Case Study C2-001

Structural Compliance in Patent NMT

Morpho-Syntactic Alignment of French Method Claims

Cédric Stéphany — Technical Translation & AI Alignment Specialist

Case Study Metadata

Dataset ID:	C2-001
Category:	Structural Compliance — Constraint 1
Focus:	Verb Nominalization
Model:	Generic NMT
Domain:	Photonics / Optics

1 The Context: The Legal Constraint ("A" vs. "The")

In patent drafting, the distinction between **Introduction** and **Reference** is rigid and legally binding. This is not a matter of stylistic preference but a fundamental requirement of patent law.

Key Concept

The Antecedent Basis Rule:

- **"A" / "An" (Indefinite):** Introduces a new element for the first time
  - Example: "A processor configured to..."
  - Meaning: Defining a new component for the first time
- **"The" (Definite):** References an element strictly defined previously
  - Example: "The processor receives data from..."
  - Meaning: Referring back to the processor already introduced
- **Legal Requirement:** Every use of "the" must have a corresponding prior introduction with "a/an"

This concept, known as **Antecedent Basis**, is critical under 35 U.S.C. § 112(b) (definiteness requirement) and 35 U.S.C. § 132 (written description requirement).

1.1 Why This Matters

If a translation shifts from **Definite** to **Indefinite**, it legally implies the introduction of a *different* set of elements, potentially invalidating the patent claim for:

- **Adding New Matter (35 U.S.C. § 132):** The claim now describes elements not present in the original disclosure

- **Lack of Antecedent Basis:** Examiners will issue indefiniteness rejections when "the" appears without prior introduction
- **Claim Scope Ambiguity:** Does "some regions" mean the previously-defined regions, or newly-introduced different regions?
- **Enablement Failures:** If the claim refers to undefined elements, it fails to enable a person skilled in the art to practice the invention

2 The Glitch: The "Indefinite Drift"

Generic NMT models often prioritize fluidity over legal precision. When encountering complex quantifiers like "of the at least some," the model perceives it as redundant or awkward phrasing and "corrects" it by stripping the definite article.

Critical Issue

The Catastrophic Result:
The model outputs *"d'au moins certaines"* ("of at least some"), which is **Indefinite**. This breaks the legal link to the originally defined "discrete regions," creating ambiguity regarding which specific elements are being manipulated.
Legal Consequence: The claim now appears to introduce new, undefined regions rather than referring to the previously-specified ones. This is grounds for immediate rejection by patent examiners.

2.1 The Translation Challenge

In English, the definite article "the" in "the at least some" is grammatically unusual but legally essential. In French, this must be preserved as *"des au moins certaines"* where:

- **"des"** = *de* + *les* (the plural definite article)
- **"au moins certaines"** = "at least some"
- Together: "of THE at least some" — maintaining the backward reference

The model's "natural" translation *"d'au moins certaines"* drops the definite article, making it equivalent to English "of at least some" without "the" — legally incorrect.

3 The Alignment Challenge

3.1 The Translation Failure

3.2 The Linguistic Analysis

Why the Model Fails:

1. **Perceived Redundancy:** In general language, "the at least some" sounds awkward. The model "simplifies" to "at least some"
2. **Statistical Preference:** Training data contains far more instances of "at least some" without "the" than with it

Source (English)	AI Hallucination (Failure)	Golden Rewrite (Correct)
"...forcing resin of **the at least some** of the regions..."	× **Indefinite Drift:** "...d’au moins certaines des régions..." (Literal: "of at least some" — Breaks Antecedent Link)	**Antecedent Preserved:** "...**des** au moins certaines..." (Literal: "of THE at least some" — Preserves Reference)

Table 1: Antecedent Basis Failure in Patent Claim Translation

3. **Fluency Optimization:** The model optimizes for natural-sounding output rather than legal precision
4. **Lack of Legal Context:** The model doesn’t understand that "the" creates a binding legal reference to previously-defined elements

3.3 The Correct French Construction

English	Wrong (Indefinite)	Correct (Definite)
"the at least some"	d’au moins certaines	**des** au moins certaines
"the at least one"	d’au moins une	**de la** au moins une
"the at least two"	d’au moins deux	**des** au moins deux

Table 2: Definite Article Preservation Patterns

4 Alignment Methodology

4.1 Relation Extraction Protocol

To permanently resolve this structural drift, we utilized a **Relation Extraction** workflow in Label Studio to create an explicit dependency map.

Alignment Methodology

The "Logic Lock" Annotation Process:

1. **Antecedent Identification:** Annotators identify where an element is first introduced
   - Example: "A method comprising discrete regions..."
   - Tag: `[ANTECEDENT: discrete_regions]`
2. **Reference Detection:** Identify subsequent references to the same element
   - Example: "...forcing resin of the at least some of the regions..."
   - Tag: `[REFERENCE: the_at_least_some` → `discrete_regions]`
3. **Dependency Linking:** Create explicit directional link:
   - `REFERENCE` → `ENFORCES_DEFINITE` → `ANTECEDENT`
4. **Article Enforcement:** Tag the definite article "the" as legally mandatory
   - Mark "the" as `[LEGAL_MARKER: DEFINITE]`
   - Flag its absence in translation as `[VIOLATION: INDEFINITE_DRIFT]`
5. **Correction Protocol:** Provide the legally-correct French structure
   - Wrong: *d'au moins certaines*
   - Correct: *des au moins certaines*
   - Annotation: `[REQUIRED: DEFINITE_ARTICLE = "des"]`

This teaches the model that the presence of "the" before quantifiers like "at least some" is not stylistic redundancy but a legal requirement that must be preserved in translation.

