PETra: A Multilingual Corpus of Pragmatic Explicitation in Translation

Doreen Osmelak[1], Koel Dutta Chowdhury[1], Uliana Sentsova[1], Cristina España-Bonet[2,3], Josef van Genabith[1,2]

[1] Saarland University, Saarland Informatics Campus, Germany
[2] German Research Center for Artificial Intelligence (DFKI)
[3] Barcelona Supercomputing Center (BSC-CNS), Barcelona, Catalonia, Spain

Abstract

Translators often enrich texts with background details that make implicit cultural meanings explicit for new audiences. This phenomenon, known as pragmatic explicitation, has been widely discussed in translation theory but rarely modeled computationally. We introduce PETra, the first multilingual corpus and detection framework for pragmatic explicitation. The corpus consists of 3,000 sentence pairs from TED-Multi and Europarl, covers twelve language pairs, and includes additions such as entity descriptions, measurement conversions, and translator remarks. We identify candidates through null alignments and refine them using active learning with human annotation. Our results show that entity and system-level (e.g., metric conversions) explicitations are most frequent, and that active learning improves classifier accuracy by 7-8 percentage points, achieving up to 0.88 accuracy and 0.82 F1 for the best transfer languages. PETra establishes pragmatic explicitation as a measurable, cross-linguistic phenomenon and takes a step towards building culturally aware machine translation.

Keywords: translation, multilingualism, explicitation

1. Introduction

When translating between languages, what is left unsaid in one culture may need to be spelled out in another. Consider an English text that refers simply to *Angela Merkel*. For a German audience, the name alone suffices. Yet a translator into, say, Arabic might render it as *Angela Merkel, former German Chancellor*, adding background knowledge that the target audience may lack. This added phrase does not alter the literal meaning but enriches the translation by making implicit, tacit knowledge in the source explicit in the target.

This process, known as pragmatic explicitation, reflects how translators bridge gaps in shared world knowledge between source and target audiences. The source community may possess contextual or cultural knowledge that is assumed but unstated, while the target audience lacks access to such background. Translators thus introduce minimal but informative cues, such as titles, descriptions, or references, to clarify meaning and preserve communicative intent. Although explicitation has long been recognized in translation studies (Vinay and Darbelnet, 1958; Nida, 1964; Klaudy, 1993; Snell-Hornby, 2006), it remains underexplored in computational research, particularly in terms of pragmatic or culturally motivated explicitations.

Most prior computational studies have focused on structural or discourse-level explicitations, such as the insertion of connectives or referential markers (Hoek et al., 2015; Lapshinova-Koltunski and Hardmeier, 2017). However, pragmatic explicitation differs in that it targets implicit knowledge rather than linguistic form—it aims to make tacit, culture-dependent meaning explicit for a new audience. Despite its communicative and interpretive significance, no large-scale, multilingual, human-validated resource currently exists for systematically studying pragmatic explicitation as a measurable phenomenon.

This paper takes a step towards filling this gap. We present a **multilingual corpus of pragmatic explicitations**, systematically identified and annotated across two large parallel multilingual datasets: Europarl, and TED-Multi. We automatically extract candidate explicitations using *null alignments*, i.e., words appearing only in one side of a translation, and refine them with named-entity recognition and part-of-speech constraints to target additions likely tied to cultural entities, roles, or institutions.

To ensure interpretive quality, we employ a human-in-the-loop annotation framework, combining expert annotation with active learning to iteratively improve coverage and precision. The resulting corpus, **PETra**, is the first multilingual, human-validated dataset for pragmatic explicitation detection. Figure 1 summarizes the overall pipeline. Our contributions are as follows:

- We formalize *pragmatic explicitation* as a computationally tractable phenomenon and re-

Correspondence to: dosmelak, koeldc@lst.uni-saarland.de The corpus and code can be found at https://huggingface.co/datasets/Doosme/PETra and https://github.com/Doosme/PETra

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8756–8766
11–16 May 2026.

