

A NER Model for Science Fiction’s Lexical Innovation

Mathilde Ducos¹ , and Frédéric Landragin² 

¹ Lattice - Langues, Textes, Traitements informatiques, Cognition, UMR 8094, Sorbonne Nouvelle University, Paris, France

² Lattice - Langues, Textes, Traitements informatiques, Cognition, UMR 8094, CNRS, Paris, France

Abstract

Speculative fiction, including science fiction, is distinguished by its ability to generate an invented lexicon composed of terms called novums. However, these lexical units are absent from Named Entity Recognition (NER) models applied to computational literary studies, mainly because of their complexity in detection (highly varied syntactic distributions, strong semantic ambiguity). We present a CamemBERT-based NER model, adapted to the analysis of science fiction narratives and enriched with a dedicated class (NOV for novum) to identify these lexical inventions. Evaluated on a specific annotated corpus, our model achieves an F1-score of 63% for the detection of the very rare NOV class, with a macro F1-score of 79% across all entities. This model thus provides a tool for studying how these lexical creations reflect or even anticipate the technological and social imaginaries of their time. The resources presented here, including the annotated corpus and code, are available online.

Keywords: Computational Literary Studies, Entity Recognition, Novum

1 Introduction

Francophone science fiction, defined as a collection of speculative narratives exploring scientific hypotheses or technological anticipations [5], is based on the fundamental question “what if?” [4]. This genre mobilizes distinct narratological and linguistic mechanisms depending on whether they are considered from the reader’s or the author’s perspective. For the reader, suspension of disbelief and the construction of a fictional encyclopedia [6] enable them to grasp the universe of the narrative. From the author’s perspective, cognitive estrangement [9] and worldbuilding are tools for creating novums, referring to an entity, object, or phenomenon that is absent from the real world but anchored in the logic of the narrative. This is how the novums “telescreen” in Orwell’s *1984* (1949) and “psychohistory” in Asimov’s *Foundation series* (1942-1993) were created.

The study of novums in a French-language corpus is justified by the historical and thematic uniqueness of French science fiction ([5]; [16]), as well as by the richness of its lexical creativity. To conduct this study, we use the FrenchSFcorpus¹, composed of 10 million tokens annotated as novums and named entities according to a BIO schema. This corpus consists of 246 full-text French-language science fiction novels and short stories that are free of copyright restrictions.

However, automatic recognition of novums poses a methodological challenge: they are not simply an extension of traditional NER classes (people, places, organizations), but a distinct category. Traditional named entities, even invented ones (names of characters, places, or companies), remain anchored in existing categories of NER models. Conversely, novums mainly refer to new concepts or objects, often specific to a single literary narrative, conveying a particular idea or invention. Identifying them therefore requires adding a dedicated NER class, which necessitates appropriate annotation and modeling.

Mathilde Ducos, and Frédéric Landragin. “A NER Model for Science Fiction’s Lexical Innovation.” In: *Computational Cultural Science Workshop 2026*, ed. by Suzanne Mpouli, Jean-Baptiste Camps, Florian Cafiero, and Chahan Vidal-Gorène. Vol. 6. Anthology of Computers and the Humanities. 2026, 23–31. <https://doi.org/10.63744/XXwEt9qrIXbC>.

© 2026 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

¹ FrenchSFcorpus: <https://github.com/mducos/FrenchSFcorpus/tree/main>

This article presents an original approach to recognizing novums in francophone science fiction texts, based on the FrenchSFcorpus corpus annotated in classic NER classes as well as in the NOV class. Our contributions are twofold:

- the adaptation of a CamemBERT-based model for this task;
- the provision of annotated and reusable resources.

This article is organized as follows: Section 2 provides an overview of the state of the art in defining novum and NER approaches applied to literary corpora; Section 3 details the integration of this new class into a NER annotation scheme; Section 4 presents the experimental results; finally, Section 5 discusses the results and implications for the analysis of technological imaginaries in science fiction.

