Dual-Metric Compliance and Quality Evaluation of Knowledge Graph Mediated Translation in Regulated Domains: An Enhanced Architectural Framework

Viveta Gene
Ionian University
viveta.gene@gmail.com

Vilelmini Sosoni
Ionian University
sosoni@ionio.gr

This paper presents the second phase of an empirical research programme evaluating knowledge graph-mediated translation (KGMT) in the regulated dermocosmetics domain. Building on Gene and Sosoni (2026), which introduced KGMT as a theoretical framework for compliance-aware multilingual communication, Phase 2 advances the research through three coordinated contributions: an indexed source design with 30 controlled compliance errors drawn from EU and US regulatory frameworks; a non-compliance in source (NCS) classification that principally distinguishes source-propagated from translation-introduced errors; and an enhanced KGMT pipeline incorporating a domain-specific style guide, a lemmatization layer, and an expanded multilingual glossary. Translations of a 500-word retinol product description into Greek and French are evaluated across three conditions—pure baseline (PB), document-augmented (DA), and KGMT—using the Post-Editing Monitoring, Assessment, Problem-Solving (PE.MAPS) dual-metric framework (Gene, 2024). KGMT achieves 100% source compliance detection recall against 0% for both baselines, and records zero penalized translation errors in both target languages under the NCS classification. Inter-condition divergence in baseline outputs from the same prompt on different inference instances demonstrates that large language models (LLMs) do not reliably or auditably guarantee compliance consistency, positioning KGMT as a deterministic, knowledge-grounded stabilization layer.

1 Introduction

Translation in regulated industries operates under a dual constraint: outputs must be linguistically accurate and simultaneously compliant with the regulatory frameworks of the target market. When source content is authored under the United States FDA guidelines for cosmetics and must be adapted for the EU—subject to Regulation 1223/2009 on cosmetic products and Regulation 655/2013 on common criteria for claims—the challenges extend beyond language. Claims permissible under one regime may constitute prohibited drug-territory language under the other (Prieto Ramos, 2014), and ingredient concentrations legal in the US market may exceed European limits set by Commission Regulation 2024/996.

LLMs, when deployed without external knowledge structures, do not reliably or auditably provide compliance guarantees in regulated translation contexts (Pan et al., 2024). While recent architectures incorporating tool access, retrieval-augmented generation (Lewis et al., 2020), or constrained decoding can partially support terminology control, they do not provide the explicit, traceable, regulation-anchored knowledge structures required where audit trails and provenance are legal requirements. The distinctive contribution of KGMT is precisely this: structured, explicit, jurisdiction-aware knowledge encoding producing deterministic, auditable outputs grounded in identified regulatory instruments (Hogan et al., 2021).

The intellectual lineage of this approach connects to foundational work at the intersection of terminology and knowledge engineering, notably Skuce and Meyer's knowledge-based term bank COGNITERM (Skuce and Meyer, 1990), which demonstrated that structured terminological knowledge could support domain-specific language processing. KGMT extends this tradition into the era of neural generation, adding compliance validation and cross-jurisdictional awareness as first-class concerns. Gene and Sosoni (2026) established the theoretical framework and a Phase 1 pilot, which revealed a detection-adoption gap: compliance flags

generated by the knowledge graph (KG) layer were not consistently integrated into the translation output. Phase 2, reported here, addresses this gap through an indexed source design, an NCS classification, and an enhanced KGMT pipeline.

Section 2 reviews the Phase 1 findings and defines compliance-aware translation. Section 3 describes the methodology, including the indexed source design, NCS classification, pipeline enhancements, and annotation protocol. Section 4 reports the empirical results. Section 5 discusses the NCS framework as a reusable contribution and the infrastructure compounding effect. Section 6 addresses limitations, and Section 7 concludes.