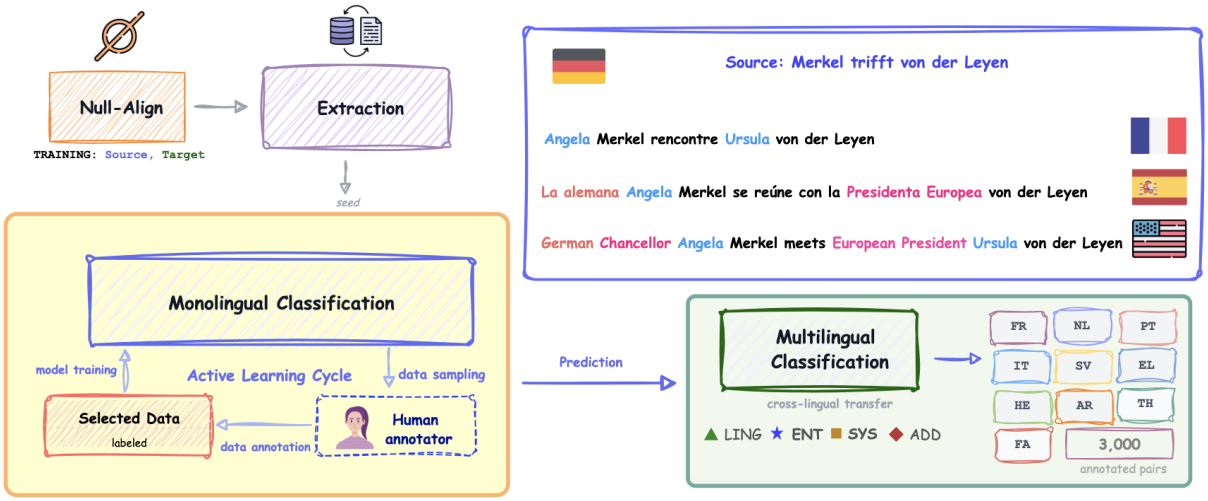


Figure 1: Main Findings. (Top left) Seed data is extracted from parallel corpora via null-alignments and refined through an active learning cycle with a human annotator, training a monolingual classifier. (Top right) Example sentence pairs illustrating pragmatic explicitation. (Bottom) The trained model transfers cross-lingually to predict explicitation labels across four broad categories— *Entities* (★ ENT), *System Transfers* (■ SYS), *Linguistic Adjustments* (▲ LING), and *Added Information* (◆ ADD) across twelve language pairs, resulting in the PETra corpus of 3,000 annotated sentence pairs.

lease PETra, the first multilingual, human-validated corpus for its study.

- We develop an active learning framework integrating linguistic heuristics with human feedback for efficient detection and annotation.
- We analyze explicitation patterns across 12 language pairs and two domains, uncovering systematic cross-linguistic tendencies in how translators encode cultural knowledge.

2. Related Work

Explicitation has long been a core concept in translation studies (Gellerstam, 1986; Baker, 1993; Laviosa, 1998; Pym, 2005; Englund Dimitrova, 2005). Early comparative stylistics framed explicitation as a systematic shift from implicit to explicit expression across languages (Vinay and Darbelnet, 1958). Nida (1964) framed similar additions as amplifications, emphasizing their role in improving readability or avoiding ambiguity derived from socio-cultural differences. Blum-Kulka (1986) conducted the first systematic empirical study, formulating the explicitation hypothesis, which broadly states that translations tend to be more explicit than non-translations. Klaudy (1993) has further distinguished obligatory, optional, pragmatic and translation-inherent explicitations, highlighting that some additions arise specifically to supply cultural or background knowledge. Pragmatic explicitations are particularly important when translators anticipate that certain cultural or contextual information may be unknown to the target readers. Such explicitations go beyond syntactic or stylistic adjustments, aiming instead to bridge knowledge gaps and support effective intercultural communication (Snell-Hornby, 2006). While structural or obligatory explicitations are primarily driven by linguistic constraints, pragmatic explicitations are motivated by the communicative needs of the target audience.

Despite its centrality in translation studies, computational research on pragmatic explicitations remains limited. Early research emphasized discourse-level explicitations, such as the insertion of connectives or cohesive markers, as in the work of Hoek et al. (2015); Lapshinova-Koltunski and Hardmeier (2017); Lapshinova-Koltunski et al. (2020) or on general translationese patterns (Teich, 2003; Baroni and Bernardini, 2006; Volansky et al., 2015; Dutta Chowdhury et al., 2020; Chowdhury et al., 2021). Krüger (2020) investigated explicitation in machine-translated texts, though the focus was primarily on general explicitation phenomena rather than those arising from cultural or contextual asymmetries. Recent research has shown growing interest in integrating cultural and contextual knowledge into multilingual systems. Han et al. (2023) introduce WIKIEXPL, a dataset of concise explicative additions extracted from Wikipedia across languages, and demonstrate their utility in improving downstream tasks such as multilingual question answering. Their work highlights the practical benefits of targeted explicitations, though it focuses on generation for specific domains rather than creating a broadly validated corpus of pragmatic explicitations. Lou and Niehues (2023) similarly proposed semi-automatic techniques for extracting ex-

planations from aligned multilingual data, while Yao et al. (2023) curated culturally specific corpora to enhance cultural grounding in machine translation.

Automated detection of explicitations poses unique challenges: linguistic explicitations can often be identified via syntactic cues or cohesion markers, while pragmatic explicitations depend on world knowledge and cultural context. Furthermore, the rarity of pragmatic explicitations (often <1% of tokens) makes annotation costly and sparse. Active learning has emerged as an effective strategy for handling such imbalances. Following recent advances in active learning for text classification (Wang and Liu, 2023; Li and Guo, 2013), we adapt a multi-label active learning framework to pragmatic explicitation detection. The resulting corpus is the first systematic resource for studying this phenomenon computationally.

3. Formalization of Cultural Explicitations

We define cultural explicitations as fragments of translated text that make implicit cultural or background knowledge explicit to the reader. They go beyond literal or stylistic reformulation by adding information that situates the text within the target audience's cultural frame. Below we outline the main conceptual dimensions; operational definitions are detailed in Section 4.2.