2 State of the art

2.1 Novum definition

A novum refers to an invented lexical unit with no reference in the real world, but whose coherence is ensured by the internal logic of the narrative. Terms such as “telescreen” (Orwell, 1949), “psychohistory” (Asimov, 1942-1993) and “terraformation” (Robinson, 1992-1996) illustrate science fiction’s ability to create expressions that are immediately understandable thanks to a cognitive logic shared between author and reader [14]. Although these words are not found in dictionaries, they are anchored in the reader’s encyclopedia [6], or even become lexicalized, as in the case of “terraformation”, and play a structuring role in the diegesis by introducing an ontological break with the world of reference.

For Suvin [14], the novum is a fictional novelty validated by logic, central to science fiction because it implies a change in the entire universe of the narrative, or at least in certain crucial aspects of it. Angenot [2] refines this definition by distinguishing between two categories of novums: those that anticipate a future state of language, and those that belong to “exolinguistics,” i.e., an extraterrestrial or parallel language. Saint-Gelais [13] uses the term “xenoencyclopedia” to emphasize the specificity of these imaginary referents, which gradually become integrated into the reader’s mental universe.

The work of Gindre [7] in lexicology and Ray [12] in translation studies has highlighted the complexity of these invented expressions: although their formation is based on lexical matrices comparable to those of the real lexicon, their use in science fiction is characterized by novel combinations and semantic networks specific to the genre.

Furthermore, unlike traditional named entities (people, places, organizations), novums do not belong to a closed inventory: their semantic role and syntactic position vary depending on the narrative. Their detection cannot therefore be reduced to a simple extension of existing NER classes (PER for character, LOC for location, ORG for organization, and MISC for miscellaneous), but requires the addition of a dedicated class (NOV for novum), capable of capturing entities that are both lexical (neologisms) and cultural (conveying a technological or social imaginary).

2.2 Named entity recognition for literary corpus

While defining novums is a lexicological challenge, their automatic detection can be integrated into a Named Entity Recognition (NER) task applied to literary corpora. Two recent tools, Propp [11] and Renard [1], stand out for their adaptation to the specificities of literary texts, but neither addresses the issue of novums.

Propp [11] is a French adaptation of BookNLP [3], a pipeline designed for the analysis of literary texts in English. Propp offers a modular processing chain, including entity recognition and

coreference resolution, with the aim of making distant reading techniques accessible to a French-speaking audience. Performance is solid for classic classes (PER: F1 = 88.4%; LOC: F1 = 56.8%), and the tool has proven useful for tasks such as literary subgenre classification, providing more interpretable features than a bag-of-words approach. However, using Propp for a NER task requires co-reference annotations, which are absent from the FrenchSFcorpus, and focuses on pronouns and co-reference chains in addition to entities.

Renard [1] is a Python library designed to extract character networks from narrative texts, with the unique feature of allowing the extraction of dynamic (evolving) networks in addition to static networks. Unlike most existing tools, Renard offers a modular architecture: each NLP subtask (entity recognition, coreference resolution, relation extraction) can be customized, allowing the pipeline to be adapted to specific types of text. After fine-tuning, performance is high for traditional named entities (PER: F1 = 83.5%; LOC: F1 = 81.8%).

However, like Propp, Renard does not support novums, which are neither proper nouns, standard common nouns, nor classic descriptors. Thus, although these two models are effective for general literary NER and character network extraction, they cannot identify novums. Adapting them to our corpus would require not only fine-tuning but also the addition of a dedicated NOV class, which is absent from existing annotation schemes. This is why it is necessary to develop a specific model, given that no model currently exists that performs this task.

