astroECR : enrichissement d'un corpus astrophysique en entités nommées, coréférences et relations sémantiques

Atilla Kaan Alkan, Felix Grezes, Cyril Grouin, Fabian Schüssler, Pierre Zweigenbaum

astroECR : enrichissement d'un corpus astrophysique en entités nommées, coréférences et relations sémantiques

Atilla Kaan Alkan[1,2], Felix Grezes[3], Cyril Grouin[1],
Fabian Schüssler[2], Pierre Zweigenbaum[1]
(1) Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique, 91405, Orsay, France.
(2) IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France.
(3) Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
{atilla.alkan, cyril.grouin, pz}@lisn.upsaclay.fr,
fabian.schussler@cea.fr, felix.grezes@cfa.harvard.edu

RÉSUMÉ

Le manque de ressources annotées constitue un défi majeur pour le traitement automatique de la langue en astrophysique. Afin de combler cette lacune, nous présentons astroECR, une extension du corpus TDAC (Time-Domain Astrophysics Corpus). Notre corpus, constitué de 300 rapports d'observation en anglais, étend le schéma d'annotation initial de TDAC en introduisant cinq classes d'entités nommées supplémentaires spécifiques à l'astrophysique. Nous avons enrichi les annotations en incluant les coréférences, les relations sémantiques entre les objets célestes et leurs propriétés physiques, ainsi qu'en normalisant les noms d'objets célestes via des bases de données astronomiques. L'utilité de notre corpus est démontrée en fournissant des scores de référence à travers quatre tâches : la reconnaissance d'entités nommées, la résolution de coréférences, la détection de relations, et la normalisation des noms d'objets célestes. Nous mettons à disposition le corpus ainsi que son guide d'annotation, les codes sources, et les modèles associés.

ABSTRACT

astroECR : an Enriched Corpus for Astrophysical Entities, Coreferences, and Relations

The lack of annotated resources poses a significant challenge for natural language processing in astrophysics. To address this gap, we introduce astroECR, an extension of the TDAC (Time-Domain Astrophysics Corpus). This corpus, comprised of 300 observation reports in English, expands the initial annotation scheme of TDAC by introducing five additional named entity classes specific to astrophysics. We enhanced annotations to include coreferences, semantic relations between celestial objects and their physical properties, and normalization of celestial object names using astronomical databases. We demonstrate our corpus's utility by providing baseline scores across four tasks : named entity recognition, coreference resolution, relation detection, and normalization of celestial object names. We provide the corpus, annotation guide, source code, and associated models to the community.

MOTS-CLÉS : Annotation de corpus, Extraction d'information, Astrophysique.

KEYWORDS: Corpus Annotation, Information Extraction, Astrophysics.

1 Introduction

Ces dernières années, le besoin de développer des systèmes d'analyse et d'extraction d'information en astrophysique a engendré une multiplication des travaux en Traitement Automatique des Langues (TAL). Les récents modèles de langue tels que astroBERT (Grezes *et al.*, 2021) et astroLLaMa (Nguyen *et al.*, 2023) sont utilisés non seulement pour identifier dans la littérature des entités spécifiques au domaine (Grezes *et al.*, 2022), mais également pour l'extraction d'information essentielles telles que les coordonnées célestes ou les propriétés physiques mesurées lors de l'observation d'objets célestes (Sotnikov & Chaikova, 2023). Néanmoins, un défi notable persiste dans la disponibilité des ressources et la diversité des annotations. En effet, la plupart des corpus existants (Becker *et al.*, 2005; Hachey *et al.*, 2005; Murphy *et al.*, 2006) ne sont pas accessibles et servent uniquement à la détection d'entités nommées. Parmi les corpus disponibles, on compte celui de la campagne d'évaluation DEAL (Grezes *et al.*, 2022) et un second plus restreint, TDAC (Alkan *et al.*, 2022), axé sur l'observation des phénomènes transitoires tels que les explosions de supernovae et les sursauts gamma. Les deux corpus partagent les mêmes classes d'entités et se limitent à une annotation en entités nommées, laissant ainsi un manque dans la diversité des annotations. Or, une extraction d'information plus complète nécessite des corpus avec des annotations comprenant les coréférences, pour identifier toutes les mentions se référant à une même entité, ou encore les relations entre les différentes paires d'entités. Par exemple, comme illustré dans la figure 1, lorsqu'il y a mention de plusieurs objets célestes dans le texte, se limiter à annoter uniquement les entités nommées ne permet pas d'établir de manière précise les liens entre les différentes propriétés physiques et les objets célestes correspondants.

FIGURE 1 – Extrait d'un rapport d'observation. A gauche, un exemple d'annotation en entités nommées uniquement, et à droite, l'annotation des entités nommées avec en plus l'annotation des mentions de coréférences et des relations sémantiques entre les objets célestes (mentions de type CelestialObject) et leurs propriétés physiques.

Afin de combler cette lacune et faciliter un spectre plus large de recherches, notre travail vise à créer un corpus annoté englobant l'annotation d'entités nommées, de mentions de coréférences, de relations astrophysiques entre les corps célestes et leurs propriétés physiques, ainsi qu'en fournissant une normalisation des noms d'objets célestes. Nous avons pour objectif de fournir à la communauté astrophysique et TAL une ressource permettant le développement de modèles d'extraction d'information. Pour ce faire, nous avons étendu la première version de notre corpus existant TDAC, afin de construire astroECR, un nouveau corpus en astrophysique plus riche en annotations.

