

Translation-Induced Label Drift in Natural Language Inference: Evidence from Nine Languages

Muhammad S. Abdo, Phakphum Artkaew, Sadia Ashraf,
Elder Henrique Attala e Paiva, Jose Benavides, Hale Karkoutli, Melissa Loeffers,
Elias Naske, Wenchen Shi, Jinyuan Wang, Sandra Kübler
Indiana University
{mabdo, skuebler}@iu.edu

Abstract

Multilingual natural language inference (NLI) benchmarks are routinely constructed by translating English datasets into target languages, based on the assumption that inference relations between premise–hypothesis pairs are preserved under translation. This assumption, however, has rarely been tested empirically at scale. We investigate this by conducting a large-scale human annotation study across nine typologically diverse languages: Arabic, German, Spanish, Portuguese, Thai, Urdu, Turkish, Mandarin, and Cantonese. We translated 1,000 English premise–hypothesis pairs, sampled across four linguistically challenging phenomena, and annotated each target-language version by a native-speaker linguist. Additionally, a native English speaker annotated the source pairs under the same protocol, allowing us to distinguish source-label disagreement from translation-induced change. While agreement between translated-language annotations and the original dataset labels ranges from 55.1% to 67.4%, agreement with the native-English annotation ranges from 57.8% to 64.2%. Entailment and Contradiction frequently shift toward Neutral, but the English control shows that some apparent drift is already present in the source items. These results call into question using transfer of English labels and highlight the need for native-language annotation in multilingual NLI evaluation.

1 Introduction

Natural Language Inference (NLI) aims to determine the logical relationship between two pieces of text. Given a *premise* p (e.g., *The cat is asleep.*) and a *hypothesis* h (e.g., *An animal is asleep.*), the task is to classify their relationship into one of three mutually exclusive categories: *Entailment* (E), where the hypothesis necessarily follows from the premise; *Contradiction* (C), where the hypothesis is incompatible (i.e., contradictory) with the premise; and *Neutral* (N), where the premise does

not provide sufficient information to infer the truth or falsity of the hypothesis (MacCartney, 2009).

However, to alleviate the bottleneck of data scarcity across languages, several NLI datasets, e.g., XNLI (Conneau et al., 2018), are constructed by translating high-resource benchmarks (typically English) into other languages. This approach assumes that inference relations remain stable after translation, an assumption that has rarely been tested empirically at scale. Translation may introduce difficult-to-disentangle shifts in meaning, cultural context, or linguistic structure. The problem may be exacerbated by machine translation (MT), which can introduce noise, semantic distortions, and structural artifacts that alter the intended inferential relations between sentences (Artetxe et al., 2020; Kabir et al., 2026).

This study aims to investigate this problem under the broad question: Are NLI labels preserved under MT into other languages? Or do these relations undergo label drift as a consequence of translation? We address this question through a large-scale empirical analysis of cross-lingual label stability across Arabic, German, Spanish, Portuguese, Thai, Urdu, Turkish, Mandarin, and Cantonese. We also compare inter-language drift to intra-language drift to be able to distinguish annotator drift from translation drift.

2 Related Work

Translating English NLI datasets into target languages while preserving the original labels has been a central methodology in multilingual NLI research. For instance, Agić and Schluter (2018) were among the early adopters of this approach by manually translating 1,332 English test pairs into Arabic, French, Spanish, and Russian. They then verified whether the original inferential relations still held true after translation. Their findings show that automatically translated test sets fell short of

performance on manually translated data by an average of 2.57%, with Arabic showing considerably larger degradation than French or Spanish, which is an early signal that translation quality is not uniform across languages.

The XNLI benchmark by [Conneau et al. \(2018\)](#) established this translation-based methodology at scale by extending MultiNLI ([Williams et al., 2018](#)) to fifteen languages through professional translation of development and test sets while directly transferring original English labels. To establish baselines, they built two datasets by machine translating the English training data into target languages (*Translate Train*) or by translating test-time input sentences back to English at inference time (*Translate Test*). Even though their validation study revealed minimal label drift, where bilingual annotators recovered the original English consensus label 85% of the time on the source data and 83% on French translations, performance varied sharply across languages, with accuracy dropping from 73.7% on English to as low as 59.3% on Urdu under *Translate Test*. This approach has been extended by subsequent multilingual NLI resources. E.g., IndicXNLI ([Aggarwal et al., 2022](#)), Myanmar XNLI ([Htet and Dras, 2025](#)), Croatian XNLI ([Obadić et al., 2023](#)), and XNLIeu ([Heredia et al., 2024](#)) all hinge, to varying degrees, on translation or post-editing. However, findings across these studies are mixed. IndicXNLI reports high label fidelity, with 82% error consistency between human and MT. The remaining studies identify systematic problems such as mistranslation, polarity shifts, omitted information, lexical ambiguity.

These studies show that label stability in multilingual NLI is neither uniform nor guaranteed. This is further corroborated by [Budur et al. \(2020\)](#), who demonstrate that even when majority-level consistency exceeds 95%, a residual inconsistency rate of 10–15% persists, owing to lexical mismatches or language-specific grammatical requirements disambiguating information. [Bui et al. \(2025\)](#) further show that the direction of translation matters, i.e., when Vietnamese served as the source language, Vietnamese-English pairs yield higher translation quality and stronger downstream NLI performance than Vietnamese-Chinese pairs. Additionally, [Rajaei et al. \(2025\)](#) demonstrate that cross-lingual stability varies across translation pipelines, and that translation reliability can be contingent on both the task and the target language. In combination, prior work shows that

translation can affect multilingual NLI reliability, but the field still lacks a large-scale, cross-lingual study that directly tests label stability through human re-annotation across multiple languages and linguistically controlled phenomena.

