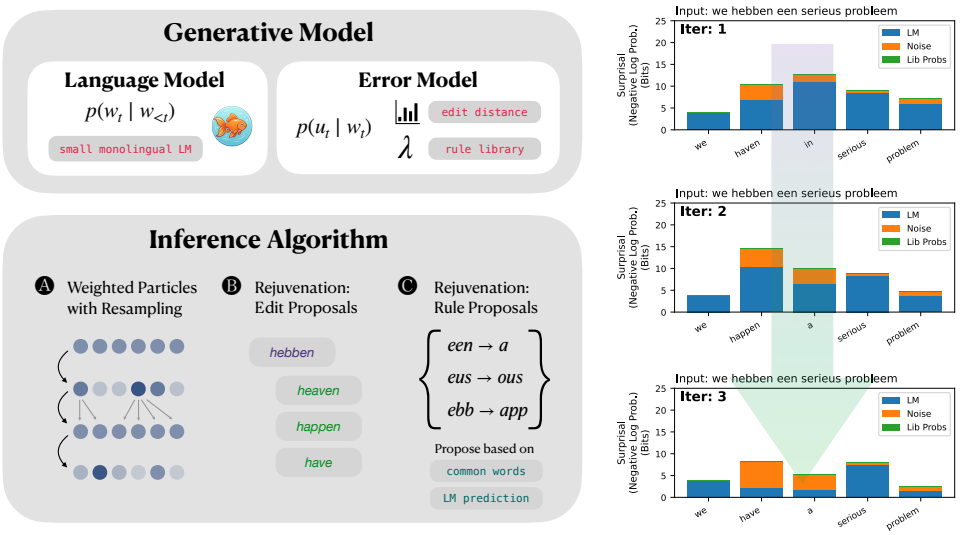

Figure 1: Model overview with example. **Generative model:** decomposed into a language model (prior) and noise model (likelihood). **Inference algorithm:** key components include SMC inference using weighted particles with resampling, and rejuvenation proposals that propose different word edits or adding/removing library rules. **Example:** Per-observation surprisal according to one particle over 3 successive iterations. Later stages of inference correspond to higher β, i.e. greater contribution of LM prior.

We consider the progression of the inference process for an example Dutch sentence: *We hebben een serieus probleem*. Figure 1 visualizes this using a barplot showing the negative log probability, or **surprisal**, assigned by the model to each word in the observed Dutch sentence, decomposed into LM surprisal (derived from language model log-probabilities), library rule surprisal (based on the description length of the rule used in a rule application, if present), and "noise" surprisal (directly related to edit distance). More formally, for a sequence of Dutch words $\mathbf{y} = y_1, y_2, \dots y_T$ and a sequence of English words $\mathbf{z} = z_1, z_2, \dots z_T$,

we plot $p(z_t \mid z_1, z_2, \ldots z_{t-1})$ and $p(z_t \mid y_t)$. When $\beta = 0$, we see that the latent English sentence (the inferred translation) is dominated by the noise likelihood. In other words, it prioritizes keeping string edit distance small with no regard to the plausibility of z as an English sentence, leading to low-probability sequences of English words which have high surface-level similarity to the Dutch sentence. As β increases, the inference process gradually increases the contribution of the English LM. In the figure, this is reflected in particles with reduced LM surprisal, possibly at the cost of increased transformation surprisal. The inferred translation of the Dutch sentence more closely resembles a plausible English sentence, rather than a word-by-word mapping based only on surface-level similarity. In particular, we observe that English words which are quite dissimilar from their Dutch counterparts are inferred, which can be explained by the use of rejuvenation moves which propose word transformations by applying the LM's next-word prediction ability the prefix of an inferred latent English sentence. If the current particle's inferred latent sentence is of low quality, such moves are unlikely to be accepted under Metropolis-Hastings. Yet once the current particle's sentence is close enough to a reasonable English sentence to start making predictions, the quality of these proposals also increases. This results in a phase-shift pattern where inference may in some cases undergo rapid movement to a high-probability region of the problem space, while in other cases it remains stuck in low-probability regions.

3 Evaluation

3.1 Data

Given that intercomprehension for humans is most likely to succeed when dealing with simple sentences with a high density of cognates, we do not expect our model to translate arbitrary sentences well from an unfamiliar language. However, we do expect to see meaningful variation in the success of the model's zero-shot approach when given sufficiently simple and cognate-rich sentences. To this end, we create a basic evaluation dataset consisting of three language pairs: Dutch-English, Italian-Spanish, and Ukrainian-Russian. For each language pair, parallel sentences are extracted from the OpenSubtitles corpus and filtered for length and quality. More information on dataset construction is found in Appendix C.

3.2 Evaluation Metrics

To compare the similarity of two L1 translations of a given L2 sentence (e.g. human-model similarity or model-ground truth similarity), we employ three distinct metrics: (1) Character-level edit distance is the number of insertions, deletions, or substitutions, needed to transform one string into another. (2) Word-level edit distance is analogous, but operates on word edits rather than character edits. (3) Sentence embedding cosine distance using SBERT (Reimers & Gurevych, 2019) for L1=English or Multilingual E5 (Wang et al., 2024) for the non-English L1 settings. For a particular item, we compare the similarity of the *distribution* of human responses to the *distribution* of model responses using the Wasserstein or Earth-Mover's Distance, as described in Appendix D. Intuitively, this captures the divergence between two discrete distributions, using one of the above three distance functions.

3.3 Ablations and Baselines

We perform inference using the following ablations and baselines:

Likelihood-only baseline without LM prior tempering. This ablation does not perform iterative inference with gradual tempering in of the LM prior. L1 candidates are chosen purely based on likelihood – i.e., by choosing the L1 word which has the smallest edit distance from the observed L2 word. This minimizes the edit distance between observation and inferred translation, but does not enforce a pressure for plausible L1 messages.

Remove rule library proposals from inference. This ablation implicitly enforces a constraint that every word in L2 must be a cognate of some word in L1, and scores mappings based

only on their edit distance. This fails to account for the ability to infer and memorize a reusable rule that is based on context and not on any surface-level similarity.

Reduce the number of particles used in SMC inference. This manipulation reflects the inverse relationship between computational resources and exactness of inference in Sequential Monte Carlo. We use 32 particles for the full model and 1 particle for the ablated model. Thus, the ablated model reflects greedy inference which only ever maintains one hypothesis at a time, while SMC with 32 particles reflects parallel processing via a weighted collection of hypotheses.

Zero-Shot Prompted LLMs. As an alternative to our Bayesian inference approach, we attempt zero-shot translation using the following prompted LLMs: `Gemma-3-1B`, `Llama-3.2-1B`, and `SmolLM-2-135M` (for English-Dutch). The prompt used can be found in Appendix G. The `SmolLM-2-135M` model is parameter-matched to the Goldfish models used by our approach; the Gemma and Llama models are much larger at 1 billion parameters. None of these baseline models are strictly trained on monolingual data, though `SmolLM` only officially supports English.