Les principales contributions de ce travail sont les suivantes :

— Nous avons augmenté la taille du corpus TDAC, en passant de 75 à 300 documents annotés. Cette augmentation comprend un ensemble plus complet d'annotations couvrant davantage d'entités nommées astrophysiques, la normalisation des noms d'objets célestes (liage référentiel), des annotations de coréférence et des relations astrophysiques. À notre connaissance, il s'agit de la première et unique ressource de ce genre dans le domaine ;

— Nous avons repris les catégories d'entités nommées du corpus TDAC et y avons ajouté cinq catégories d'entités nommées supplémentaires ;
— Nous démontrons l'utilité de ce corpus en réalisant des expériences sur quatre tâches d'extraction d'information, pour lesquelles nous avons développé des modèles et fourni de premiers scores. En perspective, ces modèles faciliteront l'annotation automatisée de documents supplémentaires ;
— Nous mettons notre corpus, notre guide d'annotation, le code associé et les modèles à la disposition de la communauté de recherche via notre dépôt GitHub [1].

2 Travaux connexes

Ressources pour la détection d'entités nommées La reconnaissance d'entités nommées implique l'identification de mentions d'entités, telles que des personnes, lieux ou organisations (Grishman & Sundheim, 1996). Il s'agit d'une tâche utile en recherche d'information (Yadav & Bethard, 2018; Banerjee *et al.*, 2019) et également pour les systèmes de question-réponse (Mollá Aliod *et al.*, 2006). En astrophysique, Becker *et al.* (2005); Hachey *et al.* (2005) ont créé l'Astronomy Bootstrapping Corpus (ABC), composé de 209 résumés d'articles scientifiques radioastronomiques en anglais pour la détection d'entités nommées spécifiques au domaine telles que les noms d'instruments astronomiques, les objets célestes, leurs types, et leurs caractéristiques spectrales. Cependant, ce corpus n'est pas accessible. Murphy *et al.* (2006) ont également élaboré un corpus de 7840 phrases issues d'articles scientifiques en anglais, définissant 43 types d'entités nommées, incluant des catégories caractérisant les objets célestes : leurs coordonnées et propriétés physiques (fréquence, luminosité). À notre connaissance, ce corpus n'est pas accessible non plus. Plus récemment, le corpus de la campagne d'évaluation DEAL (Grezes *et al.*, 2022) a été rendu public, devenant l'un des premiers corpus accessibles en astrophysique [2]. Il comprend des extraits de texte intégral et des sections de remerciements provenant d'articles d'astrophysique en anglais, annotés spécifiquement pour la campagne, avec 31 catégories d'entités nommées. Il est divisé en trois sous-ensembles : entraînement (1753 documents), développement (1366 documents) et test (2505 documents). Dans un précédent article, nous avions introduit TDAC (Alkan *et al.*, 2022), le seul corpus annoté en entités nommées construit à partir de rapports d'observation astronomique en anglais, se concentrant sur un vocabulaire spécifique à l'astronomie (l'étude des phénomènes transitoires). Accessible [3], il se compose de 75 documents, dont 25 circulaires du réseau GCN (Barthelmy *et al.*, 1995) de la NASA, 25 télégrammes astronomiques (Rutledge, 1998) et 25 AstroNotes issus du Transient Name Server (Gal-Yam, 2021).

Ressources pour la résolution des coréférences La résolution de coréférences est une tâche visant à identifier toutes les mentions dans un texte se référant à une même entité (Jurafsky & Martin, 2023; Zheng *et al.*, 2011). Comparée à la détection d'entités nommées, la tâche de résolution de coréférences dans les documents en astrophysique a reçu moins d'attention. Kim & Webber (2006) se sont penchés sur la résolution d'anaphores dans la littérature astrophysique. Les auteurs se sont uniquement concentrés sur la classification automatique du pronom "*they*" dans les articles, en distinguant ceux qui se réfèrent à des recherches citées et ceux qui ne le font pas. Leur système repose sur un classificateur d'entropie maximale avec des caractéristiques basées sur la distance

1. https://github.com/AtillaKaanAlkan/astroECR
2. https://huggingface.co/datasets/adsabs/WIESP2022-NER/
3. https://github.com/AtillaKaanAlkan/TDAC

entre les citations précédentes et les types de verbes associés au pronom en question. Brack *et al.* (2021) ont construit le corpus STM pour la résolution de coréférences. Le corpus se compose de dix disciplines scientifiques (dont onze résumés annotés en astrophysique). Les auteurs ont comparés plusieurs approches existantes dans la littérature, avec notamment l'utilisation de modèles de type BERT pour la résolution des coréférences (Joshi *et al.*, 2019), mais également en s'inspirant de la méthode proposée par Luan *et al.* (2018) via l'utilisation de représentations de mots ELMo (Peters *et al.*, 2018).