3 Data and Annotation

3.1 Data Sources

For our evaluation dataset, we sampled 1,000 English sentence pairs by drawing 200 pairs from each of five established NLI resources: MultiNLI ([Williams et al., 2018](#)), ANLI ([Nie et al., 2020](#)), WANLI ([Liu et al., 2022](#)), NLI Fever ([Nie et al., 2019](#)), and LingNLI ([Parrish et al., 2021](#)). Each of these 1,000 pairs was selected to represent one or more linguistic phenomena known to complicate semantic interpretation and inference. These phenomena are conditionality (300 instances), intensionality (300 instances), modality (300 instances), and comparative constructions (100 instances) ([MacCartney, 2009](#); [Richardson et al., 2019](#); [Haruta et al., 2020](#)).

Each of these phenomena introduces a distinct source of inferential complexity. Conditional constructions make Entailment contingent on hypothetical antecedents ([Adams, 1965](#)), intensionality shifts evaluation from real-world referents to hypothetical contexts ([Liefke, 2024](#)); modality weakens the strength of assertion through expressions of necessity or possibility ([Moltmann, 2020](#)); and comparative constructions introduce scope ambiguities that affect what can be entailed from a premise ([Haruta et al., 2022](#)).

Stratifying the four phenomena was rule-based. We identified candidate conditional items using lexical markers such as *if* and *when*. For intensional candidates, we curated a list of predicates such as *believe*, *know*, *try*, and *want*, and for modality we used modal auxiliaries and related expressions. Finally, comparative/quantifier scope used cues such as *more*, *less*, *few*, *at least*, and related scalar expressions within the same clause. These rules constitute retrieval heuristics, and each pair was assigned to a single phenomenon: hence labels should not be interpreted as indicating that an item contains only one phenomenon.

3.2 Machine Translation

We then translated this dataset into nine target languages using the GPT-4o-mini API, which we selected after preliminary testing showed that several

Language	E%	N%	C%	Valid	Excl.	Total
AR	23.93	54.07	22.00	982	18	1000
CN	41.17	32.08	26.75	957	43	1000
DE	32.42	43.94	23.64	990	10	1000
EN	40.12	33.33	26.55	987	13	1000
ES	26.29	50.72	22.98	966	34	1000
PT	35.45	33.71	30.84	976	24	1000
TH	41.27	28.48	30.25	962	38	1000
TR	39.98	35.17	24.86	873	127	1000
UR	41.07	32.29	26.64	991	9	1000
ZH	37.29	43.75	18.96	944	56	1000

Table 1: Distribution of human NLI annotations by target language. Percentages are calculated over valid annotations only.

other models such as seamless-m4t-v2-large (Barraut et al., 2023), NLLB-200 3.3B (Costa-Jussà et al., 2022), and Qwen2-72B-Instruct (Yang et al., 2024) produced translation artifacts and unacceptable hallucination rates. No post-editing was performed prior to annotation.

3.3 Human Annotation

To ensure a high quality of annotation, for each target language, the translated pairs were annotated by a native speaker with formal training in linguistics using standard NLI guidelines. Each annotator was instructed to judge inference relations based solely on the translated text, without access to the original English sentences, original labels, or the four linguistic phenomena. Labels included Entailment, Contradiction, Neutral, and a reserved *NonSense* category for cases where the annotators found the translation incoherent or uninterpretable. The annotations were carried out via a web-based annotation interface. Before the annotations, annotators completed a calibration phase where the label definitions and representative examples were discussed. Annotators proceeded in self-paced sessions, which included periodic spot checks. Data-related questions answered without revealing the source labels. A native English speaker re-annotated the 1,000 source pairs under the same protocol. Of the 1,000 pairs, 987 received a valid E/N/C judgment by the English annotator, and agreement with the original dataset label was 63.0% ($\kappa = 0.44$).

Table 1 reports the label distribution across languages after excluding *NonSense* annotations. We will use the term *original dataset labels* to refer exclusively to the NLI labels assigned to the original English sentence pairs. We use *native-*

English annotation for the monolingual control and *translated-language annotation* for target-language judgments. All data used in the study and annotation guidelines are publicly available.¹

3.4 NonSense Annotations

Among the 9,000 translated pairs, 359 (4.0%) were labeled as *NonSense*. Urdu has the lowest rate of pairs judged as *NonSense* (0.9%), whereas Turkish has the highest (12.7%), followed by Cantonese (5.6%) and Mandarin (4.3%). Conditionality contributes 164 of the 359 exclusions and has the highest phenomenon-level rate (6.1%); the rates for comparatives/quantifier scope, intensionality, and modality are 3.7%, 2.8%, and 3.2% respectively. Note that *NonSense* does not uniquely diagnose translation failure. Some excluded examples are already malformed at source. For instance, one pair has the English hypothesis “*Third World debt has a xx.*” Other instances show recognizable translation corruption, e.g., in Turkish, “cork floors” becomes “sponge floors.” We therefore exclude *NonSense* from the agreement estimates.

4 Methodology

4.1 Evaluation Framework

We adopt a multi-dimensional framework to evaluate label stability cross-linguistically under MT: Every translation is compared against two English references: (i) the original dataset label and (ii) the independent native-English annotation. Under this framework, we distinguish between *label preservation*, *semantic drift*, and *annotation uncertainty* as follows:

Label preservation measures the extent to which English-reference labels are preserved after translation using exact agreement (accuracy), Cohen’s κ , and multiclass Matthews correlation coefficient (MCC).

Cross-language label divergence evaluates agreement among language annotations using pairwise agreement, full-agreement and disagreement rates, and agreement with each of the two English references: the original dataset label and the native-English annotation.

Per-label stability isolates which NLI classes are most vulnerable to drift by examining per-class recall across languages.