4 Human Experiment

We perform a behavioral experiment to establish intercomprehension norms for the 100 sentences in each of Dutch, Italian, and Ukrainian; sentences in these languages were presented to native English, Spanish, and Russian speakers, respectively. For each language, we recruit 200 participants on Prolific.

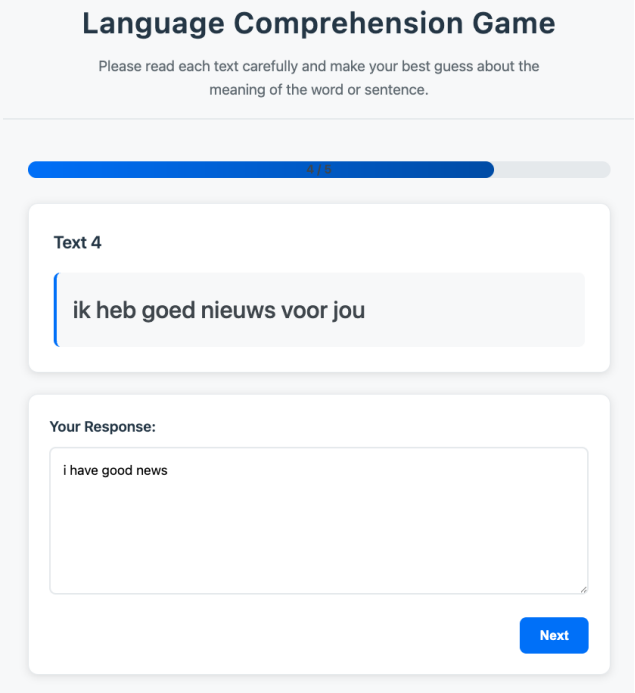

Figure 2: Experimental paradigm. Participants see 5 trials and submit an inference for each one.

Dutch sentences were presented to self-reported English-speaking monolinguals. Italian sentences were presented to Spanish speakers; those who reported prior knowledge of Italian were filtered out (30 exclusions). Ukrainian sentences were presented to Russian speakers; those who reported prior knowledge of Ukrainian were filtered out (58 exclusions). The sentences are divided into 20 disjoint lists, each consisting of 5 unique items. Participants are assigned to experimental lists pseudo-randomly while ensuring balanced numbers of

participants across lists. All participants provide informed consent at the beginning of the study, and the study was conducted according to an approved IRB protocol at the authors' institution. Participants each saw 5 experimental items. In total, the expected duration of the study was 5 minutes, and participants were paid $1.25. Participants were instructed: "Please read each text carefully and make your best guess about the meaning of the sentence" (or the translation-equivalent in their primary language). Participants were explicitly told not to use any artificial intelligence (AI) tools, translation apps, or other outside help during the task. Additional steps were taken to prevent use of outside tools: we blocked the ability to select and copy or paste text.

Figure 2 shows a screenshot of the experimental interface for the human behavioral experiment.

5 Results

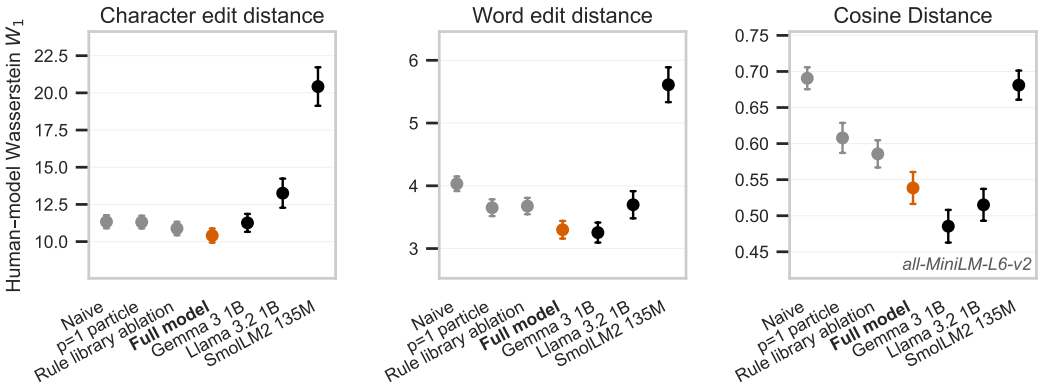

Figure 3: Dutch (L2) to English (L1).

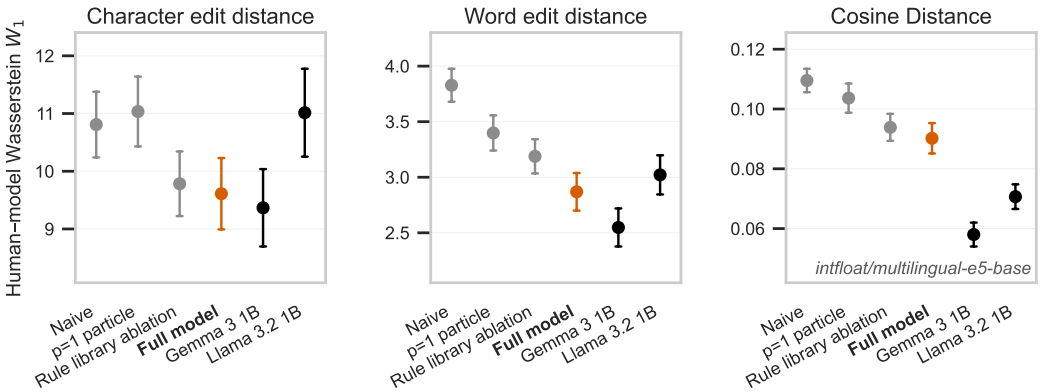

Figure 4: Italian (L2) to Spanish (L1).