3.1. Conversion of Systems

Texts are embedded in sociocultural systems—political, administrative, educational, and measurement frameworks—that often have no direct equivalents across languages. When translators adapt these constructs, they convey cultural correspondences rather than lexical substitutions.

■ **Measurement systems.** The best example of such cases is probably the differences between customary and metric systems which often require translation choices that add or replace information:

- replacement by conversion (e.g., "1 mile" → "1.6km"; "40 knots" → "75 km / h")
- adding conversions ("120 miles" → "120 miles *(193 km)*")
- adding a unit ("5 degrees" → "5 degrees *C*")
- adding dimension of a measurement ("8 cm" → "8 cm *high*")
- adding approximation terms (e.g., "3 feet" → "*about* one meter")
- replacing figurative use (e.g., "hundreds of miles" → "hundreds of kilometers")

We exclude purely orthographic variants that do not alter meaning (e.g., "lb" → "pound", ; "°" → "degree"; "F" → "Fahrenheit").

■ **Currencies.** Additional cases for currencies include:

- adding a toponym ("$" → "*US* dollar"; "105.000 pounds" → "105.000 *British* pounds")
- currency nicknames ("buck" → "dollar")

■ **Education systems.** Educational systems differ substantially across cultures, reflecting distinct institutional hierarchies, grading scales, and terminologies. Translating between education-related terms often requires interpreting underlying structural or conceptual differences, rather than performing a direct lexical substitution. Typical cases include:

- school names (e.g., "college" → "Universität")
- year names (e.g., "sophomore" → "second year of university")
- grades and exam names (e.g., "A" → "very good")
- elite and support systems (e.g., "Ivy League" → "universities of excellence")

■ **Administrative bodies.** Administrative bodies and political institutions reflect culture-specific governance structures. Transferring between these is not a pure translation, thus replacing or adapting such terms signals an attempt to orient the translation toward the target culture's institutional framework.

- district namings (e.g., German "Bundesland" vs. US "State" and "County" vs. French "Department")
- traffic systems (e.g., "highway" vs. "Autobahn"; "subway" vs. "métro")
- authorities and official institutions (e.g., US "IRS" vs. German "Finanzamt"; US "National Institute of Health" vs. German "Robert Koch Institut")

These principles extend to other societal systems, including political parties, sports leagues, and public agencies, wherever system correspondence is culturally inferred.

3.2. Named Entities

Named entities are often culturally bound and require clarification or adaptation in translation. Cases include:

★ **Replacing entity names.** Widely known entities are often represented through acronyms (e.g., "E.U."). Culturally salient entities are often shortened to colloquial or generic nouns (e.g., "the Wall"). And sometimes entities are commonly referred to by other but similar entities (e.g., "America" for

"U.S.A."). The choice of abbreviation and replacement depend on the background knowledge within the respective culture. Abbreviations might be specific to the source culture (e.g., "NIH", "the Mall"), ambiguous or misleading across cultures ("the Wall" → "the Berlin Wall" vs. "the Western Wall") or in other cases more recognizable to the target audience (e.g., "FBI"), in such cases replacement facilitates understanding. Replacing entities conveys implicit cultural knowledge, highlighting the perceived equivalence between two entities and/or involve audience-oriented cultural mediation.

- replacing entity (e.g., "United Kingdom" → "Great Britain")
- colloquial forms (e.g., "Aussie" → "Australian")
- acronym expansion (e.g., "NIH" → "National Institute for Health")
- acronym collapsing (e.g., "Federal Bureau of Investigation" → "FBI")
- entity nouns (e.g., "the Mall" → "the National Mall"; "the States" → "the United States of America")

★ **Description and specifications.** Entities are often reduced to their identifying core (e.g., "Golden Gate"). Translators may reintroduce descriptive specifications to clarify meaning, as well as replace or augment a term with a functionally similar term of the target culture or descriptive explanations.

- hypernym (e.g., "Golden Gate" → "Golden Gate *Bridge*")
- toponym (e.g., "Cannery Row" → "Cannery Row *(Californie)*")
- ethnonym ("President Carter" → "*American* President Carter")
- matching target entity (e.g., "FDA" → "American *Ministry of Health*")
- explanatory description (e.g., "Ivy League" → "universities of excellence"; "Mayor Bloomberg" → "the Mayor of New York"; "EIA" → "*US energy authority* EIA")

Pure Translations. We explicitly exclude direct translations (e.g., "Germany" → "Deutschland"), nominal component translations (e.g., "Golden Gate Bridge" → "Puente Golden Gate") and quotation marks, as these are pure lexical adaptions that do not add background information. Further, we exclude acronym expansion and acronym collapsing in cases where both entity and acronym are equally familiar across languages (e.g., "U.N." → "United Nations"; "U.S.A." → "United States"), as the choice between them reflects stylistic or linguistic conventions rather than cultural adaptation.

3.3. Terminology

▲ Some lexical items carry culturally loaded meanings or require disambiguation. Translators may add hypernyms ("hybrid" → "hybrid *car*"), substitute hyponyms ("worker" → "commuter"), or expand acronyms ("AI" → "IA *(intelligencia artificial)*"). We exclude simplifications (e.g., "on the Earth" → "on the planet") and idiomatic rephrasing.