Label Transition Dynamics models semantic drift by analyzing how labels change from each of

¹<https://github.com/Muhsabry/NLI10>

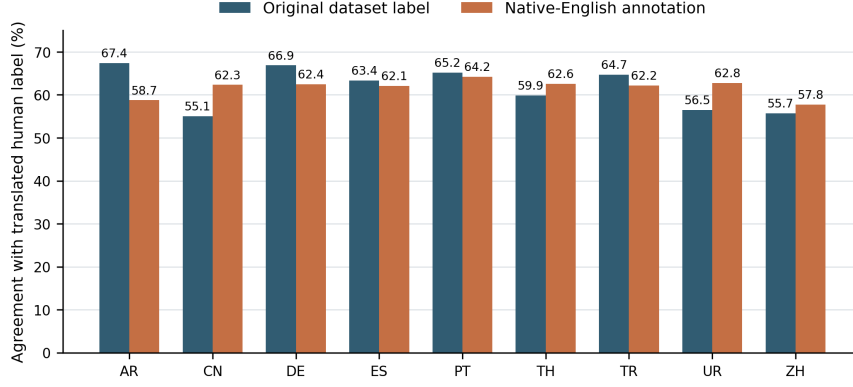


Figure 1: Agreement between translated-language human annotations and two English references.

the two English baselines, i.e., the original dataset label and the native-English annotation, to the translated-language human annotation. It specifically examines whether transitions between label classes are directionally biased rather than symmetric.

Phenomenon-driven drift focuses on whether the divergence of labels is associated with either of the four linguistic phenomena by comparing human label-change rates under both English references.

Systemic vs. language-specific drift assesses whether drift patterns are stable across languages by computing the variance of evaluation metrics, where low variance indicates systemic rather than language-specific effects.

Model-human agreement uses three pre-trained multilingual NLI models (mDeBERTa, XLM-R, and MiniLM) and an English-supervised DeBERTa-v3-large baseline as diagnostic tools to assess whether systems that are trained for NLI assign the same labels as human annotators. Native English annotations serve as a matched monolingual comparison (rather than the upper bound). Since the DeBERTa checkpoint was fine-tuned on the same five source NLI datasets, it is treated as a descriptive zero-shot transfer baseline.

Together, these metrics capture not only whether labels change, but how and under which conditions such changes do occur.

5 Results and Analysis

5.1 Label Preservation

This dimension compares the labels assigned to the translated sentence pairs with both English references. Figure 1 reports exact agreement across the

Language	Original dataset label				Native-English annotation			
	<i>n</i>	Agr.	κ	MCC	<i>n</i>	Agr.	κ	MCC
AR	982	.674	.484	.489	972	.587	.381	.403
CN	957	.551	.327	.334	945	.623	.426	.426
DE	990	.669	.489	.490	980	.624	.432	.437
ES	966	.634	.424	.426	955	.621	.431	.449
PT	976	.652	.477	.483	966	.642	.460	.462
TH	962	.599	.404	.417	951	.626	.431	.432
TR	873	.647	.465	.472	862	.622	.422	.423
UR	991	.565	.348	.356	980	.628	.433	.433
ZH	944	.557	.316	.319	933	.578	.354	.359
Mean	960	.616	.415	.421	949	.617	.419	.425

Table 2: Exact agreement (Agr.), Cohen’s κ , and multi-class MCC under the two English reference conditions.

target languages. It shows that translated pairs often receive labels different from the English labels.

Agreement with the original dataset labels ranges from a high of 67.4% for Arabic to a low of 55.1% for Mandarin, with a mean of 61.6%. Agreement with the native-English annotation ranges from a high of 64.2% for Portuguese to a low of 57.8% for Cantonese, with a mean of 61.7%. These figures indicate meaningful categorical divergence from both English references.

Table 2 reports chance-corrected Cohen’s κ and multiclass Matthews correlation coefficient (MCC). The same pattern appears under both metrics: mean κ is 0.415 against the original labels and 0.419 against the native-English annotation; mean MCC is 0.421 and 0.425 respectively. These measures supplement exact agreement by accounting for the three-class label distribution.

5.2 Cross-language Label Divergence

This evaluation is concerned with whether the same English premise–hypothesis pairs receive convergent NLI labels when translated into different languages. Since each target language was labeled by a single annotator, the agreement mea-

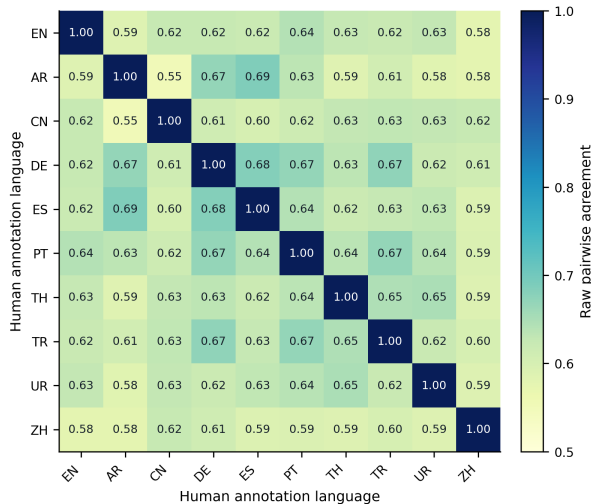


Figure 2: Pairwise raw agreement among the native-English (EN) and 9 translated human annotations.

asures do not refer to agreement among multiple annotators within the same language. Figure 2 shows that pairwise agreement among the nine translated-language annotations ranges from 55.1% to 68.6%, with a mean of 62.4%. Pairwise agreement between the English annotator and each translated-language annotator ranges from 57.8% to 64.2%.

Full cross-language convergence is much lower: among the 750 items with valid E/N/C judgments in all nine translations, all nine annotators select the same relation for only 123 items (16.4%). This pattern indicates that translated versions of the same source item often lead to different NLI judgments across languages; however, the remaining 83.6% should not be interpreted as a per-language error rate.

5.3 Per-label Stability

Next, we investigate which source relations are most susceptible to change. Figure 3 reports label recall under both references. Against the original dataset labels, aggregate recall is 71.8% for Entailment, 56.8% for Neutral, and 58.7% for Contradiction. Against the native-English annotation, the corresponding values are 63.3%, 62.2%, and 58.5%. Contradiction is therefore least stable under the English-human baseline, whereas Neutral is slightly less stable than Contradiction under the original-label baseline.