2 From Framework to Empirical Validation

A knowledge graph is a graph-structured data resource in which nodes represent real-world entities and edges encode the relationships between them, with the aim of accumulating and conveying domain knowledge (Hogan et al., 2021; see Figure C1, Appendix C). KGMT operationalises this structure for translation by injecting KG-derived terminological mappings, compliance constraints, and disambiguation rules into the LLM prompt at inference time. The Phase 1 theoretical framework and pilot evaluation are reported in Gene and Sosoni (2026); Section 2.1 summarises the key findings relevant to the present study.

2.1 Phase 1 and the Detection-Adoption Gap

Gene and Sosoni (2026) introduced KGMT as a grounded prompting architecture in which KG-derived terminological guidance, compliance constraints, and disambiguation rules are injected into the LLM prompt at inference time, following retrieval-augmented generation paradigms (Lewis et al., 2020; Pan et al., 2024). The Phase 1 evaluation translated a 500-word retinol product description into Greek and French, comparing KGMT against PB and DA conditions using GPT-5 (OpenAI, 2025). KGMT achieved a source compliance detection recall of 24% and a translation quality score (TQS) of 83.9% (Greek) and 88.2% (French), below both baselines. The detection-adoption gap was identified as the primary research priority: the KG was functioning as a source-side compliance checker but its outputs were not consistently adopted by the generation layer.

2.2 Defining Compliance-Aware Translation

A gap in Phase 1 was the absence of an explicit task definition. Compliance-aware translation, as operationalized here, refers to target-language content that is simultaneously (a) semantically faithful to the source, (b) terminologically consistent with domain glossaries (Bowker, 2002; Moorkens et al., 2018), and (c) compliant with the regulatory frameworks of the target jurisdiction. These three dimensions correspond respectively to accuracy, fluency, and compliance in the PE.MAPS framework (Gene, 2024). Compliance-aware translation must also perform source-side auditing: identifying and flagging violations before they propagate into the target. This is distinct from standard translation quality assessment and requires the NCS classification introduced in Section 3.2.

3 Methodology

3.1 Source Text, Indexed Error Design, and Classification Criteria

The source text is a 500-word retinol night cream product description representing a US market product subject to FDA cosmetics guidelines, with 30 compliance errors embedded and indexed [1]–[30] using bracketed notation, plus one unindexed C8 warning-omission error (31 evaluable violations total). Each error was classified against explicit EU regulatory criteria before indexing: C1 violations (drug claim, n=10) were assessed against Regulation 655/2013 Article 4 and Annex I criteria 2 and 6; C3 violations (unsubstantiated superlative, n=19) against Annex I criterion 4 requiring substantiation for comparative or absolute claims; C2 (prohibited concentration, n=1) against Regulation 2024/996 Annex III (retinol RE limit 0.3% for face leave-on products); C5 mismatches (n=2) against study data in the source; and C9 missing evidence (n=3) against Annex I criterion 3 requiring substantiation for efficacy claims. Classifications were established prior to annotation and provided to both annotators as decision criteria, ensuring the ground truth is theoretically grounded rather than post-hoc. Annotation was performed by two evaluators with professional regulatory translation expertise; full details of the cross-annotation protocol and inter-annotator agreement are reported in Section 3.5. The indexing design provides an objective

detection recall ground truth independent of annotator judgement.

The human reference translation was produced by the first author, a professional translator with regulatory specialization, working from the source text and the full KG knowledge base (glossary, compliance rules, style guide) but without access to any system output. This mirrors the conditions of a KG-informed human translation and provides a ceiling benchmark for the automated conditions. Producing the reference prior to and independently of the system outputs prevents anchoring bias in the annotation.

3.2 NCS Classification

The Non-Compliance in Source (NCS) classification addresses a methodological problem identified in Phase 1 and in prior quality assessment frameworks including multidimensional quality metrics (MQM) (Lommel et al., 2014) and TAUS DQF: the conflation of source-originated with translation-introduced compliance violations. This conflation penalizes translation systems for failures that belong to the content creation process, not the translation process. Table 1 defines the three NCS states and their penalty logic. Under NCS, KGMT is penalized only for TRANS_INTRODUCED errors; SRC_FLAGGED instances are recorded for detection recall analysis but do not reduce TQS.