5.1 Model demonstrates graded success at intercomprehension task

A comparison of the distribution of evaluation scores in humans and models, across evaluation metrics and model settings, is shown in Figures 3 to 5. We note that the progression of distance metrics goes from most low-level (character edit distance) to most high-level (SBERT embedding cosine distance). Several general patterns emerge. The particle ablation illustrate the contribution of parallel processing (multiple particles > single particle) Likewise, the rule library ablation demonstrates that the ability to infer symbolic rewrite

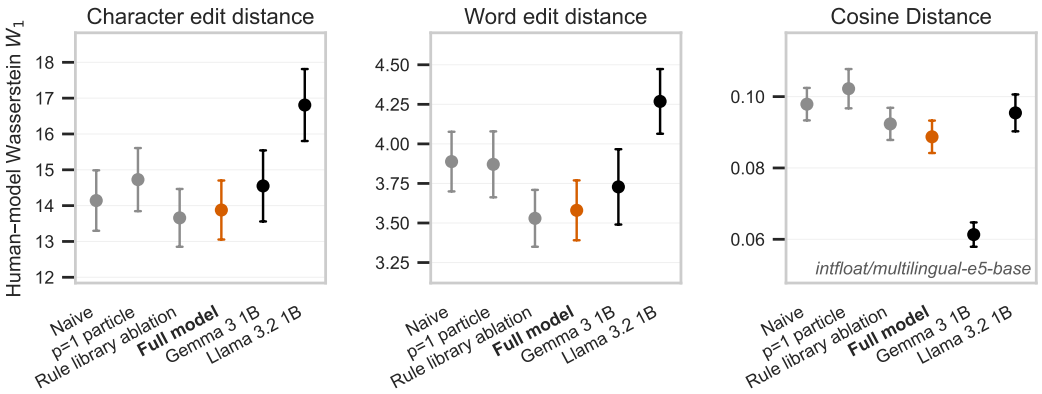

Figure 5: Ukrainian (L2) to Russian (L1).

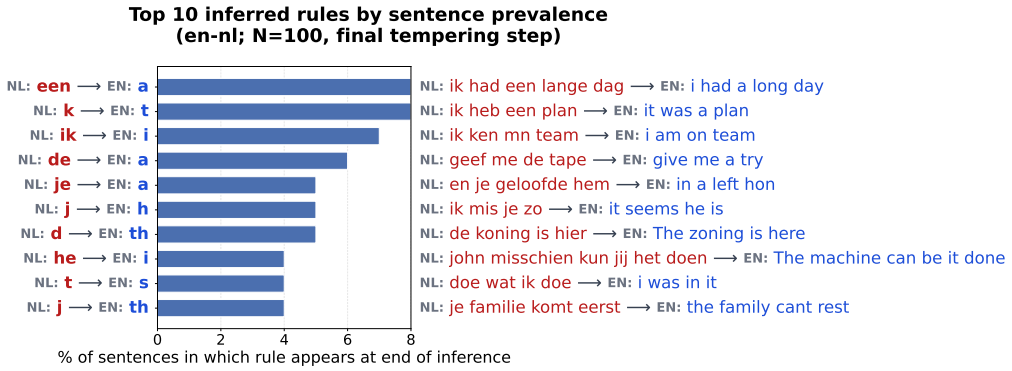

Figure 6: Top inferred subword rules for Dutch → English. The patterns show both correct inferences (*een* → *a*), as well as incorrect inferences (*k* → *t*).

rules leads to greater alignment of model inferences with human inferences. Comparing the Naive model to the Full model, we observe that the Full model (which reflects a gradually tempered L1 LM prior) consistently outperforms the Naive model (which uses only form-based similarity) at matching human inferences.

5.2 Inferred rules capture both cognate structure and red herrings

A direct byproduct of SMC inference with the provided generative model is a distribution over inferred rules. We visualize the most commonly inferred rules in Figure 6. For each

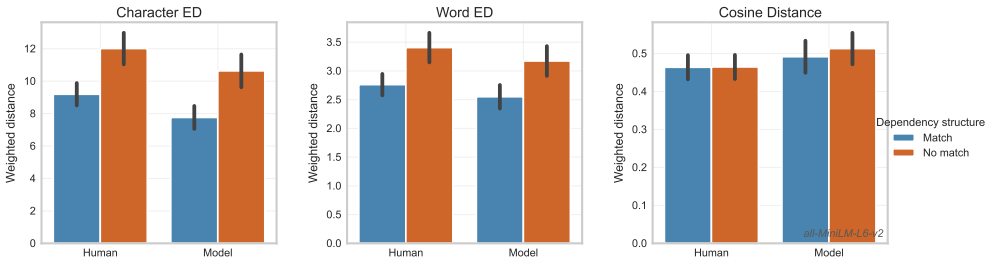

Figure 7: Comparison of distance between human/model inference and ground truth translation, as a function of dependency structure match. Both humans and models show a closer ground-truth match when dependency structures are matched.

Dutch→English re-write rule, one randomly chosen exemplar item is shown (a Dutch sentence in the evaluation corpus, and the inferred English sentence formed using that rule). Full-word rewrite rules like *een→a* and *ik→i* are common. The subword rule *d→th* reflects the correct cognate correspondence between English *th* and Dutch *d*. As would be expected, however, there are also incorrect inferred rules, or "red herrings". Because the model is operating in a very low-data setting (single sentence observations), and has no prior knowledge of cognate relationships between languages, rules may be inferred that results in plausible-seeming inferences, even if they are not veridical. Additionally, we note that for both the full model and human participants, exact character- or word-level match to the ground-truth translation is higher when the items in the language pair share a dependency structure, as opposed to having distinct structures Figure 7.

Top inferred rules for the Italian-Spanish and Ukrainian-Russian language pairs are shown in Appendix F.

5.3 Comparison to zero-shot prompted LLMs

In many cases, our approach achieves comparable or even closer alignment to human inferences, compared with much larger prompted LLMs, despite these larger models not being trained on strictly monolingual data. In the English-Dutch experiments, we performed zero-shot prompting with the parameter-matched `SmolLM` model, which aligns with human inferences substantially worse on every metric. Our approach is successful specifically because it decomposes intercomprehension into language modeling and symbolic inference, and only uses the small Goldfish model for what they are trained to do: capture the prior over messages in L1. Our model tends to be outperformed by Gemma and Llama when judged by the most high-level, semantic metric (SBERT). This suggests that our model's word-for-word translation approach fails to capture some aspects of human intercomprehension, which may be better modeled using gist-based approaches.

6 Discussion

6.1 Cognitively plausible model of intercomprehension

The model described in this paper is intended as an algorithmic-level model of a cognitive phenomenon, not as a state-of-the-art model for machine translation. Clearly, higher performance on translation can be achieved by training models on large multilingual (and possibly parallel) datasets. Yet the problem of zero-shot decipherment of an unfamiliar language, using only knowledge of a related language and general-purpose inference strategies, reflects a cognitive task with ecological validity and which requires flexible, real-time probabilistic inference. Our key innovation is to extend psycholinguistic models of noisy-channel inference (Levy, 2008; Gibson et al., 2013; Clark et al., 2025a), typically applied to out-of-distribution utterances *within* a language, to a much more general case of inference across potentially *different* languages. Key to achieving this is an efficient approximate inference algorithm that incorporates a cognitively plausible generative model of language transformations, and heuristics for performing custom inference within the SMC framework. The result is a model which requires no exposure to Dutch, Dutch-specific heuristics, or training on parallel corpora in order to achieve human-like intercomprehension ability.