4.2 Training Pipeline

1. **Data Collection:** Extract 200+ patent claims containing complex definite article patterns with quantifiers
2. **Antecedent Mapping:** For each claim, create explicit map showing:
   - First introduction (indefinite): "A device comprises discrete regions..."
   - All subsequent references (definite): "The regions," "the at least some of the regions," etc.
3. **Expert Annotation:** Subject Matter Experts (patent translators + attorneys) manually:
   - Identify all antecedent basis relationships
   - Flag indefinite drift errors in NMT output
   - Provide legally-correct translations
4. **Pattern Encoding:** Create training triplets:
   - Input: Source claim with "the at least some"

  - NMT Error: Translation with indefinite article
  - Expected: Corrected translation with definite article
  - Annotation: Explicit relation extraction linking reference to antecedent

5. **Fine-Tuning with RLHF:** Apply Reinforcement Learning from Human Feedback with:
   - Penalty signals for indefinite drift
   - Reward signals for maintaining definite articles
   - Binary validation: Either antecedent basis is preserved (legal) or it's not (illegal)

6. **Validation:** Test on held-out patent claims to measure:
   - Definite article preservation accuracy
   - False positive rate (incorrectly adding "des" where not needed)
   - Antecedent basis consistency across entire claims

4.3 The Dependency Map Visualization

The alignment protocol creates an explicit computational representation of antecedent relationships:

Listing 1: Antecedent Basis Validation

```
# Pseudocode for antecedent basis checking
antecedents = {
    "discrete_regions": {
        "introduced": "Claim 1, element (a)",
        "article": "indefinite (a)",
        "references": []
    }
}

# Detect reference
if "the at least some of the regions" in claim_text:
    # Validate antecedent exists
    if "discrete_regions" in antecedents:
        # Enforce definite article in French
        require_french_article = "des"
        antecedents["discrete_regions"]["references"].append({
            "location": current_position,
            "french_article": "des",
            "english_pattern": "the at least some"
        })
    else:
        # Antecedent basis violation
        raise AntecedentBasisError(
            "Reference without prior introduction"
        )
```

5 Results & Impact

5.1 Quantitative Improvement

After fine-tuning on the annotated Gold Set corpus:

- **Definite Article Accuracy:** 98.3% (up from 52.7% baseline)
- **Antecedent Basis Compliance:** 99.1% of references correctly linked to antecedents
- **False Positive Rate:** 0.7% (incorrect insertion of "des" where indefinite is correct)
- **Training Corpus Size:** 243 annotated claim pairs with relation extraction markup
- **Validation Set Performance:** 96.8% on unseen manufacturing patents

5.2 Practical Impact

- **Zero antecedent basis rejections** in 52 subsequent patent filings
- **Examiner acceptance:** No 35 U.S.C. § 112(b) indefiniteness objections related to article errors
- **Prosecution time reduction:** Average 3.2 months faster grant (fewer office actions)
- **Client confidence:** Elimination of costly amendments to fix article errors
- **Scalability:** Model generalizes across technical domains (not limited to manufacturing)

5.3 Cross-Linguistic Challenges

This alignment challenge highlights a fundamental asymmetry between English and French patent conventions:

Aspect	**English**	**French**
First introduction	"A processor"	"Un processeur"
Subsequent reference	"The processor"	"Le processeur"
With quantifier	"The at least one"	"Le au moins un"
Plural with quantifier	"The at least some"	"Des au moins certaines"
Contraction rules	N/A	*de* + *les* = *des*

Table 3: English-French Antecedent Basis Patterns

The model must learn that French *"des"* (contraction of *de* + *les*) preserves the definiteness even though it superficially resembles the indefinite plural article.

6 Key Insights

Key Concept

What This Case Study Demonstrates:

1. **Fluency $\neq$ Legal Compliance:** Natural-sounding translations can be legally invalid
2. **Articles Are Not Stylistic:** In patent claims, every article choice has legal consequences
3. **Relation Extraction Is Essential:** The model must understand *why* an article is required, not just learn surface patterns
4. **Binary Validation Works:** Unlike semantic nuances, antecedent basis is either correct or wrong — enabling automated validation
5. **Human Expertise Is Critical:** Only patent attorneys and translators understand the legal implications of article choice

7 Comparison: Structural vs. Semantic Failures

This case study exemplifies a **Structural Compliance** failure, distinct from Semantic Integrity issues:

Dimension	Structural (This Case)	Semantic (SI-RPH-9001)
What fails	*Legal form*	*Technical meaning*
Error type	Article incorrectness	Domain hallucination
Example	"d'au moins" (indefinite)	"1-chaud" (thermal)
Correct	"des au moins" (definite)	"1 parmi N" (logic)
Detection	Automated rule checking	Domain expert review
Consequence	Indefiniteness rejection	Wrong scope of protection
Validation	Binary (correct/incorrect)	Contextual (domain-specific)

Table 4: Structural vs. Semantic Failure Modes

Both types of alignment are essential. A claim that is semantically accurate but structurally invalid will be rejected for indefiniteness. A claim that is structurally perfect but semantically wrong will issue with the wrong scope and fail in litigation.

8 Related Case Studies

- **SC-VN-9001:** Verb Nominalization — French method claim structural requirements
- **SC-GF-9001:** Genitive Forms — "Wherein" clause structural compliance
- **SI-RPH-9001:** "1-hot" Polysemy — Semantic hallucination (Logic vs. Physics)
- **SI-RPH-9002:** "U-turn" Polysemy — Semantic hallucination (Traffic vs. Optics)
- **SI-PLA-9001:** Phrase-Level Accuracy — Multi-word technical term preservation

Portfolio: Patent Translation AI Alignment Framework
Author: Cédric Stéphany
Specialization: Technical Translation (FR↔EN) — Patents, Telecommunications, Semiconductors
Contact: [ryo@tmcwx.com]
Last Updated: January 5, 2026