NCS State	Definition	Penalty	Example (this study)
TRANS_INTRODUCED	Source compliant; translation introduces the violation	Penalized	PB/DA: 'stimulates collagen' not mitigated
SRC_REPRODUCED	Source non-compliant; translation reproduces without flagging	Penalized (QA failure)	PB/DA: 1.0% retinol reproduced, no flag raised
SRC_FLAGGED	Source non-compliant; translation flags and mitigates	Not penalized	KGMT: 1.0% retinol → NCS-C2 flag + compliant alternative

Table 1: NCS classification states, penalty logic, and corpus examples.

3.3 KGMT Pipeline Enhancement

Figure 1 illustrates the KGMT four-stage pipeline. Phase 2 injects three enhancements as a package into the KG retrieval and prompt-construction layer: (1) a comprehensive style guide aligned with EU Interinstitutional Style Guide conventions for Greek and French, covering register disambiguation (marketing versus clinical), typography, and punctuation; (2) a morphological post-processing lemmatization layer addressing gender and number agreement in Greek and French; and (3) an expanded multilingual glossary with explicit CORRECT/INCORRECT pairs encoding register-specific terminology, superlative mitigation patterns, drug claim alternatives, and qualifier preservation rules, each linked to a PE.MAPS error code. All three enhancements were injected as a simultaneous package; component-level attribution is acknowledged as a limitation and is deferred to an ablation study. For a full illustration of the underlying KG architecture, see Figure C1 in Appendix C.

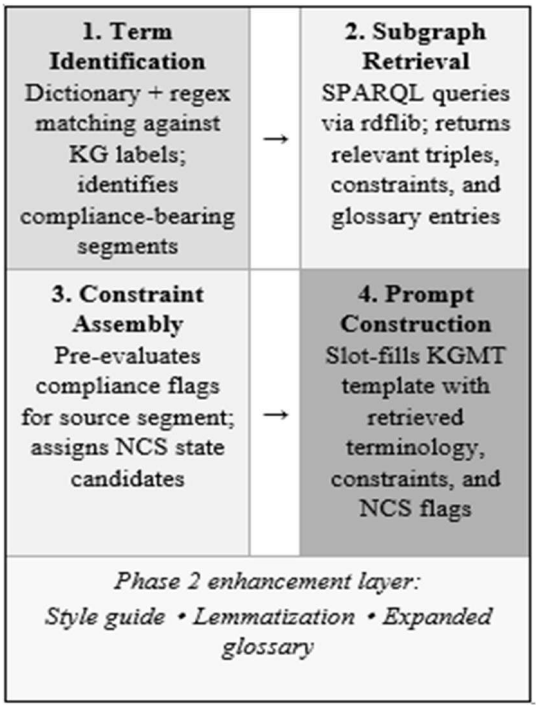

Figure 1: NCS classification states, penalty logic, and corpus examples.

3.4 Experimental Conditions and Prompt Design

Three conditions were evaluated. The Pure Baseline (PB) uses a generic translation instruction with no domain resources: 'You are a professional translator specializing in cosmetics. Translate the following from English to [language]. Maintain accuracy and appropriate register.' The Document-Augmented Baseline (DA) appends the raw glossary and compliance reference tables to the same instruction without structured retrieval. KGMT uses a structured prompt template populated by the four-stage pipeline (Figure 1), providing retrieved terminology mappings, NCS-flagged compliance constraints, and style guide rules specific to the source segment. All conditions use GPT-5 (OpenAI, 2025) at temperature 0.3, top-p 0.95. Each of the 58 segments was processed once per condition (n=1 per segment per condition; 174 output instances per language). Output stochasticity is treated as an empirical variable rather than a confound, as its measurement is a research objective.