The model outlined in this paper instantiates resource-rational inference (Lieder & Griffiths, 2020), where the number of SMC particles is a loose proxy for cognitive resources used to process a sentence. Unlike past work which used SMC to model incremental language comprehension in terms of sequential, word-by-word processing (Levy et al., 2008; Clark et al., 2025b), our model applies SMC to iterative subproblems in a different sense: by gradually tempering in the contribution of an L1 prior, we model a comprehender who moves from inferences about surface-level similarity towards richer inferences that integrate priors about intended messages. Crucially, the ability to ground inference in similarity as a starting point allows successive improvements that unfold over time, modeling intercomprehension not as strictly left-to-right reading, but as an iterative refining and sharpening of inferences in a non-sequential way. A computational model of intercomprehension provides a useful tool

for testing hypotheses about the strategies and algorithms used by human comprehenders (which can be added or ablated from the current model), and for explaining phenomena such as asymmetric intelligibility (i.e. cases where speakers of Language A can understand Language B better than speakers of Language B can understand Language A).

6.2 Parallel processing and symbolic rules as core model components

Our ablations analysis demonstrates that two specific model features contribute to improved scores on the evaluation metrics. First, the use of parallel processing with 32 particles in an SMC inference framework produced improved scores relative to single-particle inference. A collection of multiple weighted particles appears better able to represent the wide uncertainty about the possible translation equivalents of a given word sequence during iterated inference, whereas a single particle may greedily commit to one option out of many and fail to more fully explore the space of alternatives. Second, the ability to infer word re-write rules enables a "stab in the dark" approach to translation; ablated runs without this feature, which were limited to proposing latent English strings using only surface form similarity, performed worse. When only a few words in a sentence remain to be translated, humans can often make an informed guess about the identities of the remaining words using context as a guide, e.g. as seen in the cloze task (Taylor, 1953; Staub et al., 2015; Kuperberg & Jaeger, 2016). If the guess results in a significantly better translation, it might be accepted even if the inferred L1 word bears little resemblance to the observed L2 word. This type of "a-ha moment" or sudden insight is studied in other contexts in cognitive science (Goodman et al., 2008; Tenenbaum et al., 2006; Kounios & Beeman, 2014; Rule et al., 2018; Dubey et al., 2025), and here we provide an algorithmic correlate of this type of inference under high uncertainty, which can nevertheless result in a successful belief update.

6.3 Augmenting small LLMs with Bayesian inference

Thanks to their pre-training, modern LLMs are strong multi-task learners (Radford et al., 2019). However, in cases where a particular problem admits decomposition into language modeling and symbolic structure, the performance of small LLMs can be improved using Bayesian inference with explicit, customizable generative models. This is consistent with past work that has leveraged SMC inference to enforce explicit constraints on language models (Loula et al., 2025; Lipkin et al., 2025). In domains with strong constraints or structural priors, or where the inputs are highly out-of-distribution relative to the training data, or where model size is a large constraint, approaches like the one employed in this work may have broader applicability beyond the cognitive modeling of intercomprehension.

6.4 Limitations and Future Work

The current work only considers intercomprehension in a reading setting, rather than a listening setting. Reading is different from listening in that orthographic cues play a large role, while phonetic cues are mediated by a reader's ability to infer how an unfamiliar word should be pronounced. Another major limitation of the generative model in its current form is the strong assumption of word order alignment between L1 and L2 sentences. Future iterations of this model can consider additional word order transformations as part of the generative model (e.g. a mapping from the positions of L1 words to the positions of their L2 equivalents), and a corresponding proposal function for inference which proposes changes to a particle's latent word order mapping. This approach would still face challenges in dealing with $L1 \rightarrow L2$ mappings that are not one-to-one, e.g. "not" (English) $\rightarrow$ "ne ... pas" (French). Additionally, allowing changes to word order between L1 and L2 sentences results in a combinatorial explosion in the number of transformations of a particular sentence, making inference significantly more difficult. However, the language pairs which have high rates of mutual intelligibility tend to be those which have similar word order rules (Gooskens et al., 2018), so this limitation may have a cognitive interpretation.

References

Karin Beijering, Charlotte Gooskens, and Wilbert Heeringa. Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm. *Linguistics in the Netherlands*, 25(1):13–24, 2008. doi:10.1075/avt.25.05bei.

Elisabetta Bonvino, Elisa Fiorenza, and Diego Cortés Velásquez. Observing Strategies in Intercomprehension Reading. Some Clues for Assessment in Plurilingual Settings. *Frontiers in Communication*, 3, 2018. ISSN 2297-900X. doi:10.3389/fcomm.2018.00029.

Maddalena De Carlo and Sandra Garbarino. Intercomprehension: Strengths and Opportunities of a Pluralistic Approach. In *The Routledge Handbook of Plurilingual Language Education*. Routledge, 2021.

J Carpenter, Peter Clifford, and Paul Fearnhead. Improved particle filter for non-linear problems. *IEE Proceedings Radar Sonar and Navigation*, 146(1):2–7, 1999. doi:10.1049/ip-rsn:19990255.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. Goldfish: Monolingual Language Models for 350 Languages, March 2026. URL http://arxiv.org/abs/2408.10441. arXiv:2408.10441 [cs].

Sihan Chen, Sarah Nathaniel, Rachel Ryskin, and Edward Gibson. The effect of context on noisy-channel sentence comprehension. *Cognition*, 238:105503, 2023. ISSN 00100277. doi:10.1016/j.cognition.2023.105503.

Thomas Hikaru Clark, Jacob Hoover Vigly, Edward Gibson, and Roger Levy. A Model of Approximate and Incremental Noisy-Channel Language Processing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47(0), 2025a. URL https://escholarship.org/uc/item/9kr1b1gm.

Thomas Hikaru Clark, Jacob Hoover Vigly, Edward Gibson, and Roger P. Levy. Resource-Rational Noisy-Channel Language Processing: Testing the Effect of Algorithmic Constraints on Inferences. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23659–23672, Suzhou, China, 2025b. Association for Computational Linguistics. doi:10.18653/v1/2025.emnlp-main.1207.

Marco Cusumano-Towner, Feras Saad, Alexander Lew, and Vikash Mansinghka. Gen: A General-purpose Probabilistic Programming System with Programmable Inference. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI 2019, pp. 221–236, Phoenix, AZ, USA, 2019. ACM. doi:10.1145/3314221.3314642.

Marco Cusumano-Towner, Alexander K. Lew, and Vikash K. Mansinghka. Automating Involutive MCMC using Probabilistic and Differentiable Programming, 2020. URL https://doi.org/10.48550/arXiv.2007.09871.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo Samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 2006. ISSN 1369-7412, 1467-9868. doi:10.1111/j.1467-9868.2006.00553.x.