3 Novum and named entity recognition

3.1 Dataset

Since novum detection is formulated as an entity recognition task, it incorporates the standard classes PER, LOC, ORG, and MISC, as well as a specific class NOV. The PER, LOC, ORG, and MISC entities were automatically annotated using the Renard model [10]. Novum annotation, on the other hand, is performed semi-automatically: all novums were identified manually for each narrative, then annotated automatically in BIO format. Here is the specific manual annotation methodology:

- Phrases describing novum are annotated, unlike periphrases, even though they are also used to for descriptions. Thus, *almanach électrique* is a novum because it forms an autonomous lexical unit, whereas a description such as “Finally, to protect the hull itself by any means necessary from potential pressure from the ice, it was encased in a steel cradle whose bands, connected by a series of St. Andrew’s crosses and stretched through cross-shaped guides, rested on iron beams embedded in a wooden frame. Thus supported, the ship had nothing to fear from pressure on its keel or sides”² (*Une Française au Pôle nord*, Maël, 1893) is not considered a lexicalized expression, even though it describes a fictional object (a cradle protecting a ship from ice pressure).
- An expression is considered non-lexicalized when it is absent from the French version of Wiktionary³, a dictionary chosen for its broad coverage—including neologisms and terms not yet included in official dictionaries, as well as for its use in projects involving the processing of polylexical expressions [8][15]. Exceptions are terms found in Wiktionary whose etymology attests to a fictional origin corresponding to the story’s publication date, or when the definition is labeled “Science Fiction,” as with the entry *ferromagnétal*⁴.

² Personal translation of “Enfin, pour préserver par tous les moyens la coque elle-même de la pression éventuelle des glaces, on l’enveloppa d’un berceau d’acier dont les bandes, reliées entre elles par une série de croix de Saint-André et étirées dans des filières en croix, reposaient sur des poutrelles également de fer, encastrées dans une gangue de bois. Ainsi soutenu, le navire ne devait rien redouter de la pression sur sa quille ou ses flancs.”

³ <https://fr.wiktionary.org>

⁴ Etymology section for *ferromagnétal* in Wiktionary: “1910. Création de de J.-H. Rosny aîné. De ferromagnétique.”

- Terms whose form has been lexicalized but whose semantics are entirely new compared to attested definitions are also considered novums (for example, *tube* referring to an underground tunnel connecting continents in *Le vingtième siècle*, Robida, 1883).
- Metaphors, expressions (such as “se mettre sur son trente-et-un”), wordplay, and other figures of speech do not qualify as novum, nor do words belonging to a language other than French (including invented languages).
- Proper nouns of fictional characters, places, or organizations, as they refer to named entities rather than concepts or objects used to construct the fictional world are not novums [12]. Thus, *ville aérienne* is a novum, but the name given to that city is not.
- Adjectives and nouns derived from proper nouns constitute novum when they are formed by derivation and denote a concept or characteristic of the fictional world [12], such as *andrëidien*.
- When an adjective alone represents a novum, two units are annotated separately: the adjective alone on the one hand, and the adjectival noun phrase on the other, the latter forming a novum in its own right. This dual annotation is justified by the fact that, depending on the noun to which it refers, a single adjective can denote distinct fictional referents: thus, the adjective *napusien* associated with *épidémie* forms the novum *épidémie napusienne*, denoting a fictional pathological phenomenon, whereas *savate napusienne* denotes an entirely different fictional object. The adjective alone thus constitutes a novum as a derivational unit carrying a fictional meaning, and each adjectival noun phrase constitutes an additional and distinct novum whose referent is determined by the combination of the two units.
- Novums are systematically inflected in the singular in the metadata to enable the automatic annotation of all their inflected occurrences in the texts.

PER	LOC	ORG	NOV	MISC
Sir Bucephalus	Londres	la Sûreté	cinébouquin	deux Canayens
Le docteur Flax	Le National Gallery	gouvernement suisse	anthroposaure	un Américain
Samuel Sun	vallées de la Somme	l’Empire de l’Espace	homme lunaire	des Arabes
monsieur Burton	au jardin des Plantes	le Secours universel	hypermagnétos	les Martiens
M. Héricourt	Bâbord-Harbour	la marine royale	pain de houille	le Suédois

Table 1: Examples of entities for each class.