3.5 Annotation Protocol and Evaluator Bias Control

Translation quality is evaluated using PE.MAPS (Gene, 2024), distinguishing Technical Errors (weighted by severity: Critical = 3, Major = 2, Minor = 1) from Preferential Modifications. TQS = 100 − (Total Technical Penalty Points / Word Count × 100). The full PE.MAPS error taxonomy with penalty weights is provided in Table B1 (Appendix B).The two annotators applied a cross-annotation protocol (Saldanha and O'Brien, 2013; Shadish et al., 2002): Annotator A annotated PB and DA primarily, with secondary review of KGMT; Annotator B annotated KGMT primarily, with secondary review of PB and DA. Neither annotator had access to the other's annotations during primary annotation. Disagreements were resolved through structured consensus using pre-established classification criteria. Inter-annotator agreement exceeded $\kappa = 0.85$ (Cohen, 1960), indicating strong reliability. The first author designed the corpus and KG but did not perform primary annotation of KGMT outputs, mitigating the principal evaluator bias risk.

4 Results

4.1 Source Compliance Detection Recall

PB and DA achieved 0% recall across all 30 indexed errors in both languages: without KG infrastructure, both conditions reproduced all source violations without flagging or mitigation. KGMT achieved 100% recall across all 31 evaluable violations. Table 2 presents a representative subset of the 30 indexed source errors. Representative correction strategies included: superlative mitigation ('REVOLUTIONARY' → 'ADVANCED' / 'NOVATRICE'), drug claim softening ('stimulates collagen at the cellular level' → 'supports natural collagen synthesis' / 'aide à soutenir la synthèse naturelle de collagène'), qualifier preservation ('reduces wrinkles' → 'reduces the appearance of wrinkles'), and modal preservation ('transforms' → 'visibly improves'). The NCS-C2 flag at segment 4 (1.0% retinol exceeding the EU 0.3% RE limit) was reproduced with an explicit NCS flag, correctly attributed to source under the SRC_FLAGGED state.

Seg	Code	Wt.	PB	DA	KGMT Action	NCS Status
2	C3	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → ADVANCED	Flagged; KGMT corrected
3	C1	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → visible results	Flagged; KGMT corrected
4	C2	3	REPRO-DUCED	REPRO-DUCED	REPRODUCED + [NCS flag]	**NCS-C2 (source error)**
18	C1	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → supports production	Flagged; KGMT corrected
19	C3	3	REPRO-DUCED	REPRO-DUCED	HEDGED → up to 3x	Flagged; KGMT corrected
22	C1	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → natural defense	Flagged; KGMT corrected
36	C1	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → reduces appearance	Flagged; KGMT corrected
37	C1+ C3	3	REPRO-DUCED	REPRO-DUCED	RESTRUCTURE D heading	Flagged; KGMT corrected
38	C3	3	REPRO-DUCED	REPRO-DUCED	SOFTENED → documented system	Flagged; KGMT corrected
43	C3	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → dermatologists	Flagged; KGMT corrected
47	C1	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → improve visibly	Flagged; KGMT corrected
48	C3	3	REPRO-DUCED	REPRO-DUCED	MITIGATED → documented results	Flagged; KGMT corrected
Total: 30 indexed source errors (+1 C8). PB/DA detection rate: 0%. KGMT detection rate: 100%. NCS-C2 (Seg 4) not penalized[source-originated concentration error reproduced with explicit NCS flag.						

Table 2: Source compliance detection by condition representative subset, n=12

4.2 Translation Quality Across Conditions

Table 3 presents full quality metrics. KGMT records TQS = 100.0% in both Greek and French under NCS classification, reflecting zero penalized technical errors. PB and DA accumulate substantial penalties from source error propagation (30 errors reproduced without flagging, classified TRANS_INTRODUCED) combined with translation-introduced accuracy (A1/A2) and fluency (F1) errors. Under NCS, the same source errors in KGMT are SRC_FLAGGED and therefore unpenalized in TQS.

