

# Les propositions atomiques : un pont entre approches neuronales et symboliques

Interprétabilité par construction

---

Luc Pommeret, Thomas Gerald, Sophie Rosset, Patrick Paroubek, Christophe Servan, Sahar Ghannay  
GDR TAL – 2026

CNRS, LISN, Orsay – Université Paris-Saclay – AMIAD, Palaiseau  
*prenom.nom@lison.fr*

1. Contexte et motivation
2. Formalisation des propositions atomiques
3. Propositionneur multilingue
4. Résultats
5. Conclusion

# Contexte et motivation

---

# Pourquoi les propositions atomiques?

**Constat :** les approches d'explicabilité sont largement **post hoc**

- Elles expliquent des décisions *déjà prises* par des modèles opaques
- Pas de garantie *structurelle* sur la traçabilité

**Notre approche :** **interprétabilité par construction**

- Décomposer les textes en unités **atomiques**
- Chaque unité est traçable et indivisible
- Applicable à plusieurs tâches de TAL/RI

## Exemple

Le chat **et** le chien sont dans la cuisine

⇓

1. Le chat est dans la cuisine
2. Le chien est dans la cuisine

1. **Formaliser** la notion de proposition atomique
2. **Concevoir** un propositionneur multilingue léger
3. **Évaluer** sur des tâches concrètes de TAL

## Tâches cibles :

### Extraction de triplets

OpenIE / ClosedIE  
SMiLER, FewRel

### Recherche d'information

RAG / RI dense  
NQ, CQuAE

### Évaluation de résumés

Corrélation humaine  
SummEval

## Formalisation des propositions atomiques

---

Une proposition atomique respecte deux propriétés :

## ATOMICITÉ

Plus petite unité sémantique : toute décomposition supplémentaire entraînerait un **changement sémantique** (augmentation de l'information).

## AUTONOMIE

Sémantiquement complète : toute entité est explicitée par son nom propre ou une description univoque. **Pas de pronoms ambigus.**

**Problème** : la littérature ne s'accorde pas sur une définition formelle unique.

⇒ Nous proposons une formalisation via la **théorie de l'information sémantique** de Carnap & Bar-Hillel (1952).

**Idée :** transférer la théorie de Shannon dans le domaine de la logique des prédicats.

Soit  $\phi$  une formule. Le **contenu sémantique** :

$$\text{Cont}(\phi) = \{w \in \mathcal{W} \mid w \models \phi\}$$

L'information sémantique :

$$I(\phi) = -\log_2 \left( \frac{|\text{Cont}(\phi)|}{|\mathcal{W}|} \right)$$

## Décomposition sûre

$\phi \rightarrow \psi$  est **sûre** si  $I(\phi) \geq I(\psi)$   
(on perd ou conserve de l'information)

## Mauvaise décomposition

$\phi \rightarrow \psi$  est **mauvaise** si  $I(\phi) \leq I(\psi)$   
(on hallucine de l'information)

## Lemmes :

- **Conjonction** ( $A \wedge B \rightarrow A$ ) : décomposition **sûre**  
Le chat **et** le chien sont dans la cuisine  $\rightarrow$  Le chat est dans la cuisine
- **Disjonction** ( $A \vee B \rightarrow A$ ) : décomposition **mauvaise**  
Le chat **ou** le chien est dans la cuisine  $\rightarrow$  Le chat est dans la cuisine  $\Rightarrow$  hallucination!
- **Implication** ( $A \rightarrow B \rightarrow A$ ) : décomposition **mauvaise**

## Théorème

Une proposition est **atomique** si et seulement si elle est une **clause d'une FNC** (forme normale conjonctive).

## Conséquence

- Pas de conjonction de sous-faits
- Peut contenir une disjonction
- Critère d'arrêt formel pour la décomposition

# Propositionneur multilingue

---

## Schéma en deux étapes :

1. **Génération** : *Qwen3-72B* produit les propositions atomiques (prompt 1)
2. **Vérification** : *Qwen3-72B* vérifie l'autonomie de chaque proposition (prompt 2)

## Distillation vers *Qwen3-0.6B* :

- 56 000 passages Wikipedia
- 6 langues : FR, EN, DE, ES, IT, PT
- Contexte : 32 768 tokens

## Comparaison

|                 | F1    | Param. |
|-----------------|-------|--------|
| Notre modèle    | 0.842 | 596M   |
| Propositionizer | 0.829 | 783M   |

⇒ Meilleur F1 BERTScore, plus petit et multilingue

**Objectif :** garantir l'atomicité des sorties via un raffinement itératif.

## Algorithme

**Require :**  $N=5$ , texte  $t$ , propositionneur  $\mathcal{M}$

```
1:  $\mathcal{P}_0 \leftarrow \mathcal{M}(t)$ 
2:  $\mathcal{A}_0 \leftarrow \{p \in \mathcal{P}_0 \mid \{p\} = \mathcal{M}(p)\}$ 
3: while  $\mathcal{P}_i \neq \mathcal{A}_i$  et  $i < N$  do
4:    $\mathcal{P}_{i+1} \leftarrow \bigcup_{x \in \mathcal{P}_i} \mathcal{M}(x)$ 
5:    $\mathcal{A}_{i+1} \leftarrow \{x \in \mathcal{P}_{i+1} \mid \{x\} = \mathcal{M}(x)\}$ 
6: end while
7: return  $\mathcal{A}_i$ 
```