Language-level profiles are also heterogeneous, e.g., original-label Neutral recall is 78.3% in Arabic but 43.3% in Thai, while native-English Contradiction recall ranges from 42.1% in Cantonese

Phenomenon	Orig. (%)	Native (%)	Δ
Comparatives / quant. scope	32.8	33.4	+0.6
Conditionality	38.4	43.0	+4.6
Intensionality	39.5	37.4	-2.1
Modality	39.1	36.4	-2.7

Table 3: Human label-change rates by sampling phenomenon and English reference.

to 70.0% in Portuguese. Hence, it is not the case that a single relation is uniformly fragile in every language; label sensitivity interacts with both language and the choice of English reference.

5.4 Label Transition Dynamics

Figure 4 reports aggregate transition matrices for both references. Against the original dataset labels, 21.0% of Entailment and 30.3% of Contradiction cases move to Neutral; 56.8% of Neutral cases remain Neutral. Against the native-English annotation, the corresponding transitions to Neutral are 28.0% and 28.6%, while Neutral stability rises to 62.2%. The tendency toward Neutral therefore survives the monolingual control, but its magnitude and the apparent vulnerability of each source class change with the reference. Neutral cases also move outward: under the native-English baseline, 19.5% become Entailment and 18.3% become Contradiction. Translation can thus weaken or strengthen inferential commitment rather than operating in only one direction.

5.5 Phenomenon-driven Drift

Table 3 presents label-change rates for the four linguistic phenomena under both English references. Overall, the rates vary across phenomena rather than being concentrated in a single category. Under the original-label reference, drift ranges from 32.8% for comparatives/quantifier scope to 39.5% for intensionality. Under the native-English reference, it ranges from 33.4% for comparatives/quantifier scope to 43.0% for conditionality. Conditionality has the highest overall drift rate under the native-English reference, while comparatives/quantifier scope has the lowest rate under both references. This pattern suggests that translation-induced label drift is shaped by specific semantic phenomena, but the strength of this association varies by reference, language, and phenomenon. The detailed language-by-phenomenon breakdown is reported in Appendix B.

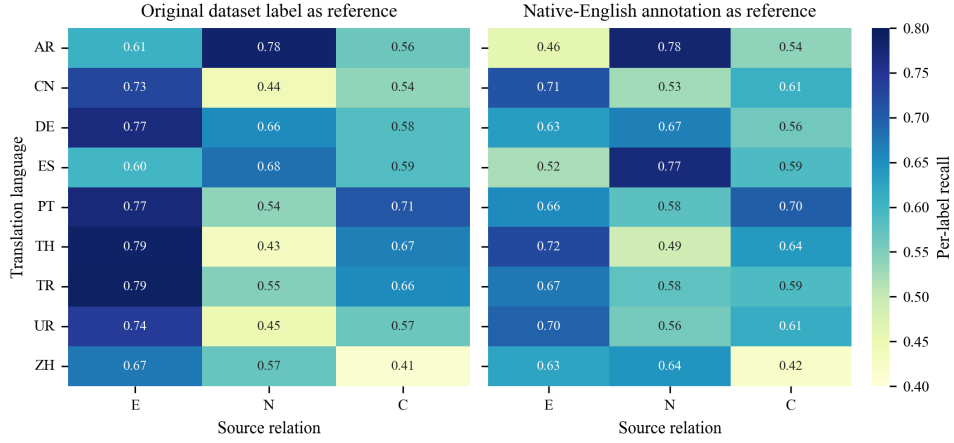


Figure 3: Per-label recall by translation language under the original-label and native-English reference conditions.

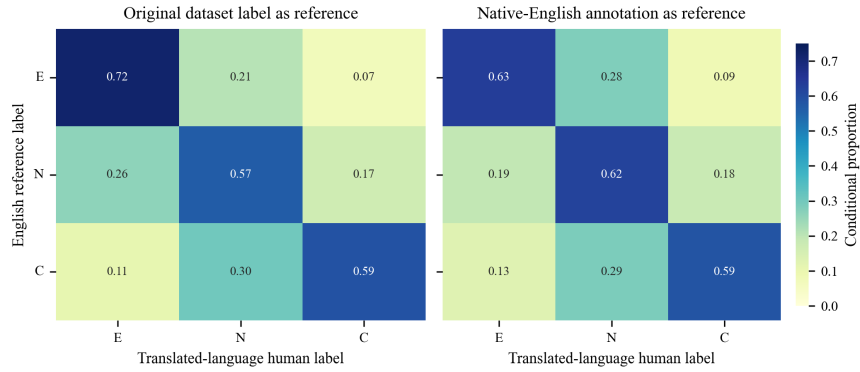


Figure 4: Row-normalized transitions from each English reference to the translated-language human annotation, aggregated across nine languages.

Metric	Mean	Std. Dev.	CV
Original agreement	0.616	0.049	0.080
Original macro recall	0.624	0.045	0.073
EN-human agreement	0.617	0.020	0.033
EN-human macro recall	0.613	0.023	0.037

Table 4: Cross-lingual variability of evaluation metrics.

5.6 Systemic vs. Language-specific Drift

We next examine whether label instability is concentrated in particular target languages or reflects a broader cross-lingual pattern.

Table 4 summarizes variation in exact agreement and macro-averaged recall across languages under both English references. Cross-language variation is modest under both conditions and is smaller when the native-English annotation is used: the coefficient of variation falls from 0.080 to 0.033 for exact agreement and from 0.073 to 0.037 for macro recall. The exact-agreement language rankings under the two references are also weakly correlated (Spearman’s $\rho = 0.033$), which

indicates that the choice of English reference substantially changes the relative ordering of languages. Nevertheless, label instability remains present across all nine translations.

5.7 Model-human Agreement

To investigate whether NLI models assign the same labels as human annotators when evaluated on the translated pairs, we evaluated the outputs of four pretrained NLI models, namely, mDeBERTa-v3-base, XLM-R-large-XNLI, MiniLMv2-L12, and DeBERTa-v3-large². The first three are multilingual while DeBERTa is included as an English-supervised, source-dataset-matched baseline. mDeBERTa is especially valuable as a diagnostic tool because it was fine-tuned on a multilingual NLI mixture derived in part from the same five English benchmarks from which our evaluation sample is drawn. If translation-invariant inference were stable, a model trained under these con-

²Hugging Face model cards: [mDeBERTa](#), [XLM-R](#), [MiniLM](#), and [DeBERTa-v3-large](#).