Category	Subcategory	Description
★ ENT	ENT-REP	Entity Replacement
	ENT-DESC	Entity Description
	ENT-SPEC	Entity Specification
	ENT-HYP	Entity Hypernym
	ENT-ACR	Entity Acronym
▲ LING	TRANS	Translation
	LING-EXPL	Linguistic Explanation
	ACR	Acronym
	HYPER	Hypernym
	HYPO-SPEC	Hyponym Specification
■ SYS	MEAS-CONV	Measurement Conversion
	MEAS-DIM	Measurement Dimension
	MEAS-SPEC	Measurement Specification
	SYS-CONV	System Conversion
	SYS-DESC	System Description
♦ ADD	ADD-INF	Additional Information
	CLEAR	Clarifying Information
	DEIX	Deixis Resolution

Table 1: Annotation Schema for Pragmatic Explicitation. Each main category is marked by a colored symbol: ★ ENT, ■ SYS, ▲ LING, ♦ ADD.

3.4. Translator remarks

♦ Translators occasionally insert short glosses: notes on wordplay ("sounds like sick brick"), contextual clarifications ("Adlai Stevenson" → "twice Democratic opponent of Eisenhower"), or literal translations ("Eat, Prey, Love" → "Prey = Beute"). Such remarks explicitly mediate cultural understanding. We exclude purely formal or stylistic changes that do not add cultural information; the operational categories are detailed in Section 4.2.

4. Corpus Construction and Annotation Framework

We build a multilingual corpus of **pragmatic explicitations** by combining large-scale parallel data with automatic alignment heuristics and human-in-the-loop annotation. This section describes the extraction pipeline and annotation framework.

4.1. Candidate Extraction via Null Alignments (EXTR)

We draw from two multilingual resources **Europarl** (Koehn, 2005) and **TED-Multi** (Ye et al., 2018). We align each bilingual sentence pair forward-directionally using `eflomal` (Östling and Tiedemann, 2016)[1] and `SimAlign` (Jalili Sabet et al.,

[1] https://github.com/robertostling/eflomal

2020)[2]. Tokens unaligned in the target are marked as *additions*, potential cases of explicitation.

We use pre-trained spaCy pipelines to perform POS-tagging and NER, as well as compound decomposing in the case of German. We extract sentence pairs containing both a named entity (PERSON, NORP, FAC, ORG, GPE, LOC, PRODUCT, EVENT, WORK_OF_ART, LAW, or LANGUAGE) and an unaligned content word (NOUN, PRON, or PROPN) in the target text, then annotate manually.

4.2. Annotation Design

For the Active Learning (AL), annotators classify candidates as:

- **TRUE:** background or cultural enrichment.
- **FALSE:** direct or figurative translation without background or cultural enrichment.
- **DISCARD:** mistranslation, or non-aligned translations.

Annotation guidelines include positive/negative examples for each language. For the corpus, we annotate each detected explicitation with respect to its **type** (what kind of information is added or made explicit) and its **style** (how this information is integrated into the translation). Table 1 provides an overview of the full annotation schema.

Type of Explicitations

We distinguish four broad categories: *Entities* (★ ENT), *System Transfers* (■ SYS), *Linguistic Adjustments* (▲ LING), and *Added Information* (♦ ADD).

***Entities* (★ ENT)** This category covers modifications of named entities that make implicit cultural knowledge explicit.

ENT-REP: *Entity replacement* — replacing an entity with another entity-term referring to the same referent (e.g., "U.K." → "Großbritannien")

ENT-DESC: *Entity description* — describing what an entity is, instead of or in addition to matching it to a target-culture entity (e.g., "EPA" → "die amerikanische Umweltschutzbehörde").

ENT-SPEC: *Entity specification* — adding a toponym or other identifying element (e.g., "Tucson" → "Tucson (Arizona)"; "Library of Congress" → "La Bibliothèque du Congrès (USA)").

ENT-HYP: *Entity hypernym* — adding a descriptive hypernym (e.g., "Golden Gate" → "Golden Gate Brücke"; "Oklahoma" → "das Musical Oklahoma").

ENT-ACR: *Entity acronym* — adding, clarifying, or expanding an acronym (e.g., "NYU" → "NYU (New York University)") or an abbreviated form (e.g., "the Mall" → "the National Mall").

***System Transfers* (■ SYS)** This category covers conversions between political, administrative, or measurement systems that differ across cultures.

MEAS-CONV: *Measurement conversion* — conversion of a measurement from the system used in the source language culture to the system used in the target language culture (e.g., "1 mile" → "1.6 km").

MEAS-DIM: *Measurement dimension* — adding dimensional information to the measurement (e.g., "3 feet" → "1m *hoch*"; "at 60 degrees" → "at *a temperature of* 60 degrees")

MEAS-SPEC: *Measurement specification* — specifying an implicit unit or scale (e.g., "5 degree" → "5 degree *Celsius*")

SYS-CONV: *System conversion* — conversion of societal, political or administrative systems of the source language culture to the one of the target language culture (e.g., "high school" → "Gymnasium")

SYS-DESC: *System description* — explanatory transfers of system terms (e.g., "freshman year" → "first year of university")

***Linguistic Adjustments* (▲ LING)** This category includes linguistic clarifications, terminological adjustments, and translator remarks that add explanatory information.