Synthèse Les corpus annotés pour le domaine astrophysique sont limités. Ces ressources se concentrent principalement sur des corpus orientés reconnaissance d'entités nommées, limitant leur usage pour des tâches plus larges en TAL. De plus, les documents concernés sont principalement des articles, restreignant la variété des sources de données que les chercheurs peuvent exploiter. Pour combler cette lacune, nous avons basé notre travail sur le corpus existant TDAC, pour construire un corpus plus riche et l'étendre à des tâches de TAL non traitées telles que la résolution de coréférences, la détection de relations astrophysiques et la normalisation des noms d'objets célestes.

3 Annotation du corpus

Dans cette section, nous décrivons le processus d'annotation des entités nommées (3.1), la normalisation des noms d'objets célestes (3.2), les coréférences (3.3) et les annotations des relations astrophysiques (3.4). Nous avons utilisé BRAT (Stenetorp *et al.*, 2012) comme outil d'annotation.

3.1 Extension des classes d'entités nommées et annotation

Nous avons adopté le guide d'annotation de TDAC en proposant une extension du schéma d'annotation avec cinq catégories d'entités supplémentaires jugées essentielles par les astronomes.

- **`Date :`** Dates et expressions temporelles se référant à une date de détection ou à la durée d'une observation. Exemple : *We report the discovery of a probable nova in M31 on a co-added 990-s R-band CCD frame taken under poor conditions on* ***2019 Mar. 12.791 UT*** $_{[Date]}$ *with the 0.65-m telescope at Ondrejov.*
- **`Reference :`** Références vers d'autres rapports d'observation, utiles pour répérer et regrouper tous les rapports concernant un même objet céleste. Exemple : *In comparison to the optical region (ref : the SALT spectrum in* ***ATel #3289*** $_{[Reference]}$*), few strong NI lines are expected in the JHK bands.*
- **`Magnitude :`** Equations et valeurs numériques qui caractérisent la luminosité des corps célestes (propriété utile pour les astronomes afin de déterminer la visibilité des objets). Exemple : *As reported to CBAT, this nearby-M31 object was discovered by Koichi Itagaki at* ***16.5 mag.*** $_{[Magnitude]}$
- **`Flux :`** Valeur numérique caractérisant l'énergie d'un corps céleste. Exemple : *The flux values ranged from* ***1.01 +/-0.06 E+11 cgs*** $_{[Flux]}$ *to* ***1.71 +/- 0.04 E+11 cgs.*** $_{[Flux]}$
- **`Redshift :`** Equations et valeurs numériques caractérisant la distance d'un corps céleste par rapport à un observateur. Exemple : *The host KUG 0180+227 is an E+A galaxy at* ***z=0.022.*** $_{[Redshift]}$

3.2 Normalisation des mentions de type CelestialObject

La normalisation consiste à désambiguïser les mentions d'entités en les reliant à leur entrée respective dans des bases de connaissances (Sevgili *et al.*, 2020). Nous normalisons les noms d'objets célestes du corpus, tels que les supernova, les sursauts gamma, et les galaxies, à leurs entrées spécifiques dans les catalogues astronomiques. En égard aux différentes conventions de dénomination des objets célestes en astronomie, cette normalisation est essentielle pour l'intégration de données provenant de divers articles et rapports d'observation. Par exemple, la galaxie d'Andromède [4] a au moins 39 désignations, chacune devant être correctement associée. Nous utilisons pour cela trois catalogues astronomiques complémentaires : SIMBAD (Wenger *et al.*, 2000), NED (Mazzarella *et al.*, 2001), et TNS (Gal-Yam, 2021).

3.3 Périmètre d'annotation des coréférences

Notre guide d'annotation détaillé est accessible via notre dépôt GitHub. Ici, nous donnons un aperçu général de nos choix d'annotation des coréférences. Nous avons également annoté les cas où un objet céleste est désigné par un autre de ses noms dans le texte. Nous avons exclu les expressions mathématiques, les quantités numériques et d'autres relations coréférentielles non associées à un objet céleste de notre schéma d'annotation. Pour clarifier cette distinction, examinons les exemples suivants :

- Coréférences annotées (une même couleur marque les éléments d'une chaîne de coréférence) :
  - *We discovered **PS19did** on MJD 58666.31 = 2019-07-02.31, at w=19.9 +/- 0.1 [...] **The new transient source** is in the galaxy **UGC 11003** [...] Adopting **the host galaxy** redshift z=0.03566 (NED) yields an expansion velocity [...] Followup observations of **this intrinsically faint transient** are encouraged.*
  - *We report on the discovery and follow-up of a very bright and highly magnified microlensing event **Gaia19bld**. [...] **It** has been detected and announced by the Gaia Science Alerts program.*
  - *We report on the NIR brightening of the intermediate redshift quasar **PKS0735+17** (z=0.424), also known as **CGRaBSJ04738+1742**.*
- Coréférences exclues du processus d'annotation :
  - *Analysis of **the data** is ongoing. We remind the community that all **Swift data** are public, and encourage **their** use.*
  - ***The observations** continued until 2019-04-26 20 :15 UT, when **they** were aborted to begin followup of.*
  - *The estimated AB magnitude is **17.6**. **This magnitude** is not corrected for the host galaxy contribution.*