Arnaud Doucet, Nando Freitas, and Neil Gordon (eds.). *Sequential Monte Carlo Methods in Practice*. Springer, New York, NY, 2001. doi:10.1007/978-1-4757-3437-9.

Peter Doye. A Methodological Framework for the Teaching of Intercomprehension. *Language Learning Journal*, 30(1):59–68, 2004. ISSN 0957-1736. doi:10.1080/09571730485200241.

Rachit Dubey, Mark Ho, Hermish Mehta, and Tom Griffiths. Aha! moments correspond to metacognitive prediction errors, 2025. URL https://doi.org/10.2139/ssrn.5475080.

Kevin Ellis, Armando Solar-Lezama, and Josh Tenenbaum. Unsupervised learning by program synthesis. *Advances in neural information processing systems*, 28, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/b73dfe25b4b8714c029b37a6ad3006fa-Abstract.html.

Edward Gibson, Leon Bergen, and Steven T. Piantadosi. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056, 2013. doi:10.1073/pnas.1216438110.

Walter R. Gilks and Carlo Berzuini. Following a moving target—Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):127–146, 2001. ISSN 1467-9868. doi:10.1111/1467-9868.00280.

Jelena Golubović and Charlotte Gooskens. Mutual intelligibility between West and South Slavic languages. *Russian Linguistics*, 39(3):351–373, 2015. ISSN 1572-8714. doi:10.1007/s11185-015-9150-9.

Noah D. Goodman, Joshua B. Tenenbaum, Jacob Feldman, and Thomas L. Griffiths. A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science*, 32(1):108–154, 2008. ISSN 1551-6709. doi:10.1080/03640210701802071.

Charlotte Gooskens and Wilbert Heeringa. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, 16(3):189–207, 2004. ISSN 1469-8021, 0954-3945. doi:10.1017/S0954394504163023. URL https://www.cambridge.org/core/product/9F99796CBB92E068C5ED1E26F3B2CD62.

Charlotte Gooskens, Vincent J. van Heuven, Jelena Golubović, Anja Schüppert, Femke Swarte, and Stefanie Voigt. Mutual intelligibility between closely related languages in Europe. *International Journal of Multilingualism*, 15(2):169–193, 2018. ISSN 1479-0718. doi:10.1080/14790718.2017.1350185.

Peter D. Grünwald. *The Minimum Description Length Principle*. The MIT Press, 2007. doi:10.7551/mitpress/4643.001.0001.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python. 2020. doi:10.5281/zenodo.1212303.

John Kounios and Mark Beeman. The Cognitive Neuroscience of Insight. *Annual Review of Psychology*, 65(Volume 65, 2014):71–93, 2014. ISSN 0066-4308, 1545-2085. doi:10.1146/annurev-psych-010213-115154.

Gina R. Kuperberg and T. Florian Jaeger. What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1):32–59, 2016. ISSN 2327-3798, 2327-3801. doi:10.1080/23273798.2015.1102299.

Roger Levy. A Noisy-Channel Model of Rational Human Sentence Comprehension under Uncertain Input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 234–243, Honolulu, Hawaii, 2008. Association for Computational Linguistics. doi:10.3115/1613715.1613749.

Roger Levy, Florencia Reali, and Thomas Griffiths. Modeling the effects of memory on human online sentence processing with particle filters. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/hash/a02ffd91ece5e7efeb46db8f10a74059-Abstract.html.

Alexander K. Lew, Tan Zhi-Xuan, Gabriel Grand, and Vikash K. Mansinghka. Sequential Monte Carlo Steering of Large Language Models using Probabilistic Programs, 2023. URL https://doi.org/10.48550/arXiv.2306.03081.

Falk Lieder and Thomas L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1, 2020. ISSN 0140-525X, 1469-1825. doi:10.1017/S0140525X1900061X.

Benjamin Lipkin, Benjamin LeBrun, Jacob Hoover Vigly, João Loula, David R. MacIver, Li Du, Jason Eisner, Ryan Cotterell, Vikash Mansinghka, Timothy J. O'Donnell, Alexander K. Lew, and Tim Vieira. Fast Controlled Generation from Language Models with Adaptive Weighted Rejection Sampling, 2025. URL https://doi.org/10.48550/arXiv.2504.05410.

João Loula, Benjamin LeBrun, Li Du, Ben Lipkin, Clemente Pasti, Gabriel Grand, Tianyu Liu, Yahya Emara, Marjorie Freedman, Jason Eisner, Ryan Cotterell, Vikash Mansinghka, Alexander K. Lew, Tim Vieira, and Timothy J. O'Donnell. Syntactic and Semantic Control of Large Language Models via Sequential Monte Carlo, 2025. URL https://doi.org/10.48550/arXiv.2504.13139.

Jiaming Luo, Yuan Cao, and Regina Barzilay. Neural Decipherment via Minimum-Cost Flow: From Ugaritic to Linear B, 2019. URL https://doi.org/10.48550/arXiv.1906.06718.

Jiaming Luo, Frederik Hartmann, Enrico Santus, Regina Barzilay, and Yuan Cao. Deciphering Undersegmented Ancient Scripts Using Phonetic Prior. *Transactions of the Association for Computational Linguistics*, 9:69–81, 2021. ISSN 2307-387X. doi:10.1162/tacl_a_00354.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., 1982.

Christian A. Naesseth, Fredrik Lindsten, and Thomas B. Schön. Elements of Sequential Monte Carlo, 2024. URL https://doi.org/10.48550/arXiv.1903.04797.

Iftekhar Naim, Parker Riley, and Daniel Gildea. Feature-Based Decipherment for Machine Translation. *Computational Linguistics*, 44(3):525–546, 2018. doi:10.1162/coli_a_00326.

Radford M. Neal. Annealed Importance Sampling, 1998. URL https://doi.org/10.48550/arXiv.physics/9803008.

Jessica Nieder and Johann-Mattis List. A Computational Model for the Assessment of Mutual Intelligibility Among Closely Related Languages, February 2024. URL http://arxiv.org/abs/2402.02915. arXiv:2402.02915 [cs].

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Sujith Ravi and Kevin Knight. Deciphering Foreign Language. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 12–21, Portland, Oregon, USA, 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1002/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019. URL https://doi.org/10.48550/arXiv.1908.10084.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978. ISSN 0005-1098. doi:10.1016/0005-1098(78)90005-5.

Joshua Rule, Eric Schulz, Steven T. Piantadosi, and Joshua B. Tenenbaum. Learning list concepts through program induction, 2018. URL https://doi.org/10.1101/321505.

Joshua S. Rule, Joshua B. Tenenbaum, and Steven T. Piantadosi. The Child as Hacker. *Trends in Cognitive Sciences*, 24(11):900–915, 2020. ISSN 1364-6613, 1879-307X. doi:10.1016/j.tics.2020.07.005.

