Guide d’annotation du corpus SCP

Merci à Motasem Alrahabi, Carmen Brando et Thierry Poibeau pour leurs bons conseils.

Dernière mise à jour le 25/05/2023 :

1. Explicitation de certains éléments
2. Ajout des remerciements

Mise à jour le 09/10/2022 :

1. Exportation et organisation des informations rassemblées sur le document de stage de l’ObTIC (accès restreint).

Par : Perrine MAUREL

I. Introduction

1. Corpus

Ce guide a été rédigé afin de formaliser l’annotation d’un corpus de textes littéraires en ligne nommé *La Fondation SCP*. Celui-ci se compose d’un large panel de textes appartenant à plusieurs genres différents, principalement ceux de l’imaginaire et de l’horreur, sous plusieurs formes différentes, généralement en prose. Concrètement, les textes se divisent en deux catégories : des textes intradiégétiques revêtant l’apparence de documents officiels appartenant à l’univers fictif qu’ils décrivent, et des textes extradiégétique avec focalisation, lesquels se rapprochent d’une narration plus classique.

Notamment, un format de textes intradiégétiques dit des “rapports” constitue une fiche informative au ton scientifique qui porte sur un sujet d’étude distinct : ce dernier reçoit dès lors une dénomination sous la forme “SCP-Numéro-Sigle”, le sigle indiquant la langue d’origine du texte. La nomenclature systématique qui résulte de ce titrage permet de repérer et d’annoter efficacement la présence d’entités narratives singulières dans les textes.

Le corpus est écrit par plusieurs mains selon un principe de libre utilisation et réinterprétation des éléments publiés : tant que les crédits de parentalité sont correctement attribués, il est possible d’utiliser les créations de quelqu’un d’autre dans ses propres écrits. Par conséquent, plusieurs entités narratives spécifiques ont un trajet atypique au sein de ce corpus pluriel : des personnages, mais aussi des lieux, des organisations voire des concepts sont réécrits par de multiples personnes et mobilisés dans divers contextes.

2. Objectifs et délimitation de ce guide

Objectifs :

1. Permettre une annotation exhaustive des entités nommées du corpus
2. Les différencier selon leurs types
3. Permettre de suivre le trajet d'une entité nommée à travers les textes (nécessite des outils extérieurs tels que ceux publiés sur le Github du projet)

Délimitation des recherches :

On veut retenir en priorité les entités nommées ayant une importance narrative dans les textes, ou fondamentale dans l'univers. On s'interroge donc sur la portée de l'entité sur deux échelles différentes : une échelle locale du texte, et une échelle globale du corpus général. Un personnage cité seulement de manière obscure dans un texte peut être annoté en raison de son importance dans d'autres textes. Le deuxième facteur demande de l'annotateurice une certaine connaissance du corpus qui n'est pas nécessairement acquise, aussi, il est plus facile sans doute de choisir d'annoter toutes les entités nommées.