Metric	PB (EL)	PB (FR)	DA (EL)	DA (FR)	KGMT (EL)	KGMT (FR)
Detection recall	0%	0%	0%	0%	**100%**	**100%**
Source C errors	30	30	30	30	**0 (NCS)**	**0 (NCS)**
Accuracy errors (A)	28	30	24	25	**0**	**0**
Fluency errors (F)	1	1	6	7	**0**	**0**
Total penalty pts	182	188	175	181	**0**	**0**
TQS (%)	71.6	73.1	72.7	74.4	**100.0**	**100.0**

Table 3: Translation quality metrics by condition and language.

4.3 Phase 1 versus Phase 2 Comparison

Table 4 presents the cross-phase comparison. The TQS reduction for PB and DA between phases reflects two factors: the stricter Phase 2 annotation protocol (A1/A2/F1 errors now systematically tracked) and natural LLM output variability across inference instances from the same prompt. This inter-instance divergence is an empirical finding: LLMs do not reliably reproduce identical terminological choices on repeated runs, which is operationally consequential in regulated workflows where consistency is a legal requirement (Castilho et al., 2017). KGMT's improvement from 83.9% to 100.0% (Greek) reflects three compounding effects: the infrastructural package, system maturation (detection recall: 24%→100%), and NCS methodological refinement.

Metric	PB Ph.1	PB Ph.2	DA Ph.1	DA Ph.2	KGMT Ph.1	KGMT Ph.2	ΔKGMT
Detection recall (comb.)	0%	0%	0%	0%	24%	**100%**	**+76pp**
Total penalty pts (EL)	83	182	78	175	102	**0**	**−102**
Total penalty pts (FR)	78	188	73	181	83	**0**	**−83**
TQS (EL, %)	86.8	71.6	87.6	72.7	83.9	**100.0**	**+16.1**
TQS (FR, %)	88.5	73.1	89.4	74.4	88.2	**100.0**	**+11.8**

Table 4: Phase 1 vs Phase 2 cross-condition comparison (combined EL + FR).

5 Discussion

5.1 NCS as a Reusable Methodological Contribution

The NCS framework addresses a systematic measurement gap identified across both phases. By separating SRC_FLAGGED from TRANS_INTRODUCED errors, NCS enables principled evaluation of two distinct competences: translation competence (does the system produce accurate, fluent, compliant output?) and quality assurance competence (does the system identify and flag source-level violations?). Conflating these in a single TQS metric, as Phase 1 did, misrepresented KGMT's capability profile. NCS is domain-agnostic and transferable to any regulated domain exhibiting cross-jurisdictional regulatory asymmetry, financial services translation, where content compliant under CFTC guidelines for US markets may violate MiFID II obligations in the EU, or where product classifications differ materially across jurisdictions, presents a structurally identical regulatory asymmetry, and the framework accommodates such domains without architectural modification.

5.2 Infrastructure Compounding and the Structure Before Scale Principle

The Phase 1–Phase 2 divergence for baseline outputs on identical prompts demonstrates that LLMs do not reliably produce consistent terminological choices across inference instances in compliance-sensitive contexts. KGMT, grounded in an explicit deterministic KG (Hogan et al., 2021), produces the same terminological guidance for the same source segment regardless of inference instance. The three Phase 2

enhancements compound: the glossary expansion eliminated A1 terminological substitution errors; the style guide resolved register inconsistencies; the lemmatization layer corrected F1 morphosyntactic errors. This supports a 'structure before scale' principle: targeted KG infrastructure additions yield disproportionate quality gains in regulated contexts, because each knowledge layer eliminates a class of errors rather than merely reducing their frequency (Gene and Sosoni, 2026).

5.3 Scope and Generalizability

The study is grounded in a single domain and a single 58-segment corpus. The dual-domain architecture of the broader research programme positions KGMT as a domain-agnostic framework: domains exhibiting structurally identical cross-jurisdictional regulatory asymmetries (financial services translation, where CFTC-compliant content may violate MiFID II, is one such domain) are accommodated by the framework. A full-interface KGMT editor implementing the NCS flagging system in real time has been developed and is the subject of a forthcoming system demonstration paper.