Adrian Staub, Margaret Grant, Lori Astheimer, and Andrew Cohen. The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, 82:1–17, 2015. ISSN 0749596X. doi:10.1016/j.jml.2015.02.004.

Wilson L. Taylor. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433, 1953. ISSN 0022-5533. doi:10.1177/107769905303000401.

Joshua B. Tenenbaum and Thomas L. Griffiths. Generalization, similarity,and Bayesian inference. *Behavioral and Brain Sciences*, 24(4):629–640, 2001. ISSN 1469-1825, 0140-525X. doi:10.1017/S0140525X01000061.

Joshua B. Tenenbaum, Thomas L. Griffiths, and Charles Kemp. Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318, 2006. ISSN 1364-6613, 1879-307X. doi:10.1016/j.tics.2006.05.009.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331(6022):1279–1285, 2011. ISSN 0036-8075, 1095-9203. doi:10.1126/science.1192788.

Jörg Tiedemann and Santhosh Thottingal. OPUS-MT – Building open translation services for the World. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada (eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 479–480, Lisboa, Portugal, 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.61/.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 Text Embeddings: A Technical Report, February 2024. URL https://arxiv.org/abs/2402.05672v1.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, 2023. URL https://doi.org/10.48550/arXiv.2201.11903.

Fei Xu and Joshua B. Tenenbaum. Word learning as Bayesian inference. *Psychological Review*, 114(2):245–272, 2007. ISSN 0033-295X. doi:10.1037/0033-295X.114.2.245.

Tan Zhi-Xuan, Jordyn Mann, Tom Silver, Josh Tenenbaum, and Vikash Mansinghka. Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems*, 33:19238–19250, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/df3aebc649f9e3b674eeb790a4da224e-Abstract.html.

A Full Generative Model Specification

The generative model defines a joint distribution over an L_1 sentence S, and a rule library $\mathcal{L}$ given an observed L_2 sentence $\mathbf{y} = (y_1, \ldots, y_T)$.

L_1 sentence prior.

$$S \sim p_{\text{LM}}(\cdot)$$

where p_{LM} is a Goldfish language model distribution over L1 word sequences. We use the following Goldfish models:

L1	Model String	Training Data Size	# Parameters
English	`goldfish-models/eng_latn_1000mb`	1000 MB	125M
Spanish	`goldfish-models/spa_latn_1000mb`	1000 MB	125M
Russian	`goldfish-models/rus_cyrl_1000mb`	1000 MB	125M

Table 1: Caption

Sentence probability is additionally weighted by $p_{\text{EOS}}(S)$, the LM probability of an end-of-sentence token following S, which favors grammatically complete sentences.

Rule library prior. A rule library $\mathcal{L}$ is a set of string rewrite rules $r = (r_{\text{in}}, r_{\text{out}})$, where r_{in} is a regex pattern and r_{out} a replacement string. The prior penalizes complex libraries:

$$\log p(\mathcal{L}) = -\lambda_{\mathcal{L}} \sum_{r \in \mathcal{L}} \big(|r_{\text{in}}| + |r_{\text{out}}|\big)$$

where $|\cdot|$ is character length and $\lambda_{\mathcal{L}} = 2.0$.

Per-word noise model. Let w_t be the t-th word of S, and let

$$\mathcal{L}(w_t) = \big\{ r(w_t) : r \in \mathcal{L},\ r_{\text{in}} \text{ matches } w_t \big\}$$

be the set of words obtainable by applying a matching library rule to w_t. For each position $t = 1, \ldots, T$:

$$\text{applyRule}_t \sim \text{Bernoulli}\left(\begin{cases} 0.9 & \text{if } \mathcal{L}(w_t) \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \right)$$

$$\log p(y_t \mid w_t,\, \mathcal{L},\, \text{applyRule}_t) = \begin{cases} \log \dfrac{1}{|\mathcal{L}(w_t)|} \cdot \mathbf{1}[y_t \in \mathcal{L}(w_t)] & \text{if applyRule}_t = 1 \\ -\lambda_{\text{noise}} \cdot \text{EditDist}(w_t,\, y_t) & \text{if applyRule}_t = 0 \end{cases}$$

where $\lambda_{\text{noise}} = 2.0$ and EditDist is the simple mean of the character-level edit distance (orthographic likelihood) or ARPAbet phoneme edit distance based on a nonce-word pronunciation module (phonetic likelihood).

Joint distribution.

$$\begin{aligned} p(S, \mathcal{L}, \{\text{applyRule}_t\} \mid \mathbf{y}) \;\propto\; & p_{\text{LM}}(S)\; p_{\text{EOS}}(S)\; p(\mathcal{L}) \\ & \times \prod_{t=1}^{T} p(\text{applyRule}_t \mid w_t, \mathcal{L})\; p(y_t \mid w_t, \mathcal{L}, \text{applyRule}_t) \end{aligned}$$

B Inference details

Tempered SMC inference target. Inference uses tempered SMC with a linear inverse-temperature schedule $\beta_k = k/K, k = 0, \ldots, K$. At step k the target is:

$$p_{\beta_k}(S, \mathcal{L} \mid \mathbf{y}) \propto p_{\text{LM}}(S)^{\beta_k}\ p(\mathcal{L}) \prod_{t=1}^{T} p(y_t \mid w_t, \mathcal{L})$$

At $\beta_0 = 0$ the target is determined solely by the edit-distance likelihood; at $\beta_K = 1$ it equals the full posterior.

Weight updates. For a sequence of β_k values $0 < \beta_1 < \beta_2 < \cdots < \beta_K = 1$, we compute an incremental importance ratio, which bridges between one target distribution and the next:

$$\frac{\pi_{\beta_k}(z)}{\pi_{\beta_{k-1}}(z)} \propto \frac{p(y \mid z) p(z)^{\beta_k}}{p(y \mid z) p(z)^{\beta_{k-1}}} = p(z)^{\Delta\beta_k}$$

Thus the weight update for each particle is given by: $w_k^{(i)} \leftarrow w_{k-1}^{(i)} \cdot p\left(z^{(i)}\right)^{\Delta\beta_k}$. At the final step of inference, the algorithm yields a weighted collection of particles corresponding to approximate samples from the final posterior target distribution $\pi_{\text{post}}(z)$.

Resampling. After each weight update, we compute the Effective Sample Size (ESS) using the formula $\text{ESS} = (\sum w^{(i)})^2 / \sum (w^{(i)})^2$. We resample the particles if the ESS drops below a preset value, using systematic resampling to reduce variance (Carpenter et al., 1999).