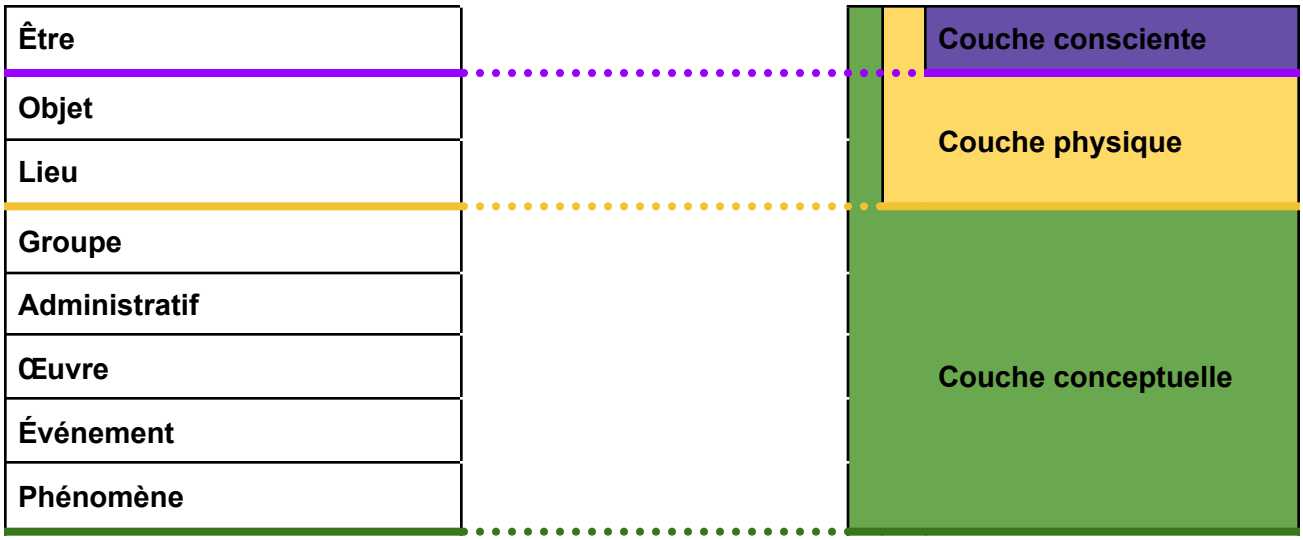
À ce stade, je fais le choix de ne récupérer que les entités nommées définies et si possible dénombrées.

On ne tient pas compte des changements d'états d'une entité nommée (Zaïre = RDC ; cadavre humain = humain) tant qu'ils ne modifient pas la nature ou la composition de l'entité (URSS =/= Russie).

II. Typage et annotation

1. Organisation des types

On a défini un total de huit classifications d'entités nommées différentes, selon une structure de dénuement progressif de chaque catégorie. Plus on descend dans la structure schématique des types, moins ledit type compte de caractéristiques variées. Cela se modélise par l'utilisation de trois "couches" qui se superposent sans pour autant se recouper entièrement.
Les types sont répartis comme suit :



2. Définitions des types

1. **Être** :

Définition :

Entité qui dispose de ou a disposé un temps d'un certain degré de volonté, lequel peut leur être prêté, programmé ou inné.

Concerne :

Les êtres humains, les créatures, les animaux, les robots qui répondent à des directives complexes, les êtres figés dans un état (cadavre, prisonniers d'un objet, robot éteint ou détruit) ou immatériels (esprit, être-concept).

Note :

Un groupe d'êtres de même nature est aussi une EN être.

À partir de là, les types ne sont plus concernés par la couche conscience (pas de volonté propre).

2. **Objet** :

Définition :

Entité dépourvue de volonté qui se caractérise par sa permanence physique et sa mobilité en tant qu'unité par des moyens humains, y compris technologiques.

Exemples :

Une statue est une entité objet ; en effet, même boulonnée au sol ou retenue par une force extérieure, elle peut être déplacée en tant qu'unité par des moyens humains.

Une œuvre titrée est un objet dans la mesure où le texte se réfère explicitement à sa forme physique et manipulable.
→ "J'ai toujours *la Bible* sur ma table de chevet." OK Objet
→ "Cette citation de la Bible fait réfléchir." NON Objet

3. **Lieu** :

Définition :

Entité dont l'existence physique unitaire, indépendamment des éventuelles infrastructures ou population qui pourraient exister en son sein, se définit par un statut inamovible par des moyens humains ET par :

1. des frontières délimitant un espace accessible, c'est-à-dire permettant un déplacement en son sein sans que l'unité du lieu en soit compromise par un déplacement de la matière.

2. OU (inclusif) son statut de lieu de transition d'un espace distinct à un autre.

Concerne :

Les pays, les bâtiments, les portails, les portes, les dimensions, l'univers, les planètes, les étoiles, les tunnels, les montagnes, les pays, les régions…

Exemple :

Si un objet traverse une étoile, l'état de l'objet importe peu : ce qui importe c'est l'état de l'étoile lors du déplacement. Une étoile est un lieu car un déplacement en son sein ne ruine pas son unité.