4. http://simbad.cds.unistra.fr/simbad/sim-id?Ident=Andromeda+Galaxy&NbIdent=1&Radius=2&Radius.unit=arcmin&submit=submit+id

3.4 Annotation des relations astrophysiques entre les mentions de type CelestialObject et leurs propriétés physiques

La détection de relation vise à établir les liens entre paires d'entités (Bassignana & Plank, 2022). Dans le cadre de notre étude, nous définissons un unique type de relation, qui relie un objet céleste à ses propriétés physiques. Dans le cas où plusieurs objets célestes sont mentionnés dans le texte, notre objectif est donc de pouvoir associer les propriétés physiques mentionnées à l'objet correspondant. Dans notre schéma d'annotation donc, seules les mentions de type `CelestialObject` sont reliées aux mentions d'entités décrivant des attributs physiques tels que `CelestialRegion` (coordonnées dans le ciel), `Flux` (énergie du corps par unité de temps), `Magnitude` (intensité lumineuse de l'objet céleste) etc.

4 Statistiques du corpus annoté

Dans cette section, nous décrivons les principales caractéristiques du corpus annoté résultant (statistiques globales et accord inter-annotateurs), et nous fournissons des tableaux comparatifs entre le corpus TDAC et notre nouveau corpus annoté astroECR.

4.1 Statistiques globales

Paramètres	**TDAC**		**astroECR**	
	Entraînement	**Test**	**Entraînement**	**Test**
# documents	59	16	210	90
# tokens	15374	3638	43481	10578
# tokens annotés	4338	1014	17392	3173
# mentions de coréférence	-	-	412	101
# chaînes de coréférence	-	-	257	65
Long. moyenne des chaînes	-	-	3,5 (+/- 2,26)	3,4 (+/- 1,61)
# relations intra-phrases	-	-	490	143
# relations inter-phrases	-	-	154	26
# total de relations	-	-	644	169

TABLE 1 – Statistiques de comparaison entre les corpus TDAC et astroECR.

La Figure 2 illustre la distribution des mentions d'entités nommées entre les corpus TDAC et astroECR. À l'issue de l'annotation, les types d'entités les plus représentées sont les classes spécifiques au domaine astrophysique telles que : les noms d'objets célestes (`CelestialObject`), les mentions de type `Magnitude` ou encore des mentions relatives à des outils d'observations tels que les télescopes (`Telescope`) et instruments (`Instrument`). Nous remarquons que les classes telles que `Software`, `Grant`, `Collaboration` ou encore `Archive` sont moins fréquentes dans les rapports d'observations.

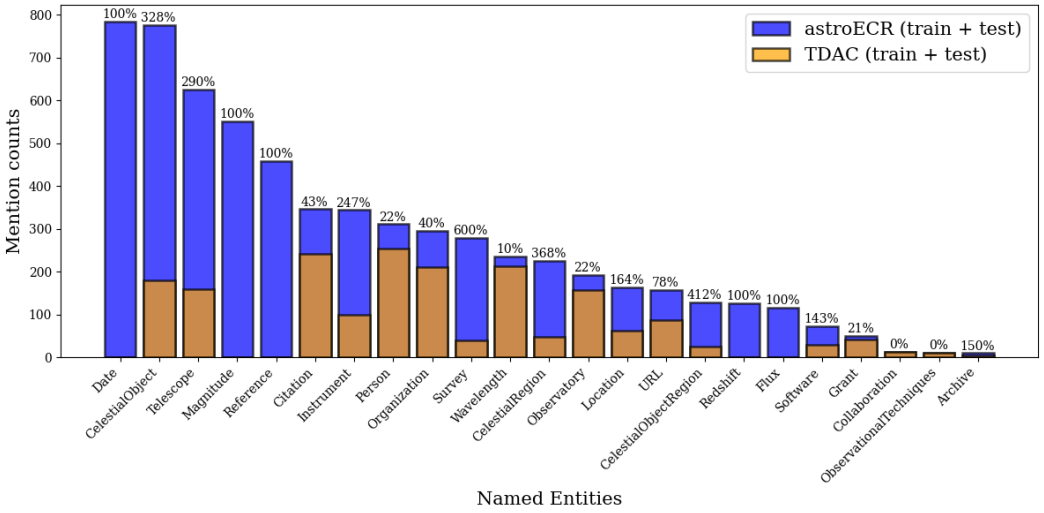


FIGURE 2 – Distribution des mentions d'entités nommées dans les corpus TDAC (en jaune) et astroECR (en bleu).

4.2 Accord inter-annotateurs et consensus

Nous avons impliqué un expert en astrophysique et un expert en TAL pour annoter un sous-ensemble du corpus (30 documents, soit 6499 unités lexicales). Les erreurs d'annotation ont été identifiées et corrigées au cours d'une phase de consensus, permettant la création du jeu de référence. Nous avons ensuite comparé les annotations des deux annotateurs avec le jeu de référence produit en utilisant les scores de précision, de rappel et de F-mesure, conformément à la méthodologie de Galibert *et al.* (2012). Les résultats du tableau 2 montrent que l'expert en astrophysique a obtenu une F-mesure plus élevée (0,94) que l'expert en TAL (0,91) par rapport au consensus (en évaluation souple), autorisant la poursuite de l'annotation par l'expert en astrophysique seul sur les 270 documents restants.