## Principe :

- Réitérer la décomposition sur les sorties
- Convergence = point fixe (aucune nouvelle proposition)
- Seuil maximal de 5 itérations

## Démo :

<https://huggingface.co/spaces/Zual/MPropositionneur-Space>

# Résultats

---

# Extraction de triplets (SMiLER / FewRel)

| Corpus | Méthode     | GLiREL      |             |             | Qwen3-0.6B  |            |             | Qwen3-4B    |            |             |
|--------|-------------|-------------|-------------|-------------|-------------|------------|-------------|-------------|------------|-------------|
|        |             | Acc         | e-rec       | r-rec       | Acc         | e-rec      | r-rec       | Acc         | e-rec      | r-rec       |
| SMiLER | Directe     | 57.3        | <b>99.5</b> | 57.6        | 30.8        | <b>100</b> | 30.8        | <b>75.7</b> | <b>100</b> | 75.7        |
|        | Proposition | 48.5        | 82.9        | 58.5        | <b>38.9</b> | 82.9       | <b>46.9</b> | 65.5        | 82.9       | <b>79.1</b> |
|        | Combinée    | <b>59.1</b> | <b>99.5</b> | <b>59.3</b> | –           | –          | –           | –           | –          | –           |
| FewRel | Directe     | 48.7        | <b>100</b>  | 48.7        | <b>42.2</b> | <b>100</b> | 42.2        | <b>67.7</b> | <b>100</b> | 67.7        |
|        | Proposition | 40.2        | 75.8        | 53.1        | 40.3        | 75.8       | <b>53.2</b> | 51.5        | 75.8       | <b>68.0</b> |
|        | Combinée    | <b>50.0</b> | <b>100</b>  | <b>50.0</b> | –           | –          | –           | –           | –          | –           |

- L'atomisation **augmente le rappel de relations**
- La méthode Combinée (Direct + Propositions) améliore la précision globale avec GLiREL
- Avec les petits LLMs : gain en précision via les propositions

## Natural Questions (NQ)

| Granularité         | R@5          | R@20         | NDCG@10      |
|---------------------|--------------|--------------|--------------|
| Passages            | 25.14        | 48.90        | 0.241        |
| Phrases             | 34.49        | 58.87        | 0.387        |
| <b>Propositions</b> | <b>38.67</b> | <b>61.53</b> | <b>0.428</b> |

**+5 points** de R@5 vs phrases

**+13 points** de R@5 vs passages

## CQuAE v2 (français)

|            | Passage     |      | Phrase      |             | Prop.       |             |
|------------|-------------|------|-------------|-------------|-------------|-------------|
|            | R@1         | R@5  | R@1         | R@5         | R@1         | R@5         |
| Global     | 41.0        | 63.3 | 41.2        | 68.8        | <b>50.4</b> | <b>70.9</b> |
| Factuel    | <b>52.8</b> | 67.2 | 46.4        | 71.2        | <b>52.8</b> | <b>74.4</b> |
| Définition | 36.5        | 59.5 | 28.6        | 55.6        | <b>50.8</b> | <b>73.0</b> |
| Cours      | 42.3        | 67.5 | 48.0        | <b>80.5</b> | <b>55.3</b> | 70.7        |
| Synthèse   | 36.8        | 62.4 | <b>43.2</b> | <b>68.0</b> | 44.0        | 65.6        |

Excellent sur faits/définitions; mitigé sur synthèse

# Évaluation de résumés (SummEval)

Reproduction des expériences de Seval-ex en remplaçant le LLM par notre propositionneur.

| Méthode          | Fluidité | Cohérence | Exact. fact. | Pertinence | Vitesse  |
|------------------|----------|-----------|--------------|------------|----------|
| <i>Qwen3-32B</i> | 0.366    | 0.232     | 0.527        | 0.266      | 3.7 r/s  |
| Notre modèle     | 0.249    | 0.270     | 0.437        | 0.319      | 20.5 r/s |

## Points forts

- Meilleure **cohérence** et **pertinence**
- **×5** plus rapide (600M vs 32B param.)

## Limites

- Baisse en fluidité et exactitude factuelle
- Compromis taille/qualité

# Conclusion

---

## Contributions :

1. **Formalisation** : proposition atomique  $\Leftrightarrow$  clause FNC (théorie de l'information sémantique)
2. **Propositionneur multilingue** : distillé (Qwen3-0.6B), récursif, 6 langues
3. **Évaluation polyvalente** : gains en extraction de triplets, RI et évaluation de résumés

**Message clé** : traçabilité et performance vont ensemble – c'est parce que chaque unité est sémantiquement élémentaire que l'extraction est plus fiable.

## Perspectives :

- Annotation humaine de l'atomicité
- Évaluation de l'efficacité du raffinement récursif
- Approches neuro-symboliques pour optimiser la génération

# Questions?

<https://huggingface.co/spaces/Zual/MPropositionneur-Space>