Metropolis-Hastings Accept-Reject Moves. Each proposal function defines a transition probability $q(z' \mid z)$ from one particle latent state to another. Using the Metropolis-Hastings algorithm, the acceptance probability for a rejuvenation move is given by:

$$\alpha = \min\left(1, \frac{p(y \mid z')\ p(z')^{\beta_k}\ q(z \mid z')}{p(y \mid z)\ p(z)^{\beta_k}\ q(z' \mid z)}\right)$$

This quantity is tracked automatically given the fully specified generative model using the Gen probabilistic programming language (Cusumano-Towner et al., 2019; 2020), which includes Metropolis-Hastings as part of the standard inference toolkit. As a practical note, since the inference process requires many calls to the language model, often with identical arguments across iterations and particles, significant efficiency gains result from utilizing a cached LM, following Lew et al. (2023). We design kernels to have the property that the quantities $q(z \mid z')$ and $q(z' \mid z)$ are both non-zero: in other words, we enforce the property that each kernel which updates a particle x to a proposed new particle x' be capable of inverting this update by calling itself on the new particle with some probability.

C Datasets

For each L1-L2 language pair, the evaluation dataset is drawn from the OpenSubtitles parallel movie subtitles dataset accessed via OPUS (Tiedemann & Thottingal, 2020) by randomly sampling 100,000 sentences. From this starting sample, we filter out sentences longer than 10 words or shorter than 4 words; where the L1 and L2 versions of the sentence differ in the number of words; which contain special symbols or non-standard punctuation; which contain capitalized words not at the beginning of the sentence (to remove proper nouns and acronyms); where the L1 translation contains words not in a wordlist of the top V L1 words according to the `wordfreq` package. We then apply dependency parsing with spaCy (Honnibal et al., 2020) to sentences (both L1 and L2), classify sentence pairs as having matching or non-matching dependency structures, and sample 150 sentences from each category to introduce variation in sentence structure match. Next, we apply an independent automated machine translation step to the L2 sentences using `meta-llama/Llama-3.3-70B-Instruct`. We filter out sentences where the machine-translated L1 output does not match the ground truth L1 sentence in OPUS. This filters out non-literal, idiomatic translations, e.g. *M'n baard is korter* (Dutch), which is translated as "I shortened my beard" in the parallel English subtitles but as "My beard is shorter" by machine translation (a more literal translation). As a final step, filtered sentences are manually reviewed to exclude potentially offensive content. The result of this process is a set of 100 parallel simple sentences, which are selected to elicit a dynamic range of intercomprehension from L1 speakers who are otherwise unfamiliar with the L2 (we note that these are not unbiased samples from the baseline distribution over L2 sentences).

D Computation of the Human–Model Wasserstein Distance

For each item we compare two distributions over sentences: the empirical distribution of human responses and the model's posterior over candidate translations. We summarize their discrepancy with the discrete 1-Wasserstein distance (earth mover's distance) under three sentence-level ground metrics. The distance is computed exactly per item by solving the optimal-transport linear program; we do *not* use the closed-form one-dimensional Wasserstein estimator.

Distributions. Let an item have n human responses $h_1, \ldots, h_n$. We place the uniform empirical measure $a = (1/n, \ldots, 1/n)$ over them, so identical responses contribute proportionally to their frequency. For model inferences, the log particle weights w_p are exponentiated and normalized, $\tilde{w}_p = e^{w_p} / \sum_{p'} e^{w_{p'}}$, each particle's final sentence is standardized (lower-cased, stripped of punctuation), and the weights of identical standardized sentences are summed. This yields a distribution $b = (b_1, \ldots, b_m)$ over m distinct model sentences $g_1, \ldots, g_m$, with $\sum_k b_k = 1$.

Ground metrics. For each pair (h_i, g_k) we form a cost C_{ik} under the three metrics of character edit distance, word edit distance, and sentence embedding cosine distance, producing three cost matrices $C \in \mathbb{R}^{n \times m}$.

Optimal transport. The 1-Wasserstein distance is the minimum-cost coupling between a and b:

$$W_1(a, b; C) \;=\; \min_{\gamma \in \mathbb{R}^{n \times m}_{\geq 0}} \sum_{i=1}^{n} \sum_{k=1}^{m} \gamma_{ik} C_{ik} \quad \text{s.t.} \quad \sum_k \gamma_{ik} = a_i, \;\; \sum_i \gamma_{ik} = b_k. \tag{1}$$

We solve this linear program exactly with the HiGHS solver (`scipy.optimize.linprog`), treating γ as nm nonnegative variables with objective vector $\mathrm{vec}(C)$. The constraint set comprises the m column-marginal equalities and $n - 1$ of the row-marginal equalities (one row constraint is omitted as redundant, since both marginals are normalized). The optimal objective value is reported as W_1; the recovered γ gives the optimal transport plan. We compute W_1 separately for each of the three cost matrices, yielding three values per item.

E Example Model Inferences

Dutch Sentence	Human Inferences	Model Inferences
slaap is het beste (sleep is the best)	slap is the best sleep is the best	sleep is a beast sleep is the best
geef haar een magisch moment (give her a magical moment)	Give her a magical moment Good have a magic moment	geek hair and magic moments gee hear a magic moment
het is een vorm van communicatie (it is a form of communication)	He is bad at communicating what is in worm can communicate	It is a form of communication hat is a form of communication
wij hebben de papieren (we have the papers)	Do we have any paper Give me the paper	we happen on paper we happen to be
het was een droom (it was a dream)	He was in a room He was in his room	it was a dream i was in doom
ik heb goed nieuws voor jou (i have good news for you)	I have good news for you I heard good news for you	i have good news for you we have good news for you
hij werkt te hard (he worked too hard)	He worked too hard He's worked too hard	the work is hard i worked it hard
we hadden kalm moeten blijven (we should have stayed calm)	We had a calm moderate blizzard We had a calm modest evening	the sudden calm met oblivion a sudden calm met oblivion
dit is een militaire post (this is a military post)	Did it happen at the military post? It is a military base	this is a military post it is a military post

Table 2: Example human and model inferences for Dutch sentences from the dataset. Each sentence is shown with two human-generated and two model-generated inferences.