Tâche	**Annotateurs**	**Stricte**			**Souple**		
		P	**R**	**F1**	**P**	**R**	**F1**
Entités nommées	Astro vs. TAL	0,65	0,59	0,62	0,84	0,92	0.88
	Astro vs. consensus	0,83	0,86	0,84	0,93	0,96	0,94
	TAL vs. consensus	0,73	0,69	0,71	0,94	0,89	0,91
Coréférences	Astro vs. TAL	0,77	0,88	0,82	0,78	0,89	0,83
	Astro vs. consensus	0,97	1,00	0,98	0,97	1,00	0,98
	TAL vs. consensus	0,74	0,89	0,81	0,75	0,90	0,82

TABLE 2 – Accord inter-annotateurs pour l'annotation des entités nommées et des mentions de coréférences entre les deux annotateurs, et comparaison avec le consensus. L'annotateur astrophysicien est dénommé "*Astro*", et l'expert en TAL est dénommé "*TAL*". Les métriques utilisées sont la Précision (P), le Rappel (R) et la F-mesure (F1). Deux modes d'évaluation : stricte et souple. En évaluation stricte, une entité annotée est considérée comme vrai positive si le type d'entité et les frontières sont correctement annotées. En évaluation souple, les frontières d'annotation ne sont pas pénalisées.

5 Expériences

5.1 Configurations expérimentales

— **Reconnaissance d'entités nommées** : Nous avons ajouté et entraîné une tête de classification aux modèles astroBERT (Grezes *et al.*, 2021) et SciBERT (Beltagy *et al.*, 2019) sur le corpus astroECR$_{\text{train}}$, puis les avons évalués sur l'ensemble de test astroECR$_{\text{test}}$. L'entraînement a été effectué sur 20 époques, avec un taux d'apprentissage $\alpha = 2,10^{-5}$, et une taille de lot d'entraînement de 4. L'entraînement a été réitéré 5 fois avec des amorces différentes.

— **Normalisation des mentions de type `CelestialObject`** : Notre système interroge d'abord la base de données SIMBAD (Wenger *et al.*, 2000) avec des requêtes ADQL[5], extrayant l' identifiant unique de chaque objet céleste, sa désignation canonique, ainsi qu'une liste de toute ses désignations. Si une source n'est pas identifiée dans la base SIMBAD, la requête s'étent à la base NED (Mazzarella *et al.*, 2001) et, si nécessaire, à la base TNS (Gal-Yam, 2021).

— **Résolution des coréférences** : Nous avons évalué F-coref (Otmazgin *et al.*, 2022), un outil de résolution des coréférences basé sur l'architecture LingMess (Otmazgin *et al.*, 2023). Nous avons choisi F-coref en raison de sa facilité d'utilisation via sa bibliothèque Python *fastcoref*[6]. Nous avons procédé à une première évaluation du modèle sans entraînement en comparant ses prédictions avec nos annotations. Ensuite, nous avons entraîné le modèle sur 50 époques en utilisant astroECR$_{\text{train}}$ et l'avons évalué sur astroECR$_{\text{test}}$. Chaque expérience a été répétée cinq fois avec différentes des amorces aléatoires.

— **Détection de relations** : Nous avons entraîné un réseau de neurones de type biLSTM, que nous avons affiné sur l'ensemble d'entraînement pendant 20 époques avec un taux d'apprentissage $\alpha = 10^{-3}$, et une taille de lot d'entraînement fixée à 128. Nous avons évalué les performances du système sur astroECR$_{\text{train}}$.

5.2 Résultats sur le corpus de test astroECR$_{\text{test}}$

Dans cette section, nous présentons et analysons les résultats obtenus sur astroECR$_{\text{test}}$, l'ensemble de test d'astroECR, à l'issue de l'entraînement sur astroECR$_{\text{train}}$, son corpus d'entraînement.

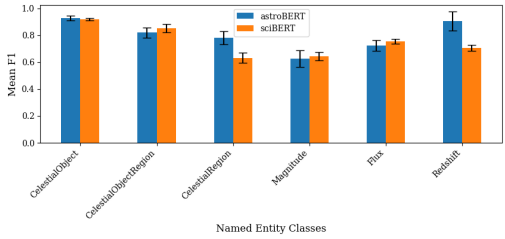
Détection d'entités nommées Le tableau 3 met en évidence la supériorité d'astroBERT par rapport à SciBERT en matière de rappel et F-mesure. En effet, une différence de 6 points sur la F-mesure globale (0,76 pour SciBERT, contre 0,82 avec astroBERT) est constatée. Une analyse plus détaillée des performances pour certaines classes spécifiques au domaine (figure 3a) montre qu'astroBERT est particulièrement plus performant dans la reconnaissance d'entités spécifiques à l'astrophysique, notamment le repérage des coordonnées célestes (`CelestialRegion`), caractérisée par une diversité de formes, ou encore pour la classe `Redshift`, de nature équationnelle. Ces résultats peuvent être attribués au fait qu'astroBERT est un modèle de langue pré-entraîné sur des textes spécifiques au domaine, ce qui renforce sa capacité à mieux repérer ces types d'entités.