6 Limitations

Five limitations bound the current findings. First, the single-domain, single-corpus design limits immediate generalizability; cross-domain replication is required. Second, the two Phase 2 additions, e.g. the style guide with integrated morphological enforcement, and the expanded compliance-aware glossary, were deployed simultaneously; the independent contribution of each cannot be determined from the current data. The modular architecture makes component-level evaluation feasible as an open direction. Third, TQS = 100.0% rests on cross-annotator confirmation that self-flagged A2 variants are morphological artefacts of the lemmatization layer; independent replication is required. Fourth, outputs were generated once per segment per condition (n=1); replication across multiple runs would characterize LLM variability more precisely. Fifth, the study uses GPT-5 as the sole LLM backend; comparative evaluation across model families remains open. The first author designed the evaluation corpus and KG, acknowledged as a potential bias source mitigated but not eliminated by the cross-annotation protocol.

7 Conclusions

This paper presents three contributions to regulated translation research. The NCS classification provides a principled, reusable framework for distinguishing source-propagated from translation-introduced compliance errors. The indexed source design provides an objective detection recall ground truth grounded in explicit regulatory criteria. The empirical evaluation demonstrates that an enhanced KGMT architecture achieves 100% source compliance detection recall and zero penalised translation errors in both Greek and French, against 0% detection and substantially lower TQS for both unaugmented conditions.

The inter-instance divergence of baseline outputs from identical prompts demonstrates that LLMs do not reliably or auditably guarantee compliance consistency—a finding with direct operational consequences for any domain where terminology consistency and regulatory traceability are legal requirements. KGMT, as a deterministic knowledge-grounded layer, stabilizes this variability. The framework is designed for domains exhibiting cross-jurisdictional regulatory asymmetry; financial services translation presents a structurally identical challenge and represents a natural domain for future evaluation. Component-level evaluation of the Phase 2 style guide and expanded glossary is feasible within the current modular architecture and represents an open direction for this or future research. Evaluation across multiple LLM backends and development of source-anchored fallback detection for semantic paraphrases not captured by the current string-matching pipeline are further open directions.

Acknowledgements

The authors gratefully acknowledge the technical support of Edwin Trebels (LangOptima) and Prasad Yalamanchi (Lead Semantics) for Knowledge Graph construction and engineering, and the second annotator, an independent professional translator specializing in regulatory and dermocosmetics translation, whose cross-annotation contribution strengthened the inter-annotator reliability of this study. This research is conducted under the postdoctoral supervision of Dr. Vilelmini Sosoni (Ionian University).

References

Bowker, Lynne. 2002. *Computer-Aided Translation Technology: A Practical Introduction*. University of Ottawa Press.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *Prague Bulletin of Mathematical Linguistics*, 108:109–120.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Gene, Viveta. 2024. Machine translation post-editing and professional translators: The contribution of training to the reduction of effort and the improvement of quality [Doctoral dissertation, Ionian University]. https://doi.org/10.12681/eadd/58063.

Gene, Viveta and Vilelmini Sosoni. 2026. Knowledge graph-mediated translation (KGMT): A theoretical framework for compliance-aware multilingual communication in regulated domains. In *Proceedings of Convergence 2026* (Surrey, 17–19 June 2026). In press. Available at: https://www.surrey.ac.uk/centre-translation-studies/convergence-2026/programme

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *ACM Computing Surveys*, 54(4), Article 71, 1– 37. https://doi.org/10.1145/3447772.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 33:9459–9474.

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12:455–463.

Moorkens, Joss, Sheila Castilho, Federico Gaspari, and Stephen Doherty (Eds.). 2018. *Translation Quality Assessment: From Principles to Practice*. Springer.

OpenAI. 2025. *GPT-5 technical report*. OpenAI. https://openai.com

Pan, Shirui, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599. https://doi.org/10.1109/TKDE.2024.3352100.

Prieto Ramos, Fernando. 2014. Quality assurance in legal translation: Evaluating process, competence and product in the pursuit of adequacy. *International Journal for the Semiotics of Law*, 28:11–30.