Un véhicule n'est pas un lieu : il est mobile par des moyens humains.
Un bâtiment préfabriqué et temporaire n'est pas un lieu (une tente, des toilettes de chantier…)

Note :

Les pays recevront de préférence comme clef leur appellation longue (trois lettres) dans le Code ISO 3166-1.

À partir de là, les types ne sont plus concernés par la couche physique (pas de forme physique).

4. **Groupe :**

Définition :

Entité dépourvue de volonté en tant que telle, mais qui qualifie néanmoins directement la réunion de plusieurs acteurs humains (ou êtres) mis en relation les uns avec les autres dans un contexte allant au-delà de la simple similarité de nature, donc un contexte situationnel, filial ou directionnel commun.

Concerne :

Les organisations, les familles…

Exemple :

Les survivants du naufrage, l'ONU, le gouvernement français, les mousquetaires…
"Les humains" n'est pas un groupe car leur seul lien est de nature. C'est une EN être.

5. **Administratif :**

Définition :

Entité dont l'existence dépend d'un corps administratif, régissant un principe sociétal ou juridique défini par des documents officiels, lesquels lui confèrent à la fois sa signification et sa portée.

Concerne :

Les devises, les lois, les dépôts légaux, les phrases d'identification officielles, les structures administratives de gouvernance (la Cinquième République), les identifiants de formatage ou d'archivage …

6. **Œuvre** :

Définition :

Entité désignant une substance intellectuelle et perceptible, ordonnée et intelligible, résultant d'un travail humain d'expression et de réflexion. Son existence ne dépend pas d'une officialisation juridique.

Concerne :

Les disciplines de recherche, les structures conceptuelles composant un texte (chapitre, paragraphe, phrase…), les documents que l'on désigne en tant qu'objet intellectuel plutôt qu'objet physique, les documents dématérialisés et numériques.

7. **Événement** :

Définition :

Entité qui se caractérise à la fois par sa ponctualité temporelle, par la nature humaine (ou être) des acteurs à son origine et par son rôle de marqueur chronologique. Le terme de ponctualité ne veut pas nécessairement exclure ici les événements récurrents, comme les solstices par exemple.

Concerne :

Les batailles et guerres, les naissances, les morts, les jours, les dates en tant que marqueurs temporels ponctuels arbitrairement créés par l'homme…

Exemples :

Un tremblement de terre naturel n'est pas un événement.
Un séisme artificiellement induit l'est.

8. **Phénomène** :

Définition :

Entité qui constitue la réunion de notions physiques ou conceptuelles dénuées de forme matérielle à proprement parler, bien qu'elles puissent avoir un effet sur la matière et les éléments matériels, notamment en induisant un changement d'état.

Concerne :

Les événements météorologiques, les concepts, les lois physiques et sociales, les comportements…

Exemples :

SCP-209-FR désigne la transformation des moutons en nuage.

III. Procédure d'annotation

1. Questions progressives

J'ai voulu créer une procédure mentale pour classer au mieux les entités nommées, soit un ensemble de questions progressives. Celles-ci reprennent très exactement l'enchaînement des couches et des types précédemment exposés.

Lorsqu'on identifie une entité nommée qui remplisse nos critères de sélection (importance au sein du texte ou de l'univers), on se pose les questions suivantes dans cet ordre précis. Le premier "oui" rencontré garantit la catégorisation, ou, s'il s'agit d'une question subsidiaire, doit induire une réflexion sur les limites de la catégorisation actuelle et comment les résoudre avant de reprendre le travail d'annotation :

1. **L'entité est-elle pourvue de volonté ?** (Être)

[Question subsidiaire A : ***L'entité est-elle consciente ?***

Si oui, c'est qu'il nous manque une catégorie de classification car toutes les classes qui suivent sont dépourvues de conscience.**]**