A sample of 1,000 sentences was compiled to evaluate the quality of the annotations: 500 sentences containing at least one NOV annotation to measure precision, and 500 sentences without a NOV annotation to estimate recall. In the first subsample, manual evaluation identified 11 errors out of 559 annotated fiction words, resulting in a precision of 98.0%. These errors stem primarily from categorical ambiguities not resolved by the automatic system, as in the case of *être artificiel*, where the word “être” can be either a noun or a verb (“c’est un être artificiel” versus “son sourire est artificiel”). In the second subsample, no errors were found, bringing the recall to 100%. Overall, the pipeline therefore demonstrates very high performance, with accuracy errors stemming from the limitations of automatic processing (notably the lack of part-of-speech consideration).

The different entity classes contain heterogeneous distributional properties, illustrated by the examples in Table 1. PER, LOC, and ORG entities correspond mainly to stable references, often proper names or lexically fixed noun phrases, although they may include determiners or titles (Dr. Flax or Sir Bucephalus).

The NOV class is clearly distinguished by its high syntactic variability. Novums can appear in the form of simple nouns (*cinébouquin*), compounds (*anthroposaure*, *hypermagnétos*), or noun phrases (*homme lunaire*, *pain de houille*). Even verbal phrases can be novums (*napusifier* or *signer électriquement*). No strong morphosyntactic constraints can therefore be imposed on this class.

Finally, the MISC class mainly includes designations of peoples, human groups, or non-individualized characters. This choice is inherited from the initial annotation scheme of Renard’s model, in which these mentions did not fall under strict named entities (PER) or organizational or geographical categories, but corresponded to collective or undefined referents, with annotators deciding to place them in this category.

This heterogeneity between classes poses a challenge for NER models based on surface indices and motivates the use of contextual models capable of integrating finer discursive and semantic regularities, such as transformers.

After annotation, the corpus comprises a total of 601,267 sentences and approximately 300,000 entity occurrences, divided into train, dev, and test sets according to an 80/10/10 split, with the additional constraint that all occurrences of a novum must be found in only one of the three sets in order to force the models to generalize. Thus, the train set contains 481,013 sentences, the dev set 60,126 sentences, and the test set 60,128 sentences.

# of occurrences...	in train set	in dev set	in test set	in total
PER	155 642	18 849	18 749	193 240
LOC	66 261	7 647	7 642	81 550
ORG	5 711	705	667	7 083
NOV	1 764	217	227	2 208
MISC	11 628	1 290	1 377	14 295
Total	241 006	28 708	28 662	298 376

Table 2: Number of entity occurrences.

As shown in Table 2, the distribution of entities is highly unbalanced: the NOV class represents less than 1% of annotated entities. In order to ensure sufficient exposure of the model to this class during training, we use two complementary strategies:

- Oversampling of the training set: sentences containing at least one novum are duplicated with a factor of 10. The training set thus increases from 1,764 occurrences of NOV to 17,640, while the dev and test sets retain their natural distribution.
- Focal loss during training, in order to reduce the influence of easy and majority examples (PER, LOC) and focus learning on errors concerning rare classes, in particular NOV.

3.2 Transformers for novums recognition

We experimented with alternative approaches including novelty detection, rule-based candidate filtering, and binary token classification. These proved inadequate due to: (1) mismatch between sentence-level anomaly detection and span-level novum identification; (2) inability of lexical rules to distinguish novums from OCR errors, dialectal variants, and onomatopoeia without unwieldy exception lists; and (3) severe performance degradation under extreme class imbalance (<1% NOV class), with F1 scores reaching only 16%. The NER formulation resolves these issues through sequence-aware contextual modeling, implicit handling of non-contiguous spans via BIO tagging, and improved class balance through multi-class supervision.

To perform entity and novum recognition, CamemBERT-base⁵ is used for supervised fine-tuning for a multi-class sequence labeling task. The hyperparameters are optimized by Bayesian optimization over 20 trials, exploring the learning rate, batch size, warm-up ratio, weight decay, and focal loss parameters (gamma, and alpha corresponds to the weights of each label). After studying the overfitting of the model, the number of epochs was set in advance to avoid excessive computation time. The best hyperparameters are reported in the appendix.