TRANS: *Translation addition* — adding an additional translation of a term or phrase (e.g., "Infinity Mushroom" → "Infinity Mushroom *(Unendlichkeitspilz)*"). This also includes adding an original Latin spelling of a name in case of differing scripts.

LING-EXPL: *Linguistic explanation* — Remarks by the translator about the translation (e.g., "loosely translated"; "indirect speech"; "transferred to our system"), linguistic remarks such as explanations of an idiom or pun (e.g., "sounds like sick brick").

HYPER: *Hypernym* — adding hypernyms (resp. broader terms) to a noun, pronoun or entity, and/or other clarifying elements (e.g., "hybrid" → "hybrid *car*")

HYPO-SPEC: *Hyponym specification* — replacing a term by a more specific term (e.g., "worker" → "commuter"), or adding a specifying term to it (e.g., "investments" → "*health* investments")

ACR: *Acronym* — adding an acronym or expansion of an acronym (e.g., "AI" → "IA *(intelligencia artificial)*")

[2] https://github.com/cisnlp/simalign

	Pairs			Types				
	P	E	T	★	■	▲	♦	Σ
TED-multi								
DE	534	105	162	146	708	177	41	1072
ES	372	47	63	161	352	49	8	570
FR	280	57	–	134	154	134	26	448
PT	211	55	–	85	104	87	33	309
NL	180	55	–	74	127	64	20	285
IT	123	47	–	56	84	49	16	205
FA	182	–	–	99	14	24	63	200
HE	109	–	–	29	35	29	30	123
SV	32	48	–	23	41	13	16	93
EL	13	59	–	38	22	13	9	82
TH	97	–	–	24	58	27	17	126
AR	65	–	–	23	10	30	17	80
Europarl								
DE_{ep}	52	34	–	100	13	48	11	172
Total				**992**	**1722**	**744**	**296**	**3754**

Table 2: Corpus statistics for PETra. Symbols match the type legend: ★ ENT, ■ SYS, ▲ LING, ♦ ADD. P=pool (classifier), E=extracted (null-alignment), T=active learning. Types are counted for individual instances of explicitation.

***Added Information* (♦ ADD)** This residual category captures all remaining translator remarks that transport additional information.

ADD-INF: *Additional information* — Adding additional (background) information that is not transported in the source text. Such as specifying time of certain events (e.g., "on the 5th of September" → "Am 5. September *[2012]*"), description of what a word means (e.g., "sarong" → "sarong *(traditional Malaysian clothing)*"), additional information about an entity (e.g., ; "Myanmar" → "Myanmar *(former Birma)*"), and more.

CLEAR: *clarifying information* — Adding discourse-related context information, that is not contained in the source text, such as replacing a pronoun by its referee (e.g., "go see him" → "go see him *[Dalai Lama]*"). This can be information known to the source audience, for example due to visual input, or prior discourse input.

DEIX: *deixis resolution* — resolving deixic references known to the audience (e.g. "here" → "in the US").

OTHER: any other type of cultural explicitation.

Style of Explicitations

Each explicitation is additionally annotated for its integration style:

R: *replace* – replacing terms in the source by terms in the target (e.g., "1 mile" → "*1.6km*").

A: *add* – adding terms in the target additionally to the content of the source text (e.g., "1 mile" → "1 mile *(1.6km)*").

Resulting Resource

The resulting resource contains sentence pairs that show cultural explicitation in the translation. Table 3 shows some examples. Each pair contains:

ID: a unique identifier of the sentence pair, containing the language combination.

Source / **Target**: the source language text and target language text. Brackets indicate parts of the text that were added or replaced by cultural explicitation in the translation.

Type: indicated by colored symbols in the target text (see §4.2).

Source: whether the pair was extracted via null-alignments (EXTR), detected by the classifier (POOL), or annotated during active learning (TRAIN).

The final resource consists of instances extracted by null-alignment extraction EXTR, plus further instances found during active learning and instances detected by the final classifier CLF.

Table 2 summarizes the quantitative profile of the PETra corpus[3]. The corpus covers 12 typologically diverse languages and includes both high-resource (e.g., English–German) and medium-resource pairs (e.g., English–Portuguese, English–Arabic), enabling cross-linguistic evaluation of explicitation behavior. Of all candidate sentence pairs, 6% (2.6% for TED, 42% for Europarl[4]) were extracted by EXTR. 3% of the sentence pairs (2.5% for TED, 7.7% for Europarl) were classified as containing cultural explicitation by CLF. Approximately 3,000 instances were manually labeled across the full dataset.

The types of explicitations vary across languages and domains. While on Europarl many explicitations concern specifications and hypernym additions to entities, TED shows a broader range, including linguistic explanations, descriptions and explanations of words and introduction of additional background information. German and Spanish show a strong prevalence of system transfers, while Hebrew and Arabic for example prefer descriptions and translations of entities and concepts.