Language	<i>n</i>	mDeBERTa		XLM-R		MiniLM		DeBERTa (EN-sup.)	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Arabic	982	0.674	0.654	0.441	0.435	0.388	0.376	0.523	0.492
Chinese CN	957	0.582	0.575	0.544	0.533	0.476	0.459	0.539	0.541
German	990	0.664	0.653	0.488	0.486	0.422	0.418	0.687	0.687
Spanish	966	0.649	0.642	0.491	0.493	0.410	0.406	0.684	0.686
Portuguese	976	0.615	0.612	0.510	0.503	0.436	0.428	0.662	0.663
Thai	962	0.616	0.610	0.556	0.536	0.469	0.453	0.361	0.353
Turkish	873	0.640	0.632	0.482	0.462	0.431	0.416	0.528	0.527
Urdu	991	0.580	0.570	0.529	0.518	0.459	0.444	0.410	0.407
Chinese ZH	944	0.583	0.566	0.503	0.488	0.456	0.436	0.543	0.539
Target mean	—	0.623	0.613	0.505	0.495	0.439	0.426	0.549	0.544
English control	987	0.609	0.606	0.562	0.553	0.492	0.491	0.648	0.649

Table 5: Model–human agreement against the independent human annotation for each language. Target mean is the unweighted mean over the nine translations; English is shown separately as a matched monolingual comparison. F1 is macro-averaged over E/N/C. Bold numerical values indicate the highest score in each row for each metric.

ditions should show strong agreement with human annotations of the translated pairs. For each target language, we compare each model’s predicted label for the translated pair against the human annotation assigned to that same pair. *NonSense* and otherwise invalid human labels are excluded, leaving 8,641 target-language pairs. We also report performance against the 987 valid native-English annotations as a matched monolingual comparison.

Table 5 reports exact agreement and macro-F1; its target-language mean gives every language equal weight. The table shows that mDeBERTa achieves the highest mean agreement with human annotations, with 0.623 accuracy and 0.613 macro-F1, outperforming XLM-R (0.505), DeBERTa-v3-large (0.549), and MiniLM (0.439). Its strongest agreement appears in Arabic (0.674), German (0.664), and Spanish (0.649), while Urdu and the two Chinese varieties show the lowest agreement. Notably, against the native-English annotations, mDeBERTa reaches 0.609 agreement, while DeBERTa-v3-large reaches 0.648, XLM-R 0.562, and MiniLM 0.492.

This matched comparison uses the same human reference condition for English and the translated languages rather than mixing original-label English accuracy with translated-human agreement. For the 987 valid English items, the English-supervised DeBERTa agrees with the original dataset labels in 92.2% of cases but with the native-English annotations in 64.8%. Among the 365 cases where the original and native-English labels differ, the model follows the original label in 314 cases (86.0%), the native-English la-

bel in 44 cases (12.1%), and the third label in 7 cases (1.9%). This strong source-label alignment may reflect learned source-dataset conventions, artifacts, or exact item exposure; without source-split provenance, these possibilities cannot be separated. The English-supervised DeBERTa-v3-large model’s English-to-target difference (0.648 versus a 0.549 target mean) needs to be interpreted descriptively rather than as a causal effect of translation or multilingual training.

Figure 5 reveals that a major source of model–human disagreement involves the *Neutral* label. For mDeBERTa, 26.6% of human Entailment and 34.1% of human Contradiction cases are predicted as Neutral, and the English-supervised DeBERTa shows a similar tendency. This mirrors the drift pattern observed in the human annotation analysis for these two models, although the DeBERTa pattern must also be considered in light of its source-label alignment. In contrast, XLM-R and MiniLM frequently predict Entailment for cases annotated as Neutral by humans; only 37.2% and 34.4% of their human Neutral cases, respectively, remain Neutral. The four models therefore do not exhibit a uniform tendency to weaken determinate inference relations toward Neutral.

5.8 Qualitative Evidence

In addition to assigning NLI labels, annotators were asked to document observations about the translated pairs. These notes, cross-referenced with original dataset labels and phenomenon tags, reveal several recurring sources of drift that cut across languages. To begin with, several modal ex-

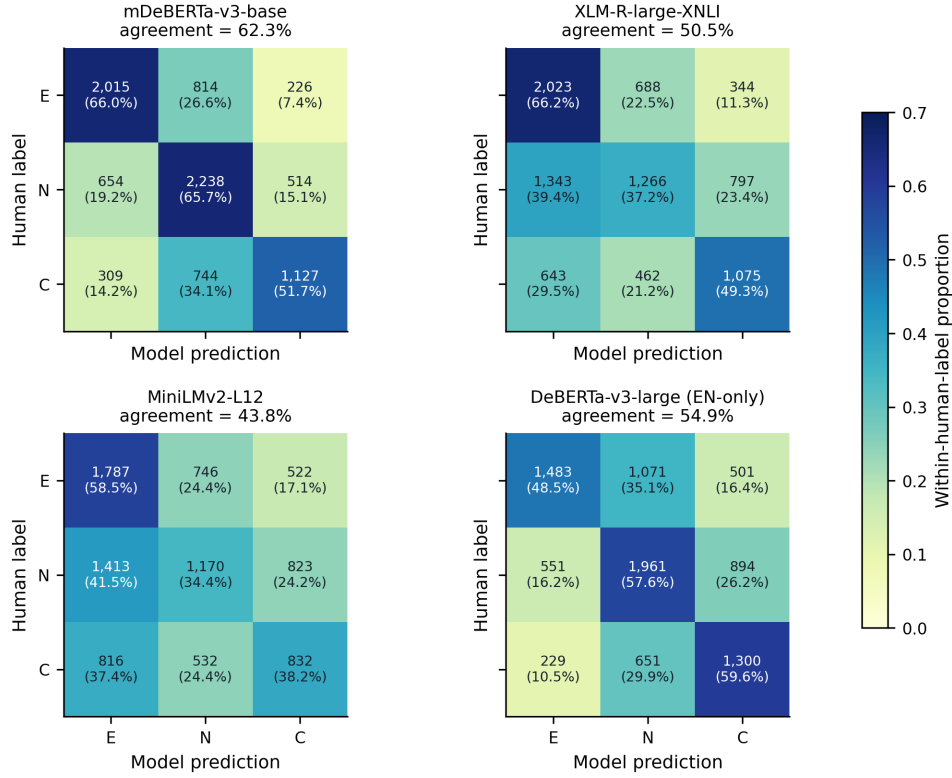


Figure 5: Aggregate confusion matrices over 8,641 target-language pairs.

