

Etude quantitative des Creepypasta : définition et caractérisation

Alexandre Lionnet-Rollin

`alexandre.lionnet@chartes.psl.eu`

`https://github.com/Rollybre/memoire-creepypasta`

23 mai 2024

Introduction

Le début des années 2000 marque l'entrée d'Internet dans la seconde phase de son existence : le Web 2.0 correspond à une période de développement sans pareils pour l'époque de nouveaux canaux de communication. L'avènement des grands réseaux sociaux tout comme les grandes plateformes d'hébergement et de partage de contenus vont permettre une nouvelle façon de dialoguer. Et parmi les nouvelles pratiques permises par ceux-ci, on trouve de nouvelles formes d'écriture, qui renouvellent aussi bien la façon d'écrire que le but des productions : écriture collaborative, écriture fragmentaire, mais aussi une écriture qui peut, à la manière d'une trainée de poudre, se répandre bien au-delà de leurs sphères de productions. Parmi celle-ci, on trouve une écriture simple, et relativement courte, prenant la forme d'un court témoignage, qui rapidement vire au cauchemar : les creepypastas.

La première occurrence du terme remonte au mois de juillet 2007 sur le forum *4Chan*¹. Cette dénomination est le fruit de la rencontre entre l'adjectif *creepy* (litt. effrayant) et l'expression *copypasta*, elle-même contraction des verbes *copy* et *paste* (litt. copier, coller), désignant un bloc de texte copié et collé sur différents forums afin de le partager. Si les *copypastas* peuvent traiter de tous les sujets, aussi bien de blagues que de nouvelles et autres informations, les creepypastas sont quant à elles cantonnées au domaine du frisson : les creepypastas sont des copypastas dont le but est d'effrayer le lecteur. Ainsi nous nous appuyerons sur cette définition de Joe Ondrak pour amorcer notre réflexion :

1. Le post a depuis longtemps été supprimé. On peut néanmoins en trouver une archive au lien suivant : <https://web.archive.org/web/20111207000647/http://chanarchive.org/4chan/b/257/creepypasta>

[coppypastas are] short pieces of prose, sometimes accompanied by an image or a video [...] meant to be copied and pasted [...] and spread on the Internet via social media, e-mails, and message boards [...] whereas coppypasta can be about almost any subject, creepypasta is usually aimed at scaring the reader and/or viewer.²(Ondrak, 2018 p.162)³

Cette première définition permet de souligner la dualité intrinsèque des CP : d'une part elles sont définies par leur aptitude à effrayer le lecteur (le caractère horrifique), et d'autre part par leur capacité à être transmise sur différents canaux.

I. Les caractéristiques formelles des CP

D'un point de vue formel, il est possible d'identifier deux caractéristiques prédominantes des CP : ce sont des productions écrites courtes et qui appuient leur narration sur la première personne. Le récit des CP prend souvent la forme d'un témoignage, d'une histoire rapportée, justifiant le rapprochement avec le folklore et les légendes urbaines⁴. L'usage de la première personne est d'autant plus important car, à l'instar du folklore ou de la légende urbaine, l'histoire, et sa narration, joue avec la frontière située entre la fiction et la réalité.

2. *Les [coppypastas sont] de courts morceaux de prose, parfois accompagnés d'une image ou d'une vidéo [...] destinés à être copiés et collés [...] et diffusés sur Internet via les médias sociaux, les courriels et les tableaux d'affichage [...] Alors que le coppypasta peut porter sur presque tous les sujets, le creepypasta vise généralement à effrayer le lecteur et/ou le spectateur.*

3. Joe Ondrak, « Spectres des Monstres : Post-postmodernisms, hauntology and creepypasta narratives as digital fiction », *Horror Studies*, 9-2 (oct. 2018), p. 161-178, DOI : 10.1386/host.9.2.161_1.

4. Trevor J. Blank et Lynne S. McNeill (éd.), *Slender Man is coming : creepypasta and contemporary legends on the Internet*, Logan, 2018.

La plupart des CP suivent un même schéma : un narrateur souvent intra-diégétique rapporte une histoire qu'il aurait vécu. Cette histoire commence souvent de façon anodine voire trivial (vie de tous les jours, quotidien banal...), puis prend un tournant troublant voire terrifiant. Ce schéma n'est pas sans rappeler un autre genre littéraire : la littérature fantastique. La distance entre un personnage *a priori* banal et un événement hors-du-commun est à la base aussi bien de la littérature fantastique que de l'esthétique des CP. T. Todorov, dans son *Introduction à la littérature fantastique* propose comme première définition du genre fantastique cette dichotomie :

Le fantastique c'est l'hésitation éprouvée par un être qui ne connaît que les lois naturelles, face à un événement en apparence surnaturelle (Todorov, 1970, p. 27)

Cette "hésitation" se trouve aussi bien du côté du personnage que du lecteur :

Le fantastique implique donc [...] l'existence d'un événement étrange, qui provoque une hésitation chez le lecteur et le héros [...] (Todorov, 1970)

Hésitation qu'on retrouve sous la forme de l'effet recherché par la CP sur le lecteur : la peur.

En plus de la façon de raconter, les thèmes en eux même sont proches de ceux employés par ces genres « littéraires » : il est question de monstres sanguinaires, d'objets retrouvés, et souvent de l'enfance, le narrateur se rappelant d'événements de sa propre jeunesse, ou faisant appel à une certaine nostalgie. Et les monstres traditionnels, s'ils sont souvent présents, laissent place dans de nombreux cas à des éléments tout aussi effrayants, mais beaucoup

plus pernicious : des éléments de notre quotidien, et particulièrement l'outil numérique.