Saldanha, Gabriela and Sharon O'Brien. 2013. *Research Methodologies in Translation Studies*. St. Jerome Publishing.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Skuce, Doug and Ingrid Meyer. 1990. Concept analysis and terminology: A knowledge-based approach to documentation. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING 1990)*, Volume 1, pages 197–201. https://aclanthology.org/C90-1011.

Appendix A. PE. MAPS Evaluation Framework

This appendix presents the PE.MAPS (Post-Editing Monitoring, Assessment, Problem-Solving; Gene, 2024) evaluation framework applied throughout Phase 2 of this study. Table A1 lists the error codes, dimensions, definitions, and penalty weights used in annotation. Table A2 summarizes the penalty weight system. The TQS formula is provided below. For the full theoretical framework, see Gene and Sosoni (2026).

Code	Category	Dimension	Definition	Weight
C1	Drug/Medical Claims	Compliance	Therapeutic claims beyond cosmetic scope (e.g., 'cures', 'treats', 'repairs')	3
C2	Prohibited Ingredient/Concentration	Compliance	Prohibited substances or concentrations exceeding regulatory limits	3
C3	Unsubstantiated Superlative	Compliance	Absolute or comparative claims without substantiation (e.g., 'best', 'only', 'maximum')	3
C5	Quantified Claim Mismatch	Compliance	Numbers or timeframes inconsistent with source data	3
C8	Warning Omission	Compliance	Required safety advisory absent (e.g., pregnancy warning for retinol)	2
C9	Missing Evidence	Compliance	Claims without required scientific or clinical substantiation	3
A1	Term Substitution	Accuracy	Incorrect domain terminology (maintains source error or uses wrong term)	3
A2	Terminological Inconsistency	Accuracy	Same concept translated with different terms across the text	2
F1	Grammar Error	Fluency	Morphosyntactic errors: gender, number, case, verb form agreement	1
F2	Style/Register	Fluency	Inappropriate register for domain (marketing vs. clinical)	1
NCS	Non-Compliance in Source	Classification	Source error flagged by KGMT but NOT penalised in TQS; recorded for detection recall only	0
—	No Error (Preferential Edit)	—	MT output acceptable; translator modification is preferential, not corrective	0

Table A1. PE. MAPS Error Code Taxonomy — Codes, Definitions, and Penalty Weights

Colour key: Red = Compliance dimension; Amber = Accuracy dimension; Green = Fluency dimension; Blue = NCS clasasification (not penalised).

Severity	Weight	Codes	Rationale
Critical (Compliance)	3	C1, C3, C5, C9	Regulatory violations; legal or health risk
Major (Warning)	2	C8	Mandatory safety warnings (2 pt as absence, not direct harm)
Major (Accuracy)	3	A1	Terminology errors impacting meaning or compliance
Minor (Accuracy)	2	A2	Inconsistencies affecting usability or cohesion
Minor (Fluency)	1	F1	Grammar errors affecting readability, not meaning
Info (not penalised)	0	NCS, F2	Source-originated or preferential edits (not scored)

Table A2. PE. MAPS Penalty Weight System

TQS Formula

$$TQS = 100 \times (1 - \Sigma[\text{errors} \times \text{weights}] / \text{word count})$$

Key principle: NCS (Non-Compliance in Source) errors are identified by KGMT but not penalised in TQS. Only translation-introduced errors reduce TQS. This design ensures that translators are not penalised for source-level compliance issues outside their control.

Appendix B. PE.MAPS Full Error Code Taxonomy

PE.MAPS (Post-Editing Monitoring, Assessment, Problem-Solving; Gene, 2024) evaluates translation quality across Compliance (C1–C9), Accuracy (A1–A7), and Fluency (F1–F4) dimensions. NCS (Non-Compliance in Source) identifies source-level violations flagged by KGMT that are excluded from TQS penalisation. TQS = 100 × (1 − Σ[errors × weights] / word count). Codes not triggered in the Phase 2 corpus are included for completeness.