pressions were found to shift inferential strength after translation. For instance, in some Arabic translations, the weaker English modal *should* was rendered as *يجب أن* (*yajibu an*, meaning “must” or “have to”). This translation makes the hypothesis express a stronger requirement than the English version. Consequently, a premise that supports what *should* occur may not entail what *must* occur, changing the relation from Entailment to Neutral. In other cases, the surface form of a modal negation was preserved after translation whereas its rhetorical force was lost, which led annotators to assign a Contradiction label to pairs that were originally Entailment. In contrast, in Urdu, expressions of possibility, uncertainty, or risk such as *could*, *might*, and *may* were rendered as stronger declarative constructions, which resulted in annotating Neutral pairs as Entailment or Contradiction. A similar pattern was also observed in Turkish. English modals such as *can*, and *could* are typically rendered using the auxiliary Turkish –(y)Abil, which may indicate possibility, ability, and requesting or suggesting something. In addition, due to the extensionality of –(y)Abil structure, the premises and hypotheses that include one of these English modal verbs are annotated in vari-

ous ways unless there are other indicators of the intended meaning.

Another language-specific interpretation has been observed with declarative sentences, which are translated to Turkish by using the generalizing modality marker *-Dir*. Besides being used for definitions and scientific facts, the generalizing modality marker *-Dir* in Turkish may indicate certainty and predictability. As a result of this contradictory nature of the marker, the premises and hypotheses that involve *-Dir* in Turkish showed mismatches in the label assignment.

Further, expressions with conditional constructions exhibited a related pattern across target languages. That is, hypothetical or preventive sentences were reinterpreted as deterministic and definite outcomes. Additionally, in intensional contexts, potential or planned actions were altered into a more concrete and factual sense.

Interrogative constructions were also found to present common challenges. In Thai, for instance, English rhetorical questions were rendered as genuine questions, and tag questions were softened by target-language markers. Both issues changed the inferential weight of the premise and hence shifted labels away from the English reference.

Other contributors to label drift were both register inconsistency and structural divergence. For instance, in Cantonese translations often shifted between Standard Written Chinese (書面語) and everyday spoken Cantonese, with only some items using colloquial markers such as 㗎, 嘅, 㗐, and 係. This register inconsistency sometimes made the premise and hypothesis feel stylistically mismatched. In addition, because Cantonese does not require articles, plural marking, or tense marking, English contrasts such as *the* vs. *a*, singular vs. plural, and past vs. present were often lost or marked inconsistently. As a result, some premise–hypothesis pairs became semantically underspecified, especially when the label depended on co-reference, number, or temporal interpretation. Further, inconsistent transliteration of named entities across premises and hypotheses compounded the problem of assigning labels when co-reference was key in making a decision.

Lexical drift was another factor that led to label instability. In German, for example, there were instances where the translation of polysemous expressions introduced semantic shifts, which changed the inferential relationship between the premises and hypotheses. For instance, the English word *Gemini* was rendered as *Zwilling*, which can also mean *twin*. Similarly, the word *single* was translated as *ledig*, which strictly means *unmarried*, rather than “not in a relationship.” Also, the mapping between the German equivalents of *if* and *when* is not homomorphic. Even though both terms show considerable overlap in usage, they do not translate directly, which introduced conditionality-related shifts in the inference relation.

Furthermore, in Spanish for instance, intensional predicates such as *believe*, *know*, and *try* produced label instability when subtle lexical or aspectual differences altered the interpretation of embedded events. In addition, some comparative constructions became semantically underspecified, especially in cases involving negation or scalar interpretations. Similarly, in Urdu, scalar expressions such as *at least* and *more than* underwent slight semantic shifts that changed the inferential relationship between premise and hypothesis despite substantial lexical overlap.

Taken together, these notes identify recurring candidate sources of label change that appear across several languages and sampling strata. Because each language was judged by one annotator

and the examples were not independently adjudicated, we treat these observations as qualitative hypotheses rather than estimates of how often each mechanism occurs.

Appendix C provide more examples of label drift under translation.

6 Conclusion

This study investigated whether MT preserves NLI inferential relations across nine typologically diverse languages. We sampled 1,000 sentence pairs from five NLI benchmarks, translated them using GPT-4o-mini, and annotated each target language set by a linguistically trained native speaker. A native-English speaker annotated the same source pairs under the same protocol. Each sentence pair contained one of four linguistic phenomena that are known to complicate Entailment judgments. Agreement with the original dataset labels ranges from 55.1% to 67.4%, while agreement with the native-English annotations ranges from 57.8% to 64.2%. Complete nine-language convergence is 16.4% on the 750 items with valid annotations in every translation, while mean pairwise agreement is 62.4%. Entailment and Contradiction judgments are frequently redirected toward Neutral, although the English control shows that some apparent instability is already present in the source items. Conditionality has the highest label-change rate under the native-English reference, while comparatives/quantifier scope has the lowest rate under both references. The model analysis also shows that model–human errors differ by classifier rather than following a single direction.

These results challenge the assumption that synthesizing multilingual NLI datasets by means of machine-translating English benchmarks yields resources that are reliable for cross-lingual evaluation. Future work therefore needs to place greater emphasis on human validation, phenomenon-aware analysis, and alternative dataset construction strategies that do not rely exclusively on projected English labels.

Limitations

First, each language, including English, was annotated by one native speaker. Calibration and spot checks improve protocol consistency but cannot estimate within-language inter-annotator agreement or separate an annotator-specific preference from a language-level effect. Multi-annotator adju-

dication is therefore the highest-priority validation step, and claims in this paper are framed as annotation-set differences rather than population-level judgments for each language.