2. **L'entité est-elle déplaçable physiquement par des moyens humains ?** (Objet)

3. **L'entité est-elle constituée d'un espace accessible ou d'un espace de transition physique ?** (Lieu)

[Question subsidiaire B : ***L'entité dispose-t-elle d'une forme physique unitaire ?***

Si oui, c'est qu'il nous manque une catégorie de classification car toutes les classes qui suivent sont dépourvues de forme physique unitaire.**]**

4. **L'entité qualifie-t-elle directement la réunion de plusieurs acteurs humains (ou êtres) mis en relation les uns avec les autres ?** (Groupe)

5. **L'entité désigne-t-elle un principe administratif ou juridique formalisé par des documents officiels ?** (Administratif)

6. **L'entité qualifie-t-elle le produit intellectuel d'une démarche raisonnée de réflexion et d'organisation conceptuelle ?** (Œuvre)

7. **L'entité est-elle ponctuelle et d'origine humaine ?** (Événement)

8. **L'entité est-elle une réunion de concepts ou de principes ou de notions immatériels ?** (Phénomène)

Si on arrive à ce point sans jamais répondre oui, il faut alors élargir certaines catégorisations ou penser à une nouvelle catégorie selon l'entité nommée étudiée.

Un exemple de son application est le cas particulier d'une entité nommée qui serait un nuage.

2. Importance de l'ordre des types :

Cette méthode a un effet secondaire involontaire. En effet, puisque la première réponse positive entraîne la catégorisation, elle implique qu'on préférera prendre au sens large dans cet ordre :

Être > Objet > Lieu > Groupe > Administratif > Œuvre > Événement > Phénomène

D'où le fait d'avoir ordonné les types du plus précis au plus large.

Une telle approche implique également un certain recoupement entre certaines catégories : en s'en tenant simplement aux définitions, il est ainsi possible de catégoriser une même entité dans deux types différents. Par exemple, un contrat peut logiquement appartenir à "Administratif" et à "Œuvre". Il s'agit là d'une démarche assumée qui permet une spécification des critères de recherche : dans le corpus, la proportion des documents administratifs est telle que j'ai fait le choix de les différencier dans une catégorie spécifique.

De fait, la catégorisation est modulable en fonction des besoins de l'annoteurice et certaines des catégories se recoupant peuvent être fracturées ou regroupées selon.

3. Approche XML TEI P5 :

Si l'annotation veut suivre les normes XML TEI P5, je propose l'utilisation de balises rs (*referring string*) dotées de trois attributs.

Nom	**Usage**	**Statut**
key	id permettant de regrouper les différentes mentions d'une même entité	"Obligatoire" dans un contexte d'annotation avancé, sinon, non
type	indique la catégorisation de l'entité	Obligatoire
source	indique le texte d'origine de l'entité si cette dernière est fictive	Facultatif

L'attribut "source" peut recevoir comme valeurs notables :

1. un lien url vers le premier conte d'apparition, ou par défaut le centre traitant de l'entité
2. "unknown" si la source est inconnue
3. "generic" si l'entité est bien fictive mais qu'elle est trop générique pour être attribuée à un texte particulier. Par exemple : "la page 234".

IV. Problématiques rencontrées

1. Cas particuliers

Dépôt légal :

Qu'est-ce ? Techniquement, pas de forme matérielle (se distingue du document qui *atteste* du dépôt légal), réunion de concepts de propriétés.

On peut le classer en phénomène je pense ?

Trop d'occurrences, on va plutôt créer une EN administratif car il y a de toute façon un besoin.

Protocole :

intuitivement, je dirais Administratif. Mais attention, car un "Protocole" techniquement désigne un événement à mettre en place… Donc protocole = événement plutôt ? Selon la méthode de dénuement progressif, on tombe plutôt sous le coup de l'administratif (post-remaniement des couches).

Mesure Humes :

Unité de mesure de la réalité. En ce sens, j'aurais tendance à dire que c'est un phénomène.