5. Experiment Settings

We adopt a multi-round, pool-based active learning framework inspired by Wang and Liu (2023),

[3]Although more named-entity instances are extracted in Europarl, they are closely tied to the source text, often reflecting simple expansions or stylistic rephrasings. In contrast, TED translations exhibit more context-driven adaptation, with explicitations that provide additional background, explanations, or cultural information for the target audience.

[4]for TED on DE, ES, IT, PT, FR, NL, EL, SV, HE, AR, FA, TH and for Europarl on DE, ES, IT, PT, FR

Lang	Source (EN)	Target	
en → de	**[The EPA]** estimates, in the United States, by volume, this material occupies 25 percent of our landfills.	Die amerikanische Umweltschutzbehörde ★schätzt, dass in den USA dieses Material, dem Volumen nach, 25% unserer Mülldeponie ausmacht.	P
en → de	So, that's happening in the **[U.K.]** with **[U.K.]** government data.	Das passiert also in Großbritannien ★mit britischen ★Regierungsdaten.	T
en → de	So basically, China is a SICK BRIC country.	China ist also ein SICK-BRIC-Land. klingt wie "kranker Ziegel" ▲	P
en → de	And this bear swam out to that seal — **[800 lb.]** bearded seal — grabbed it, swam back and ate it.	Und dieser Bär schwamm zu dieser Robbe hin — eine 350 Kilo ■ schwere ■, bärtige Robbe — schnappte sie, schwamm zurück und aß sie.	P
en → de	You run up to **[15,000 feet]**, descend **[3,000 feet]**.	Dann rennt man auf eine Höhe von ■ 4.500 m ■hinauf, und wieder 900 m ■bergabwärts.	E
en → de	It's a classic tale of Eat, Prey, Love.	Die klassische Mär von 'Eat, Prey, Love'. Prey = Beute ▲	P
en → nl	So this is a Yellowstone, you know, of Saturn.	Dit is een Yellowstone (natuurpark in de VS ★ bekend van de geisers ♦), van Saturnus.	E
en → nl	You must have gotten your education **[here]**.	U bent vast opgeleid in de VS ♦ en niet in India ♦.	E
en → es	The **[seniors and juniors]** are driving the **[freshmen and the sophomores]**.	Los alumnos del último grado ■llevan a los de primero ■.	P
en → es	They haven't learned that in Monterey.	Eso no lo han aprendido en Monterey (California, EE.UU. ★; sede de la conferencia ♦).	P

Table 3: Examples of pragmatic explicitation in TED-Multi. Colored symbols mark the category: ★ ENT (entity), ■ SYS (system), ▲ LING (linguistic), ♦ ADD (added info). Source: P=pool (classifier), E=extracted (null-alignment), T=active learning.

tailored to multi-label explicitation classification. The process iteratively refines the labeled dataset through uncertainty-driven human feedback. Our annotated seed set forms the initial training data for the active learner. Subsequent candidate samples for annotation are drawn automatically from the unlabeled pool using hybrid query strategies.

5.1. Active Learning Workflow (CLF)

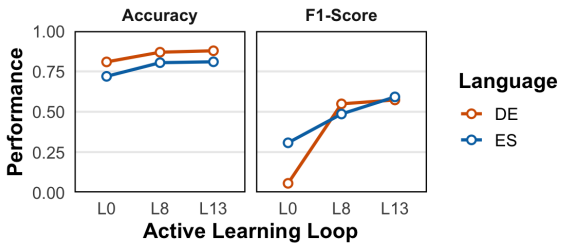
Data Splits. We randomly sample a set of 400 annotated examples from the extracted data, and split it into 100 samples **seed data** and 300 samples **test data**. We ensure 1/3 of the seed data to be positive (explicitation) and 2/3 negative (non-explicitation) cases. The test set remains fixed and unseen to the classifier throughout, serving only for final evaluation. The remaining samples serve as our training **pool**.

Model. We fine-tune `bert-base-multilingual-cased` for binary classification. Each instance consists of a sentence pair separated by the `[SEP]` token. Training uses the `Transformers` library with default settings for 10 epochs. A base classifier is trained on the seed set described above.

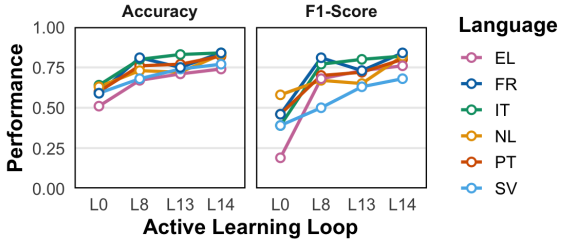
Query Strategies. As pragmatic explicitations are sparse and diverse, we design a two-phase querying approach combining diversity-based exploration and uncertainty-based refinement. Each AL cycle queries 100 new instances, distributed across 2–3 of the strategies below following a heuristic sequence to balance diversity.