Optimization focuses on the macro F1-score at the entity level. Unlike micro F1-score, whose result is largely dominated by the majority classes, macro F1-score puts all classes on the same level, regardless of the number of entities they contain. It therefore makes it possible to evaluate the model’s ability to detect rare and emerging entities. Although this choice tends to penalize the majority classes, the points lost on PER and LOC are largely offset by those gained by the minority classes.

4 Results

	Precision	Recall	F1-score
PER	92.00	94.03	93.00
LOC	80.73	85.55	83.07
ORG	68.66	71.91	70.25
NOV	67.02	60.29	63.48
MISC	83.00	87.23	85.06
micro F1-score	87.90	90.77	89.31
macro F1-score	78.28	79.80	78.97

Table 3: Transformer results by class for the NER task.

The results obtained for the NER task on the FrenchSFcorpus corpus reveal very solid performance for traditional classes, with a micro F1-score of 89.31% and a macro F1-score of 78.97%. The PER (F1 = 93.00%), LOC (F1 = 83.07%) and MISC (F1 = 85.06%) classes achieved high scores, reflecting the models’ ability to detect characters, groups of characters, and peoples. In contrast, the ORG class (F1 = 70.25%) showed good performance, and the NOV class obtains the worst but good results per class with F1 = 63.48%.

As these two classes are the rarest (5,711 in the train set and 17,640 after the factor 10 data augmentation, respectively), these results are encouraging and suggest that a larger corpus would lead to higher results. The diversity of forms that novums can take and their ambiguity also prevent this class from achieving the same performance as PER or LOC.

5 Discussion

The task of recognizing novums differs from traditional NER approaches in several major ways. First, the scarcity of examples in annotated corpora limits the model’s ability to generalize. With only 2,208 occurrences of NOV across all sets, compared to tens of thousands for other classes, the model suffers from a lack of training data, resulting in lower recall. Besides the low number of NOV occurrences, it is also the only class that has only different occurrences across the three sets to force the model to generalize this class. In addition, novums are characterized by a highly varied distribution: some are common nouns, others are compound terms, morphological neologisms,

⁵ <https://huggingface.co/almanach/camembert-base>

lexical creations from fictional languages, etc. This diversity makes it difficult to define stable lexical or syntactic patterns, unlike traditional named entities.

Another challenge lies in the semantic and contextual ambiguity of novums. Unlike entities such as PER or LOC, which rely on stable referents and clear linguistic markers (capital letters, determiners, etc.), novums can be confused with common nouns or technical terms, especially in a genre such as science fiction where the boundary between reality and fiction is often blurred. For example, a term such as “hypermagnetos” could be interpreted as a technical common noun, while “anthroposaure” could be confused with an existing academic name of an existing species by a transformer. This ambiguity requires the model to have a detailed understanding of the narrative context, which exceeds the capabilities of traditional NER approaches, even those adapted to literary corpora.

6 Conclusion and future works

This model presents a first systematic approach for the automatic recognition of novums in French-language science fiction texts, integrating a dedicated NOV class into a CamemBERT-based NER model. The results obtained, with an F1-score of 63.48% for the NOV class and an overall macro F1-score of 78.97%, demonstrate the feasibility of the task, despite the challenges posed by the rarity of examples, their morphosyntactic variability, and their semantic ambiguity. These encouraging results for a novel and complex class confirm the value of adapting NER tools to the specificities of literary corpora, and more particularly to the lexical creativity of science fiction.

In addition to the value of automatically detecting novums, this task opens up several perspectives for the literary analysis of science fiction narratives. It allows us to study the circulation of novums across texts, authors, and periods, and to identify patterns in their reuse or transformation. Such an approach also allows us to study how these novums are influenced by real-world technological advancements (such as the arrival of electricity in homes, followed by the advent of robotics or artificial intelligence, for example).

In terms of the more immediate perspectives, the FrenchSFcorpus corpus could be enriched by automatically annotating new science fiction stories, thereby significantly increasing the number of examples for the NOV class and improving the robustness of the model. Exploring data augmentation techniques adapted to minority classes could also help to better capture the diversity of novums.