Hier :

On peut exclure le problème en disant que ce n'est pas une entité nommée définie.

SCiPnet :

Réseau internet, qu'on ne peut pas classer dans l'EN administratif.

On veut une EN qui puisse l'accueillir mais qui ne soit pas trop précise dans sa définition afin de nous être utile quand même. Donc je crée une EN Oeuvre.

Devise "Sécuriser Confiner Protéger" :

Est-ce une EN administrative ? Oui, si elle est officielle et non personnelle. C'est le cas pour la Fondation SCP. Mais une devise personnelle d'un individu ne rentrerait pas dans la catégorie.

République

Très compliqué celui-ci parce que c'est à la fois un groupe et un principe administratif régissant ce groupe… On va faire différer selon son officialisation et son statut.

Par exemple, la République populaire de Chine désigne davantage le gouvernement du pays, donc EN groupe.

Mais si on parle de la Cinquième République, alors on parle d'une République spécifique du gouvernement Français, donc EN administratif.

Défense, Culture, Economie…

"Les rapports de la Défense/Culture/Economie" -> que sont defense, culture etc…?

Catégorisé en Oeuvre.

Nuage

Un nuage est-il un phénomène ?

→ Pas un être (pas de volonté)

→ Pas un objet (inamovible par les humains)

→ Pas un lieu (on ne peut pas se déplacer à l'intérieur sans compromettre l'unité du lieu)

→ Pas un événement (aucune origine humaine, pas de ponctualité temporelle)

→ Pas un groupe (ne qualifie pas des acteurs humains ou êtres)

→ Phénomène (réunion de notions physiques qui agissent sur la matière)

Vévé et symboles visuels et graphiques

L'existence d'un vévé mentionné dans un texte amène la question des symboles graphiques : bien qu'ils ne s'y réduisent pas, il est difficile de les séparer de leur forme physique.

On assimile ça au cas des œuvres textuelles : un exemplaire est déplaçable, donc objet, mais le fondement même du texte est une œuvre. De même pour les graphes.

Réseau 5g

Forme physique de la technologie mais pas des ondes… Mélange objet et phénomène ?
Comme il y a une forme physique, on assimile au "réseau aux normes 5G" donc des technologies -> objet

"La 5G" est une norme, donc plutôt administratif en l'état (conclusion atteinte à l'aide de Yoann Dupont).

L'aire de Broca

Partie du cerveau qui ne se qualifie pas vraiment en être ni en lieu. Alors on se rabat sur l'annotation de "Broca" simplement, de Paul Broca.

SCPs : désignation générale

Certains textes parlent “des SCPs” en général pour désigner les objets d’étude.
De même, dans le texte SCP-204-FR, la dénomination SCP-204-FR-02 désigne à la fois des objets et des lieux, ce qui entre en contradiction directe avec ma façon d‘étiqueter le corpus.

Ce cas est rare et demande une adaptation de l’étiquetage plutôt que celle des définitions d’entités nommées. Par défaut, on va donc basculer sur une désignation “administrative” pour expliquer que la nature même de l’objet désigné est variable et que le dénominateur commun demeure la désignation administrative utilisée formellement par une organisation.

2. Regroupement des catégories

Administratif et Oeuvre

Concrètement, la définition d’Oeuvre englobe Administratif. On a choisi une focalisation sur ce dernier concept en raison de la nature du corpus (des rapports officiels).

Lieu, Administratif et Groupe

Dans le cas des pays, notion de population et de territoire qui se mélangent. On prend par défaut la notion de territoire, donc de Lieu, mais il est très possible que certaines configurations amènent à une catégorisation différente.

Administratif et groupe se recoupent également car certains groupes ont aussi une existence administrative.

Phénomène et événement

La nature humaine des acteurs à l’origine est le seul critère déterminant entre les deux, et il a été arbitrairement décidé, ce qui implique que mes recherches et mes définitions pourraient être inutiles en dehors de mon corpus, ce qui est dommage.