Second, all translations were generated by GPT-4o-mini without post-editing. The observed changes may reflect this specific LLM-based translation pipeline, and the present study cannot determine whether the same patterns would occur with another MT system or with professional human translation. A controlled comparison using additional systems or a human-translated subset requires new target-language judgments and remains future work.

Third, the corpus intentionally oversamples difficult semantic phenomena. The English reannotation controls for some inherent source-item instability, but a single English annotator cannot fully characterize monolingual ambiguity.

Fourth, our original dataset is sampled from five different datasets, which means five different annotation processes and instruction. Thus, it is likely that the original labels already show label drift.

Finally, the sample may not generalize to all genres or domains represented in larger NLI benchmarks. NonSense labels also combine malformed source items, incoherent translations, and difficult-to-interpret pairs; without adjudication, their causes cannot be cleanly separated.

Acknowledgments

AI assistants were used for limited technical support in formatting L^AT_EX tables and figures and in programming and debugging code for data processing. The authors reviewed and verified the resulting code, analyses, and manuscript changes and retain full responsibility for the work. Further details and the associated reproducibility code are documented in the project repository.³

References

- Ernest Adams. 1965. The logic of conditionals. *Inquiry*, 8(1-4):166–197.
- Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. [IndicXNLI: Evaluating multilingual inference for Indian languages](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10994–11006, Abu Dhabi, United Arab Emirates.

- Željko Agić and Natalie Schluter. 2018. [Baselines and test data for cross-lingual inference](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. [Translation artifacts in cross-lingual transfer learning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online.

- Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, and 1 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

- Emrah Budur, Rıza Özçelik, Tunga Gungor, and Christopher Potts. 2020. [Data and representation for Turkish natural language inference](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8253–8267, Online. Association for Computational Linguistics.

- Tran Bao Bui, Linh Thi-Thuy Nguyen, and Tin Van Huynh. 2025. VietX-NLI: A cross-lingual natural language inference dataset with Vietnamese as the source language. In *2025 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6. IEEE.

- 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*, pages 2475–2485, Brussels, Belgium.

- Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean and Anna Sun Maillard, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Searley Jarrett, and 18 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

- Izumi Haruta, Koji Mineshima, and Daisuke Bekki. 2020. [Combining event semantics and degree semantics for natural language inference](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1758–1764, Barcelona, Spain (Online). International Committee on Computational Linguistics.

- Izumi Haruta, Koji Mineshima, and Daisuke Bekki. 2022. [Implementing natural language inference for comparatives](#). *Journal of Language Modelling*, 10(1):139–191.

³<https://github.com/Muhsabrys/NLI10>

- Maite Heredia, Julen Etxaniz, Muitze Zulaika, Xabier Saralegi, Jeremy Barnes, and Aitor Soroa. 2024. [XNLIeu: A dataset for cross-lingual NLI in Basque](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4177–4188, Mexico City, Mexico.
- Aung Kyaw Htet and Mark Dras. 2025. [Myanmar XNLI: Building a dataset and exploring low-resource approaches to natural language inference with Myanmar](#). *Language Resources & Evaluation*, 59:3267—3310.
- Mohsinul Kabir, Tasnim Ahmed, Md Mezbaur Rahman, Polydoros Giannouris, and Sophia Ananiadou. 2026. [Semantic label drift in cross-cultural translation](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 3714–3724, Palma de Mallorca, Spain.
- Kristina Liefke. 2024. Intensionality and propositionalism. *Annual Review of Linguistics*, 10(1):85–105.
- Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. [WANLI: Worker and AI collaboration for natural language inference dataset creation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Bill MacCartney. 2009. *Natural Language Inference*. Ph.D. thesis, Stanford University.
- Friederike Moltmann. 2020. Truthmaker semantics for natural language: Attitude verbs, modals, and intensional transitive verbs. *Theoretical Linguistics*, 46(3-4):159–200.
- Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. [Adversarial NLI: A new benchmark for natural language understanding](#). In *Proceedings of ACL*, pages 4885–4901. Association for Computational Linguistics.
- Leo Obadić, Andrej Jeršec, Marko Rajnović, and Branimir Dropuljić. 2023. [C-XNLI: Croatian extension of XNLI dataset](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2258–2267, Toronto, Canada. Association for Computational Linguistics.
- Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Agarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. [Does putting a linguist in the loop improve NLU data collection?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Sara Rajaei, Rochelle Choenni, Ekaterina Shutova, and Christof Monz. 2025. [An empirical analysis of machine translation for expanding multilingual benchmarks](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1–30, Suzhou, China. Association for Computational Linguistics.
- Kyle Richardson, Hai Hu, Lawrence S. Moss, and Ashish Sabharwal. 2019. Probing natural language inference models through semantic fragments. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of NAACL-HLT*, pages 1112–1122. Association for Computational Linguistics.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. [Qwen2 technical report](#). *Preprint*, arXiv:2407.10671.

A Preliminary Model Diagnostic

Before collecting the matched human annotations, we tested mDeBERTa-v3-base-xnli-multilingual-nli-2mil7⁴ on 5,000 pairs from an Arabic MT version of LingNLI.⁵ Comparing model predictions with the original English labels yielded 58.14% exact agreement and 57.63 macro-F1, with many errors involving Neutral (as illustrated in Figure 6). This diagnostic motivated the human study but cannot distinguish model error from translation-associated change.