F Inferred rules and dependency structure match analyses for non-English language pairs

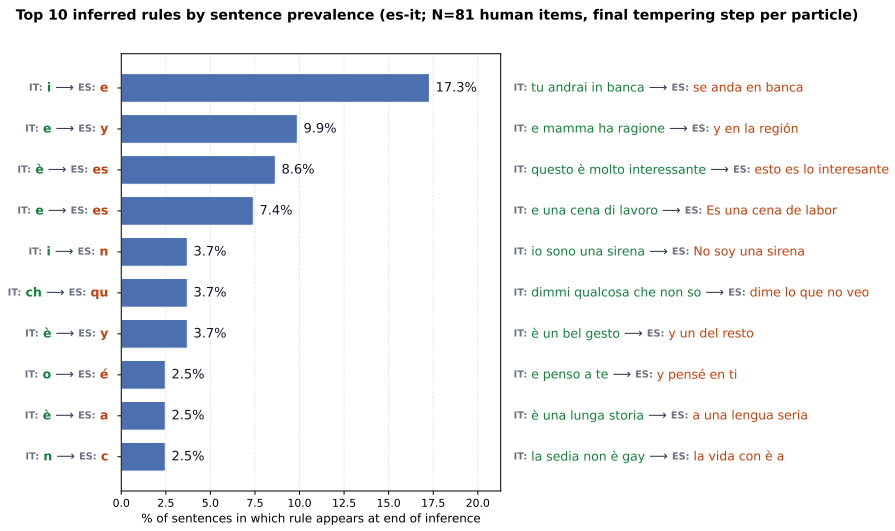

Figure 8: Top inferred subword rules for Italian → Spanish. The patterns show both correct inferences (*ch* → *qu*), as well as incorrect inferences (*i* → *n*).

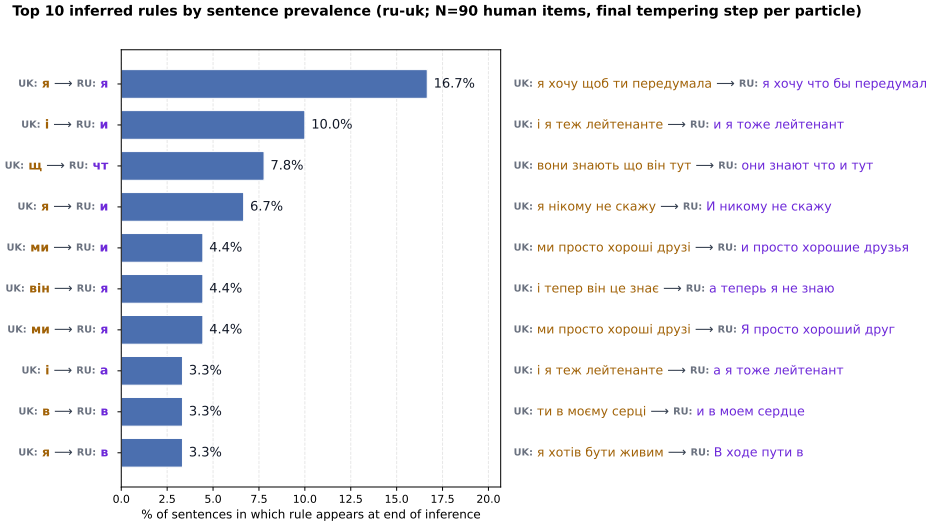

Figure 9: Top inferred subword rules for Ukrainian → Russian. The patterns show both correct inferences (щ → чт), as well as incorrect inferences (він → я).

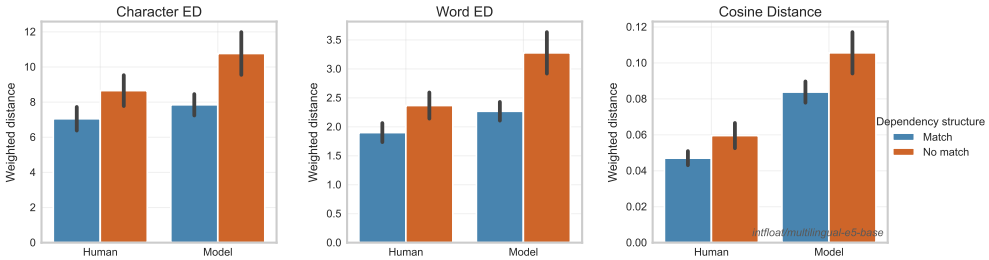

Figure 10: Comparison of distance between human/model inference and ground truth translation, as a function of dependency structure match, for the Spanish-Italian language pair. Both humans and models show a closer ground-truth match when dependency structures are matched.

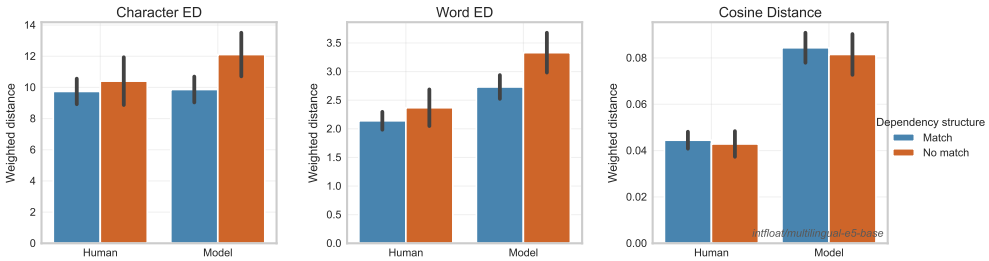

Figure 11: Comparison of distance between human/model inference and ground truth translation, as a function of dependency structure match, for the Russian-Ukrainian language pair. Both humans and models show a closer ground-truth match when dependency structures are matched.

G Zero-Shot Prompting of the LLM Baselines

The LLM baselines (Llama 3.2 1B, Gemma 3 1B, and SmolLM2 135M) are prompted zero-shot: each stimulus sentence is sent as a single user turn to the model's chat endpoint, with no system prompt, no in-context examples, and no conversation history. The prompt names only the target (reference) language and instructs the model to emit the translation alone; it does not name the source (stimulus) language, so the model must recognize the unfamiliar language on its own, enabling more fair comparison to our model. Each item is translated independently in a fresh context.

Zero-Shot Translation Prompt (Template)

```
> Translate '{sentence}' into {reference_name}. Only produce the {reference_name}
translation and no other output.
```

Here `{sentence}` is the stimulus sentence verbatim and `{reference_name}` is the human-readable name of the reference language (*English*, *Spanish*, or *Russian*). Instantiating the template for one item from each of the three language pairs gives:

Zero-Shot Translation Prompt (Instantiated Examples)

Dutch → English

```
> Translate 'calvin had een idee' into English. Only produce the English translation
and no other output.
```

Italian → Spanish

```
> Translate 'abbiamo già abbastanza problemi' into Spanish. Only produce the Spanish
translation and no other output.
```

Ukrainian → Russian

```
> Translate 'а я так і не отримала чек' into Russian. Only produce the Russian
translation and no other output.
```

Decoding and serving. All three baselines are served locally through an OpenAI-compatible chat-completions endpoint in LM Studio[4]. We decode greedily, with temperature 0, nucleus sampling parameter top-$p = 0.9$, and a cap of 512 generated tokens; the model's response is stripped of surrounding whitespace and used as the translation with no further post-processing. The specific checkpoints are `Llama-3.2-1B-Instruct`, `gemma-3-1b-it`, and `SmolLM2-135M-Instruct` (bf16 weights).

[4]Version 0.4.16, https://lmstudio.ai/