5. Le language ADQL est basé sur le langage SQL, avec quelques extensions pour prendre en charge des requêtes spécifiques à l'astronomie, notamment pour des requêtes sur les coordonnées célestes.

6. https://pypi.org/project/fastcoref/

Modèle	Précision	Rappel	F-mesure
SciBERT	0,84 (0,01)	0,70 (0,01)	0,76 (0,01)
astroBERT	0,83 (0,01)	0,81 (0,01)	0,82 (0,01)

TABLE 3 – Performance moyenne (avec écart-type) des systèmes de REN fondés sur SciBERT et astroBERT. Les modèles ont subi cinq entraînements distincts avec diverses amorces sur l'ensemble d'entraînement d'astroECR, puis ont été évalués sur le jeu de test d'astroECR. Les métriques utilisées sont la Précision, le Rappel, et la F-mesure.



(a) F-mesure moyenne (avec écart-type) de classes d'entités nommées spécifiques au domaine.

Résolution des coréférences, normalisation des noms d'objets célestes, et détection de relations
Les résultats du tableau 4 montrent que le système de base F-coref a une précision très faible (0.09) et un rappel élevé (0.26), entraînant un faible score F1 (0.13). Le modèle manque de connaissances spécifiques au domaine pour résoudre avec précision les coréférences dans ce contexte. Cependant, en affinant le modèle (astroFastCoref), il a pu apprendre des motifs spécifiques à l'astrophysique, le rendant plus efficace dans la résolution des coréférences liées aux objets célestes en atteignant un score F1 CoNLL de 0.53.

Modèle	CoNLL		
	Précision	Rappel	F1
F-coref	0,09 (0)	0,26 (0)	0,13 (0)
astroFastCoref	0,67 (0,01)	0,44 (0,01)	0,53 (0,01)

TABLE 4 – Précision moyenne, rappel et F-mesure (avec écart-type) du système F-coref évalué sur l'ensemble de test de notre corpus avec et sans affinage. Chaque expérience a été exécutée cinq fois (sur 50 époques lors de l'affinage) avec différentes amorces aléatoires.

Catalogue	Précision
SIMBAD	60,39
SIMBAD + NED	71,28
SIMBAD + NED + TNS	80,19

TABLE 5 – Précision d'un système de normalisation des noms d'objets célestes à l'aide de bases de données astronomiques.

Précision	Rappel	F1
0,77	0,80	0,79

TABLE 6 – Performance d'un système biLSTM de détection de relation entre un objet céleste et une propriété physique.

Le Tableau 6 présente les performances de notre système de détection de relations. La F-mesure de 0,79 suggère une performance satisfaisante dans l'identification de relations entre les objets célestes et les propriétés physiques.

5.3 Analyse des gains obtenus sur $\text{TDAC}_{\text{test}}$

Dans cette section, nous analysons l'intérêt du corpus d'entraînement astroECR pour la détection d'entités nommées sur le corpus TDAC. La Figure 3 illustre l'évolution des performances sur $\text{TDAC}_{\text{test}}$, le jeu de test du corpus TDAC (Alkan *et al.*, 2022), en fonction de différents corpus d'entraînement utilisés. Pour cela, nous comparons l'ensemble d'entraînement de référence $\text{TDAC}_{\text{train}}$, avec les deux ensembles d'entraînement $\text{DEAL}_{\text{train}}$ et $\text{astroECR}_{\text{train}}$. Pour ces deux ensembles, nous faisons varier la taille du corpus d'entraînement par incrément de 25%. Les résultats obtenus montrent clairement l'intérêt d'astroECR, permettant d'améliorer la F-mesure globale moyenne d'environ 4 points. Dans le cadre de notre travail, l'entraînement d'un système sur le corpus DEAL ne permet pas l'amélioration de la F-mesure. Ceci peut s'expliquer par la nature différente des documents du corpus DEAL (articles scientifiques), qui possèdent des propriétés différentes des rapports d'observation.

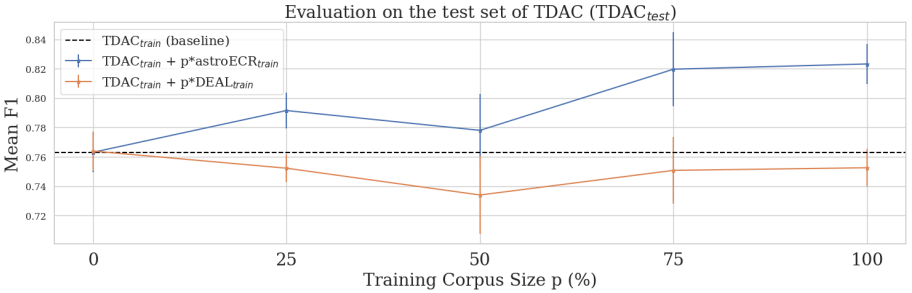


FIGURE 3 – Evaluation sur $\text{TDAC}_{\text{test}}$ d'un système de détection d'entités nommées à base d'un modèle astroBERT pour la détection d'entités nommées en fonction de la taille du corpus d'entraînement.