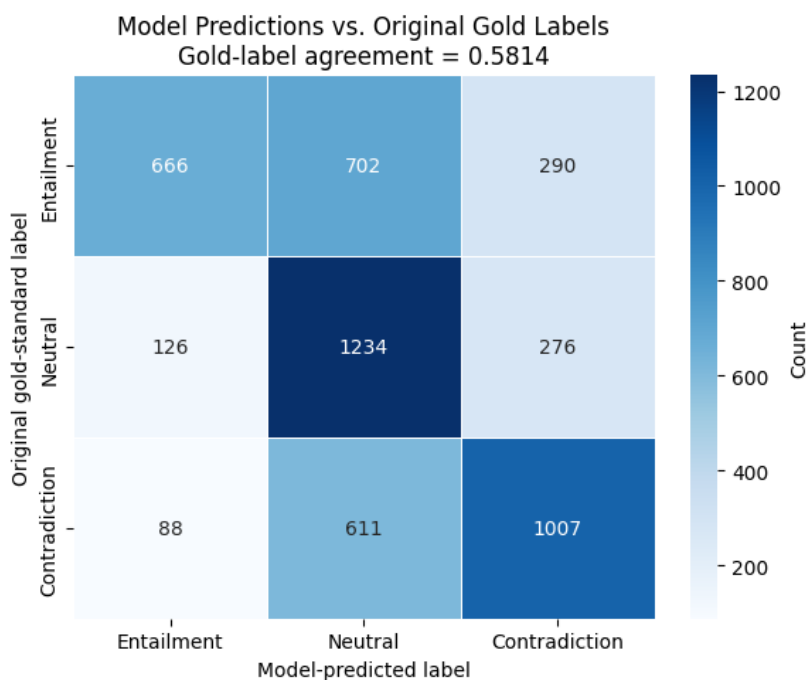


Figure 6: Preliminary Arabic LingNLI diagnostic with mDeBERTa.

⁴<https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7>

⁵<https://huggingface.co/datasets/MoritzLaurer/multilingual-NLI-26lang-2mil7>

B Language-by-Phenomenon Detail

Table 6 reports the target-language count, excluded NonSense count, exact preservation under each English reference, and the most frequent non-stable transition under the native-English reference. The original sampling design contains 100 comparative/quantifier items and 300 items in each other stratum per language.

Lang.	Sampling phenomenon	Valid	NonSense	Original agr.	EN-human agr.	Main change
AR	Comparatives / quantifier scope	100	0	.630	.600	E→N
AR	Conditionality	294	6	.684	.507	E→N
AR	Intensionality	297	3	.717	.577	E→N
AR	Modality	291	9	.636	.675	E→N
CN	Comparatives / quantifier scope	98	2	.653	.724	N→E
CN	Conditionality	276	24	.525	.557	E→N
CN	Intensionality	289	11	.557	.635	N→E
CN	Modality	294	6	.534	.639	N→C
DE	Comparatives / quantifier scope	100	0	.710	.670	C→N
DE	Conditionality	294	6	.687	.588	E→N
DE	Intensionality	297	3	.660	.618	E→N
DE	Modality	299	1	.645	.652	E→N
ES	Comparatives / quantifier scope	89	11	.742	.730	E→N
ES	Conditionality	286	14	.647	.546	E→N
ES	Intensionality	298	2	.607	.639	E→N
ES	Modality	293	7	.614	.641	E→N
PT	Comparatives / quantifier scope	98	2	.612	.612	E→C
PT	Conditionality	285	15	.691	.606	E→N
PT	Intensionality	297	3	.636	.666	N→C
PT	Modality	296	4	.642	.662	N→C
TH	Comparatives / quantifier scope	99	1	.727	.737	N→C
TH	Conditionality	281	19	.584	.598	E→N
TH	Intensionality	291	9	.550	.625	N→E
TH	Modality	291	9	.619	.615	N→C
TR	Comparatives / quantifier scope	90	10	.722	.678	N→E
TR	Conditionality	242	58	.661	.567	E→N
TR	Intensionality	267	33	.581	.643	E→N
TR	Modality	274	26	.675	.631	N→C
UR	Comparatives / quantifier scope	100	0	.630	.670	C→E
UR	Conditionality	294	6	.524	.610	E→N
UR	Intensionality	297	3	.582	.628	E→N
UR	Modality	300	0	.567	.630	N→E
ZH	Comparatives / quantifier scope	93	7	.634	.570	C→N
ZH	Conditionality	284	16	.542	.550	E→N
ZH	Intensionality	292	8	.551	.608	E→N
ZH	Modality	275	25	.553	.577	E→N

Table 6: Detailed language-by-phenomenon results.

C Examples of Translation Label Drift

Table 7 presents three cases in which the original dataset label and the native-English annotation agree, but the translated-language annotator assigns a different label. The examples illustrate lexical, modal-pragmatic, and direct content changes introduced by translation.

Item	English pair	Target-language pair	Labels	Why the label changes
LING2078 (German)	P: A woman of my acquaintance recently announced that she has a boyfriend and wants everyone to introduce the fellow by that title. H: The woman is <i>single</i> .	P: Eine Frau, die ich kenne, hat kürzlich angekündigt, dass sie einen Freund hat und möchte, dass jeder ihn mit diesem Titel vorstellt. H: Die Frau ist <i>ledig</i> .	Original: C; English human: C; German human: E.	English <i>single</i> denotes not being in a relationship here, whereas German <i>ledig</i> specifically denotes being unmarried. The boyfriend therefore no longer contradicts the translated hypothesis, which the German annotator judged as Entailment.
LING2913 (Spanish)	P: How <i>could</i> the Mormons have practiced polygamy? H: The Mormons practiced polygamy.	P: ¿Cómo <i>pudieron</i> los mormones practicar la poligamia? H: Los mormones practicaban la poligamia.	Original: N; English human: N; Spanish human: E.	The Spanish question favors the factual reading “How were they able to practice polygamy?”, which presupposes that the practice occurred more strongly than the English modal question. This supports the translated hypothesis.
Fever20541 (Portuguese)	P (relevant clause): Seattle is the largest <i>city</i> in both the state of Washington and the Pacific Northwest region of North America. H: Seattle is the largest <i>person</i> in the state of Washington.	P (relevant clause): Seattle é a maior <i>cidade</i> do estado de Washington e da região do Noroeste Pacífico da América do Norte. H: Seattle é a maior <i>cidade</i> do estado de Washington.	Original: C; English human: C; Portuguese human: E.	The translation changes the erroneous English noun <i>person</i> to <i>city</i> . The Portuguese hypothesis consequently repeats information stated in the premise, turning a Contradiction into Entailment.

c

Table 7: Three item-level examples of translation-associated label change.