Combined Querying (8 iterations). We alternate between the following strategies:

- **High-Confidence Positives:** selects high-probability positives based on lexical, POS, and alignment features, with random sampling above a fixed confidence threshold.
- **Embedding Clustering:** identifies unexplored regions via mini-batch k-means on sentence embeddings; one centroid per cluster is selected.
- **Diverse Seed Expansion:** expands around distant seed positives by retrieving diverse, nearby unlabeled instances in embedding space.
- **Nearest Positive Neighbors:** retrieves unlabeled instances closest (by cosine distance) to current positives; expanded to all positives over iterations.



(a) Monolingual Evaluation



(b) Cross-Lingual Evaluation

Figure 2: Classifier accuracy and F1 at successive active learning stages (L0–L14), averaged over 5 random seeds. L0: before active learning, L8: after 8 cycles of combined strategies, L13: after additional 5 cycles of uncertainty-only strategies, 14: after adding additional 1000 samples of the German and Spanish pools.

- **Low-Confidence Sampling:** selects uncertain cases with lowest model confidence.

Uncertainty Sampling (5 iterations). After coverage stabilizes, we switch to uncertainty-driven sampling by selecting instances with posterior probabilities closest to 0.5 for the positive label.

Annotation and Iteration. Queried samples are manually annotated using the pragmatic explicitation schema (§4.2), merged with existing labels, and used to re-train the model. We run 13 AL cycles in two phases. In the first phase (L1–L8), combined querying (100 samples per cycle) maximizes coverage of the explicitation space. We switch to uncertainty-only sampling for the second phase (L9–L13; 50 samples per cycle) once diversity-based strategies yield diminishing returns. Finally, we construct an extended training set (L14) by adding 1,000 manually annotated instances from the German and Spanish pools, as these provide a more diverse range of explicitation types (Table 2), strengthening generalization to rarer categories before cross-lingual transfer.

6. Results and Discussion

We assess four versions of the classifier: trained on German and Spanish seed data only (L0), seed and data obtained from combined querying strategies (L8), seed and all data from active learning (L13), and seed and all data obtained during active learning plus additional 1000 annotated instances from German and Spanish pool (L14).

Monolingual Evaluation. Figure 2a shows the performance of the Spanish and German classifiers at three stages of AL, as indicated on the x-axis. Active learning improves performance by roughly 7–8 percentage points, yielding approximately 1,150 high-confidence labeled examples for EN–DE and EN–ES. These examples are then used for large-scale automatic annotation of TED-Multi and Europarl, followed by manual spot checks.

Cross-Lingual Evaluation. We further train a cross-lingual classifier on the combined German and Spanish examples and evaluate it on 6 languages: Portuguese, Italian, French, Dutch, Swedish and Greek. The test sets for these languages each contain 100 annotated instances extracted via null-alignment methods[5].

Figure 2b shows that cross-lingual transfer improves steadily across all languages. Accuracy and F1 scores are lowest at L0 and increase with each stage of active learning. The training languages, German and Spanish, start with relatively high scores and continue to improve, with German reaching 0.88 accuracy and 0.57 F1, and Spanish 0.81 accuracy and 0.59 F1 in L13. Languages close to the training data, such as Portuguese (similar to Spanish) and Dutch (similar to German), consistently achieve high scores, with L14 reaching 0.82–0.83 in accuracy and 0.80–0.82 in F1. Italian and French also perform well, showing strong cross-lingual transfer, while Swedish shows moderate gains. Greek, being more distant typologically, exhibits lower scores throughout, with accuracy peaking at 0.74 and F1 at 0.76 in L14.

Analysis. The results from the PETra experiments demonstrate that pragmatic explicitation is a consistent and measurable phenomenon across languages. Table 2 shows that entity-related and system level explicitations occur most frequently, reflecting translators' efforts to make culturally bound references and institutional systems more accessible to target readers. Linguistic adjustments (▲ LING) and additional-information cases (♦ ADD) appear less often but capture meaningful contextual enrichments, such as translator remarks and clarifying glosses.

[5]These datasets only contain very clear cases of explicitations, while the German and Spanish test sets contain also more subtle cases, due to the proficiency of the annotator.

Active learning proved essential for detecting these relatively rare phenomena. As illustrated in Figure 2a, classifier performance improved by about seven to eight percentage points over successive annotation cycles, resulting in roughly 1150 high-confidence labeled examples for English–German and English–Spanish each. Figure 2b further shows that classifier performance steadily improves across successive annotation cycles. Two clear trends emerge: (i) performance improves consistently with more annotated data, and (ii) typological similarity to the seed languages strongly influences transfer success, with L14 achieving the best results across all languages. Table 3 illustrates the qualitative range of the corpus, spanning entity descriptions, measurement conversions, system transfers, and translator glosses across three target languages. The diversity of examples confirms that PETra captures interpretable instances of culturally motivated explicitation.

Typological Patterns. While all languages in PETRA exhibit a wide range of pragmatic explicitations, we observe systematic differences in how they occur across language families. For instance, the category ■ MEAS-DIM (adding dimension information to measurements) is particularly common in Germanic languages compared to Romance and Semitic languages, showing that dimensional information may carry higher communicative salience for speakers of those languages.