Classe	**$\text{TDAC}_{\text{train}}$**				**100% $\text{astroECR}_{\text{train}}$**				$\Delta F1$ (%)
	N	P	R	F1	N	P	R	F1	
`CelestialObject`	130	0,88	0,94	0,90	519	0,94	1,0	0,97	+ 7,7
`CelestialRegion`	20	0,31	0,23	0,26	149	0,64	1,0	0,78	+ 200
`Observatory`	60	0,54	0,58	0,64	101	0,80	0,67	0,72	+ 12,49
`Database`	36	0,75	0,81	0,78	79	0,77	0,90	0,83	+ 6,4

TABLE 7 – Comparaison des gains obtenus par classe sur le jeu de test $\text{TDAC}_{\text{test}}$ en fonction du corpus d'entraînement utilisé. Les métriques utilisées sont la Précision (P), Rappel (R) et la F-mesure (F1). N correspond au nombre de mentions d'entités de la classe dans le corpus d'entraînement.

Le Tableau 7 présente en détail les performances de certaines classes essentielles du domaine, notamment le repérage des noms d'objets célestes (`CelestialObject`), des coordonnées dans le ciel (`CelestialRegion`), des installations astronomiques impliquées dans l'observation (`Observatory`), ainsi que des bases de données astronomiques (`Database`). La plupart des classes bénéficient d'une amélioration des performances. Toutefois, la classe `CelestialRegion` est celle qui a le plus tiré profit de l'enrichissement du corpus. En effet, nous constatons une nette amélioration de la F-mesure (passant de 0,26 à 0,78). Cette progression significative s'explique par une augmentation marquée du rappel et une hausse plus modérée de la précision.

6 Conclusion et perspectives

Nous avons cherché à remédier au manque de données annotées, en élargissant le corpus TDAC (Alkan *et al.*, 2022) de 75 à 300 documents annotés. Notre corpus devient ainsi une ressource unique dans le domaine, proposant des annotations plus riches en entités nommées astrophysiques, dont cinq catégories nouvellement définies. Nous avons également annoté les coréférences et les relations entre objets célestes et leurs propriétés physiques, tout en normalisant les noms d'objets célestes à l'aide de bases de données astronomiques. À travers des expérimentations sur le corpus test d'astroECR, nous avons développé des modèles et fourni des scores de référence, soulignant l'utilité de notre ressource pour l'annotation automatisée de futurs documents. Nous avons démontré que l'augmentation de la taille du corpus améliore notablement la détection d'entités nommées sur le corpus de test de TDAC. Notre objectif est de mettre à disposition de la communauté TAL une ressource propice à des études complémentaires, telles que la résolution de coréférences scientifiques ou la détection de relations. Notre corpus peut également enrichir d'autres corpus spécialisés tels que ceux proposés par Chaimongkol *et al.* (2014) et Brack *et al.* (2021). À l'avenir, les modèles développés pourraient être utilisés à des fins d'extraction d'information comme suggéré par Sotnikov & Chaikova (2023). Nous mettons à disposition notre corpus, guide d'annotation, code, et modèles pour les deux communautés.

Références

ALKAN A. K., GROUIN C., SCHUSSLER F. & ZWEIGENBAUM P. (2022). TDAC, the first corpus in time-domain astrophysics : Analysis and first experiments on named entity recognition. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, p. 131–139, Online : Association for Computational Linguistics.

BANERJEE P. S., CHAKRABORTY B., TRIPATHI D., GUPTA H. & KUMAR S. S. (2019). A information retrieval based on question and answering and ner for unstructured information without using sql. *Wirel. Pers. Commun.*, **108**(3), 1909–1931. DOI : 10.1007/s11277-019-06501-z.

BARTHELMY S. D., BUTTERWORTH P. S., CLINE T. L., GEHRELS N., FISHMAN G. J., KOUVELIOTOU C. & MEEGAN C. A. (1995). BACODINE, the real-time BATSE gamma-ray burst coordinates distribution network. *Astrophysics and Space Science*, **231**, 235–238.

BASSIGNANA E. & PLANK B. (2022). What do you mean by relation extraction ? a survey on datasets and study on scientific relation classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, p. 67–83, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-srw.7.

BECKER M., HACHEY B., ALEX B. & GROVER C. (2005). Optimising selective sampling for bootstrapping named entity recognition. In *In Proceedings of the ICML Workshop on Learning with Multiple Views*, p. 5–11.

BELTAGY I., LO K. & COHAN A. (2019). Scibert : A pretrained language model for scientific text. In *EMNLP* : Association for Computational Linguistics.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BRACK A., MÜLLER D. U., HOPPE A. & EWERTH R. (2021). Coreference resolution in research papers from multiple domains. *CoRR*, **abs/2101.00884**.

CHAIMONGKOL P., AIZAWA A. & TATEISI Y. (2014). Corpus for coreference resolution on scientific papers. In *Proceedings of the Ninth International Conference on Language Resources*

and Evaluation (LREC'14), p. 3187–3190, Reykjavik, Iceland : European Language Resources Association (ELRA).

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

GAL-YAM A. (2021). The TNS alert system. *Bulletin of the AAS*, **53**(1). https ://baas.aas.org/pub/2021n1i423p05.