Code	Category	Dimension	Definition	Weight
C1	Drug/Medical Claims	Compliance	Therapeutic claims beyond cosmetic scope (e.g., 'cures', 'treats', 'repairs')	3
C2	Prohibited Ingredient/Conc.	Compliance	Prohibited substances or concentrations exceeding regulatory limits	3
C3	Unsubstantiated Superlative	Compliance	Absolute or comparative claims without substantiation (e.g., 'best', 'only', 'maximum')	3
C4	INCI Name Violation	Compliance	Incorrect or non-standard translation of INCI nomenclature	3
C5	Quantified Claim Mismatch	Compliance	Numbers or timeframes inconsistent with source data	3
C6	Qualifier Loss	Compliance	Removal of hedging language (e.g., 'may', 'appears to', 'visible results')	2
C7	Conditional→Definitive Shift	Compliance	Modal or conditional rendered as definitive (e.g., 'can reduce' → 'reduces')	2
C8	Warning Omission	Compliance	Required safety advisory absent (e.g., pregnancy warning for retinol)	2
C9	Missing Evidence	Compliance	Claims without required scientific or clinical substantiation	3
A1	Term Substitution	Accuracy	Incorrect domain terminology: wrong target term used for source concept	3
A2	Terminological Inconsistency	Accuracy	Same concept translated with different terms across the text	2
A3	Spelling/Morphological Error	Accuracy	Target term misspelled or morphologically incorrect	1
A4	Omission	Accuracy	Source content absent from target without justification	2
A5	Addition	Accuracy	Content added in target not present in source	2
A6	Mistranslation	Accuracy	Semantic shift: target meaning diverges from source meaning	2
A7	Number/Unit Error	Accuracy	Numbers, percentages, or measurement units rendered incorrectly	3
F1	Grammar Error	Fluency	Morphosyntactic errors: gender, number, case, or verb form agreement	1
F2	Style/Register	Fluency	Inappropriate register for domain context (marketing vs. clinical)	1
F3	Punctuation/Mechanics	Fluency	Punctuation, capitalisation, or spacing errors	1
F4	Coherence	Fluency	Logical flow problems affecting readability	1
NCS	Non-Compliance in Source	Classification	Source error flagged by KGMT; NOT penalised in TQS; recorded for detection recall only	0
—	No Error (Preferential Edit)	—	MT output acceptable; translator modification is preferential, not corrective	0

Table B1. PE.MAPS Complete Error Code Taxonomy

Colour key: Red = Compliance (C1-C9); Amber = Accuracy (A1-A7); Green = Fluency (F1-F4); Blue = NCS (not penalised in TQS).

Appendix C. KGMT Knowledge Graph: Dermocosmetics Source Text

Figure C1 presents the knowledge graph applied to the dermocosmetics source text in Phase 1 (Gene and Sosoni, 2026), implemented in TextDistil v.3.0 (Lead Semantics). The graph integrates four node classes: ingredient entities (blue), regulatory references (purple), regulatory constraints (orange), and concept/claim-type nodes (teal). Compliance flags generated by KGMT are rendered as green triangular nodes labelled [N] CATEGORY, where N is the source-text segment identifier and CATEGORY denotes the PE.MAPS compliance code (C1, C2, C3, C6) or KG-internal classifier (DRUG_MECH, SUPER, OVERPREC) that triggered the flag. Red edges denote violation, instance-of, and constraint-failure relations; grey edges denote structural and disambiguation relations. The graph encodes 58+ terminology entries across ingredient classifications, regulatory claim categories, concentration limits, and cross-jurisdictional EU/US distinctions. This architecture underpins the Phase 2 enhanced pipeline described in Section 3.3.

Figure C1. Knowledge Graph applied to the dermocosmetics source text (TextDistil v.3.0, Lead Semantics). Triangular nodes (green) represent compliance-flagged concepts; circular nodes represent terminology and regulatory entities. Node labels for flags are formatted [segment-id] CATEGORY (e.g., [5] C2 denotes segment 5, prohibited concentration).