European languages usually convert customary units to metric ones, while Semitic languages often keep the original measurement systems, reflecting broader cultural conventions around measurement and localization. On the other hand, German translations tend to assume more cultural background knowledge, using fewer explicitation strategies overall, while Semitic languages more often add descriptive explanations of concepts, entities, or linguistic forms (for example, explicitly stating that a term is adjectival, as in “Aristotelian science”).

7. Conclusion

This paper introduces PETra, the first multilingual corpus and detection framework for pragmatic explicitation. Through a combination of automatic extraction, active learning, and human annotation, we created a resource that captures how translators bridge world-knowledge gaps across languages and domains. The results demonstrate that pragmatic explicitation is both detectable and consistent across typologically diverse language pairs. Our active-learning framework improved classifier accuracy by approximately eight percentage points and achieved strong cross-lingual performance, confirming that such additions follow recognizable patterns of cultural mediation.

Future work will expand the dataset to additional languages, explore context-aware and multimodal signals, and investigate applications in machine translation and cultural adaptation modeling. By documenting pragmatic explicitations at scale, PETra opens a new empirical direction for studying how translators make culture explicit, and represents a first step toward more culturally sensitive computational translation systems.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1102 Information Density and Linguistic Encoding. CEB acknowledges her AI4S fellowship within the “Generación D” initiative by Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by NextGenerationEU through PRTR.

Limitations

While PETra covers a diverse set of language pairs, including both European and non-European languages, it still reflects the availability of existing parallel corpora and therefore remains biased toward high-resource languages and formal domains such as TED talks and parliamentary proceedings. As a result, the observed explicitation patterns may not generalize to low-resource, oral, or literary contexts. The automatic extraction methods capture surface-level additions effectively but may overlook more implicit pragmatic or cultural adaptations. When used for analyzing multilingual large language models, the dataset can reveal systematic differences in cultural representation, yet it cannot isolate the underlying causes of such differences (e.g., training data imbalance or alignment procedures). Consequently, results should be interpreted as diagnostic indicators rather than exhaustive measures of cross-cultural representation.

Bibliographical References

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and technology: In honour of John Sinclair*, pages 233–250.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3).

Shoshana Blum-Kulka. 1986. Shoshana blum-kulka: Explicitation and translation. *Target*. (classic work discussing explicitation hypothesis).

Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2021. Tracing source language interference in translation with graph-isomorphism measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*.

Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2020. Understanding translationese in multi-view embedding spaces. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 6056–6062.

Birgitta Englund Dimitrova. 2005. Expertise and explicitation in the translation process.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english translation studies in scandinavia. *Lund: CWK Gleerup*.

HyoJung Han, Jordan Boyd-Graber, and Marine Carpuat. 2023. Bridging background knowledge gaps in translation with automatic explicitation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9718–9735, Singapore. Association for Computational Linguistics.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted JM Sanders. 2015. The role of expectedness in the implicitation and explicitation of discourse relations. In *Proceedings of the second workshop on discourse in machine translation*, pages 41–46.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1627–1643, Online. Association for Computational Linguistics.

Kinga Klaudy. 1993. Explicitation in translation. In *Translation and Meaning*. Routledge. (discusses obligatory/optional/pragmatic/translation-inherent explicitation).

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

Ralph Krüger. 2020. Explicitation in neural machine translation. *Across Languages and Cultures*, 21(2):195–216.

Ekaterina Lapshinova-Koltunski and Christian Hardmeier. 2017. Discovery of discourse-related language contrasts through alignment discrepancies in english-german translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 73–81.

Ekaterina Lapshinova-Koltunski, Marie-Pauline Krielke, and Christian Hardmeier. 2020. Coreference strategies in english-german translation. In *3rd Workshop on Computational Models of Reference, Anaphora and Coreference, December 2020, Barcelona, Spain (online)*, pages 139–153.

Sara Laviosa. 1998. Universals of translation. *The Routledge Encyclopedia of Translation Studies*, pages 288–291.

Xin Li and Yuhong Guo. 2013. Active learning with multi-label svm classification. In *IjCAI*, volume 13, pages 1479–1485.

Renhan Lou and Jan Niehues. 2023. Audience-specific explanations for machine translation. *arXiv preprint arXiv:2309.12998*.

E Nida. 1964. Toward a science of translating. with special reference to principles and procedures involved in bible translating text.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Anthony Pym. 2005. Explaining explicitation. *Beyond descriptive translation studies*, pages 29–44.

Mary Snell-Hornby. 2006. The turns of translation studies. *New Paradigsms or Shifting Viewpoints*.

Elke Teich. 2003. *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*, volume 5. Walter de Gruyter.

Jean-Paul Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais: méthode de traduction*. Didier.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Mengqi Wang and Ming Liu. 2023. An empirical study on active learning for multi-label text classification. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 94–102, Dubrovnik, Croatia. Association for Computational Linguistics.

B Yao, M Jiang, D Yang, and J Hu. 2023. Empowering llm-based machine translation with cultural awareness. arxiv. *Preprint posted online on May*, 23.

Qi Ye et al. 2018. Word-level quality estimation for machine translation. In *Proceedings of WMT*.