GALIBERT O., ROSSET S., GROUIN C., ZWEIGENBAUM P. & QUINTARD L. (2012). Extended named entities annotation on OCRed documents : from corpus constitution to evaluation campaign. In *International Conference on Language Resources and Evaluation*, Istanbul, Turkey. HAL : hal-01831254.

GREZES F., BLANCO-CUARESMA S., ACCOMAZZI A., KURTZ M. J., SHAPURIAN G., HENNEKEN E. A., GRANT C. S., THOMPSON D. M., CHYLA R., MCDONALD S., HOSTETLER T. W., TEMPLETON M. R., LOCKHART K. E., MARTINOVIC N., CHEN S., TANNER C. & PROTOPAPAS P. (2021). Building astrobert, a language model for astronomy & astrophysics. *CoRR*, **abs/2112.00590**.

GREZES F., BLANCO-CUARESMA S., ALLEN T. & GHOSAL T. (2022). Overview of the first shared task on detecting entities in the astrophysics literature (DEAL). In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, p. 1–7, Online : Association for Computational Linguistics.

GRISHMAN R. & SUNDHEIM B. (1996). Message Understanding Conference- 6 : A brief history. In *COLING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*.

HACHEY B., ALEX B. & BECKER M. (2005). Investigating the effects of selective sampling on the annotation task. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, p. 144–151, Ann Arbor, Michigan : Association for Computational Linguistics.

JOSHI M., LEVY O., ZETTLEMOYER L. & WELD D. (2019). BERT for coreference resolution : Baselines and analysis. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 5803–5808, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1588.

JURAFSKY D. & MARTIN J. H. (2023). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. USA, 3rd édition.

KIM Y. & WEBBER B. (2006). Implicit reference to citations : a study of astronomy. *ERPANET*.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LUAN Y., HE L., OSTENDORF M. & HAJISHIRZI H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3219–3232, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1360.

MAZZARELLA J. M., MADORE B. F. & HELOU G. (2001). Capabilities of the NASA/IPAC extragalactic database in the era of a global virtual observatory. In J.-L. STARCK & F. D. MURTAGH, Éds., *SPIE Proceedings* : SPIE. DOI : 10.1117/12.447177.

MOLLÁ ALIOD D., VAN ZAANEN M. & SMITH D. (2006). Named entity recognition for question answering. In L. CAVEDON & I. ZUKERMAN, Éds., *Proceedings of the Australasian Language Technology Workshop, ALTA 2006, Sydney, Australia, November 30-December 1, 2006*, p. 51–58 : Australasian Language Technology Association.

MURPHY T., MCINTOSH T. & CURRAN J. R. (2006). Named entity recognition for astronomy literature. In *Proceedings of the Australasian Language Technology Workshop 2006*, p. 59–66, Sydney, Australia.

NGUYEN T. D., TING Y.-S., CIUCĂ I., O'NEILL C., SUN Z.-C., JABŁOŃSKA M., KRUK S., PERKOWSKI E., MILLER J., LI J., PEEK J., IYER K., RÓŻAŃSKI T., KHETARPAL P., ZAMAN S., BRODRICK D., MÉNDEZ S. J. R., BUI T., GOODMAN A., ACCOMAZZI A., NAIMAN J., CRANNEY J., SCHAWINSKI K. & UNIVERSETBD (2023). Astrollama : Towards specialized foundation models in astronomy.

OTMAZGIN S., CATTAN A. & GOLDBERG Y. (2022). F-coref : Fast, accurate and easy to use coreference resolution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing : System Demonstrations*, p. 48–56, Taipei, Taiwan : Association for Computational Linguistics.

OTMAZGIN S., CATTAN A. & GOLDBERG Y. (2023). LingMess : Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2752–2760, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.202.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 2227–2237, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1202.

RUTLEDGE R. E. (1998). The Astronomer's Telegram : A Web-based Short-Notice Publication System for the Professional Astronomical Community. *Publications of the Astronomical Society of the Pacific*, **110**(748), 754–756. DOI : 10.1086/316184.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SEVGILI Ö., SHELMANOV A., ARKHIPOV M. Y., PANCHENKO A. & BIEMANN C. (2020). Neural entity linking : A survey of models based on deep learning. *CoRR*, **abs/2006.00575**.

SOTNIKOV V. & CHAIKOVA A. (2023). Language models for multimessenger astronomy. *Galaxies*, **11**(3). DOI : 10.3390/galaxies11030063.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). brat : a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 102–107, Avignon, France : Association for Computational Linguistics.

WENGER M., OCHSENBEIN F., EGRET D., DUBOIS P., BONNAREL F., BORDE S., GENOVA F., JASNIEWICZ G., LALOË S., LESTEVEN S. & MONIER R. (2000). The SIMBAD astronomical database. *Astronomy and Astrophysics Supplement Series*, **143**(1), 9–22. DOI : 10.1051/aas :2000332.

YADAV V. & BETHARD S. (2018). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 2145–2158, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

ZHENG J., CHAPMAN W. W., CROWLEY R. S. & SAVOVA G. K. (2011). Coreference resolution : A review of general methodologies and applications in the clinical domain. *Journal of Biomedical Informatics*, **44**(6), 1113–1122. DOI : https ://doi.org/10.1016/j.jbi.2011.08.006.