References

- [1] Amalvy, A., Labatut, V., and Dufour, R. “Renard: A Modular Pipeline for Extracting Character Networks from Narrative Texts”. In: *Journal of Open Source Software* 9, no. 98 (2024), p. 6574. DOI: 10.21105/joss.06574.
- [2] Angenot, Marc. *La Parole Pamphlétaire. Contribution à la typologie des discours modernes*. Payot, 1982.
- [3] Bamman, David, Popat, Sejal, and Shen, Sheng. “An annotated dataset of literary entities”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, ed. by Jill Burstein, Christy Doran, and Tamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 2138–2144. DOI: 10.18653/v1/N19-1220. URL: <https://aclanthology.org/N19-1220/>.
- [4] Bérard, Sylvie. *La science-fiction : de la perspective à la prospective*. Vol. 8. Les Cahiers de la CRSDD, 2012.

- [5] Bréan, Simon. *La Science-fiction en France : Théorie et histoire d'une littérature*. Presses Université Paris-Sorbonne, 2012.
- [6] Eco, Umberto. *Lector in fabula*. Librairie générale française, 1985.
- [7] Gindre, Philippe. *La formation des néologismes dans la littérature de science-fiction d'expression anglaise contemporaine*. PhD thesis. Université de Franche Comté, 1999.
- [8] Krstev, Cvetana, Stanković, Ranka, Marković, Aleksandra, and Ikonić Nešić, Milica. "Progress in SR-ELEXIS Semantic Annotation: Focusing on Multiword Expressions, Named Entities, and Sense Repository". In: *3rd General Meeting Hungarian Research Centre for Linguistics, Budapest, Hungary 29-30 January 2025*. European Cooperation in Science and Technology. 2025.
- [9] Langlet, Irène. *La science-fiction, lecture et poétique d'un genre littéraire*. Armand Colin, 2006.
- [10] Maurel, Perrine, Amalvy, Arthur, Labatut, Vincent, and Alrahabi, Motasem. "Du repérage à l'analyse : un modèle pour la reconnaissance d'entités nommées dans les textes littéraires en français". In: *Digital Humanities 2025*. Lisbonne, Portugal, July 2025. URL: <https://hal.science/hal-05128746>.
- [11] Mélanie-Becquet, Frédérique, Barré, Jean, Seminck, Olga, Plancq, Clément, Naguib, Marco, Pastor, Martial, and Poibea, Thierry. "BookNLP-fr, the French Versant of BookNLP. A Tailored Pipeline for 19th and 20th Century French Literature". In: *Journal of Computational Literary Studies* 3 (1 Nov. 2024), pp. 1–34. ISSN: 2940-1348. DOI: 10.48694/jcls.3924. URL: <https://jcls.io/article/id/3924/>.
- [12] Ray, Alice. *Traduire les termes du futur : Analyse du traitement des termes-fictions dans la traduction de l'anglais au français de la littérature de science-fiction*. PhD thesis. Université d'Orléans, 2019.
- [13] Saint-Gelais, Richard. *L'Empire du pseudo. Modernités de la science-fiction*. Nota Bene, 2005.
- [14] Suvin, Darko. *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre*. Yale University Press, 1979.
- [15] Überrück-Fries, Till, Savary, Agata, and Dryjańska, Agnieszka. "Sailing through multiword expression identification with Wiktionary and Linguse: A case study of language learning". In: *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*. 2024, pp. 248–262.
- [16] Vas-Deyres, Natacha. *Ces Français qui ont écrit demain : utopie, anticipation et science-fiction au XXe siècle*. Honoré Champion, 2013.

A Appendix

Number of epochs	30
Learning rate	3e-5
Batch_size	32
Max length	64
Weight decay	0.01
Warm-up ratio	0
Gamma	2.15
Alpha B-PER	10
Alpha I-PER	15
Alpha B-LOC	20
Alpha I-LOC	15
Alpha B-ORG	150
Alpha I-ORG	90
Alpha B-NOV	480
Alpha I-NOV	440
Alpha B-MISC	40
Alpha I-MISC	65
Alpha O	1

Table 4: Best hyperparameters, used to train de final model.