Notons par exemple la CP *Candle Cove* qui prend la forme d'une discussion sur un forum entre plusieurs utilisateurs se remémorant une émission dérangeante de leur jeunesse. Les souvenirs des uns et des autres s'ajoutent, aussi bien que les détails perturbants, jusqu'à ce qu'un utilisateur souligne le fait que sa mère s'inquiétait de le voir regarder la statique de la télévision lorsque celui-ci prétendait y voir ladite émission.

La perversion du numérique que cela soit du jeux-vidéo en passant par les vidéos et le plus souvent Internet, prolonge la remarque sur une frontière réalité/fiction friable : les objets les plus terrifiants sont souvent des objets qu'on côtoie au jour le jour sans se rendre compte de leur « potentiel horrifique ».

Ainsi la mobilisation du genre de l'*Analog Horror* est assez fréquente.

Ce genre de l'horreur s'appuie sur une déformation et/ou sur la perversion de l'outil numérique dans toute sa diversité, dans le but de créer un climat dérangeant né du contraste entre l'aspect réconfortant du numérique et sa déformation⁵.

On peut trouver une bonne illustration de ce phénomène dans les productions du type *Found Footage*. Ces productions prennent la forme de vidéos, souvent déclarées comme des vidéos perdues puis retrouvées par un tiers. La vidéo réplique souvent le format VHS. Le format VHS est une norme d'enregistrement vidéo sur bande magnétique, en l'occurrence la cassette. Si la majorité des consommateurs (nous reviendrons plus tard sur ceux-ci) aujourd'hui n'ont pas connu directement l'âge d'or des VHS, celle-ci est facilement

5. Jessica Balanzategui, « Creepypasta, 'Candle Cove', and the digital gothic », *Journal of Visual Culture*, 18-2 (août 2019), p. 187-208, DOI : 10.1177/1470412919841018.

associé aux productions remontant à l'époque de leur parents, et donc potentiellement aux premières expériences de visualisation de contenu vidéo durant l'enfance, faisant ainsi du format VHS un vecteur de nostalgie.

Si les exemples sont nombreux dans le domaine du cinéma (on peut ici penser aux films *The Blair Witch Project*, ou bien encore *Paranormal Activity* plus récemment), on en retrouve plusieurs occurrences parmi les CP les plus iconiques : les séries de vidéos inspirées par des CP comme la série *Marble Hornets* ou *The Backrooms* en sont les parfaits représentantes.

Or en plus d'un thème, le numérique fait partie intégrante de l'identité des CP. Rappelons ainsi notre définition de départ : les CP sont avant tout des productions littéraires sur Internet destinées à la diffusion sur Internet.

Ces objets nativement numériques ne le sont pas seulement par leur existence sur Internet mais aussi par la forme qu'ils prennent : Il n'est pas rare que les CP soient des productions multimodales, faisant intervenir d'autres médium, comme la vidéo, la photographie ou la musique.

Notons par exemple la CP *Ben Drowned*⁶ : classique du genre, cette CP raconte l'histoire d'un jeune homme jouant au jeu-vidéo *The Legend of Zelda : Majora's Mask* acheté dans une mystérieuse brocante.



FIGURE 1 – Une statue invoqué par le joueur : elle ne bouge pas et le suit durant sa partie, générant un malaise nouveau.

BEN Drowned fait usage de tous les tropes mentionnés jusqu'ici : La narration

6. https://creepypasta.fandom.com/wiki/BEN_Drowned

est assurée par un narrateur intradiégétique, qui se trouve être un jeune homme, à la première personne ; une suite d'évènements paranormaux ont lieu sans que le narrateur ou le lecteur soient en mesure d'expliquer ; le jeu dont il est question est un jeu célèbre, souvent parmi les premières expériences vidéoludiques de beaucoup, et donc en un sens nostalgique. S'ajoute à cela la forme de cette CP : le récit du narrateur est entrecoupé d'extrait vidéo issu du jeu, montrant les différents bugs ou distorsion absente du jeu original.

Cette CP est emblématique car elle met en lumière la dualité fondamentale des creepypastas : puisant à la fois dans le folklore et la tradition littéraire fantastique, ainsi que dans une culture numérique, tant par ses thèmes que par sa forme assumant cette modernité numérique, les creepypastas opèrent comme un carrefour entre deux univers distincts. En s'érigeant comme récit recombinaut⁷ mêlant le neuf au vieux, cette convergence peut être analysée à la lumière du concept de culture de la convergence, développé par Henry Jenkins⁸.

Henry Jenkins a principalement exploré la convergence entre une culture industrielle traditionnelle, dominée par de grandes sociétés, et une culture numérique émergente, caractérisée par la collaboration et la participation active des individus. Toutefois, dans le contexte des creepypastas, cette hiérarchie traditionnelle n'est pas aussi pertinente. En effet, les creepypastas sont des œuvres nativement numériques, voire intrinsèquement connectées, où le rôle des médias traditionnels n'est pas aussi central. Ainsi, la dynamique de convergence se manifeste davantage à travers une analogie entre les médias et canaux traditionnels d'une part, et les grandes creepypastas qui

7. Naomi Jacobs, *The character of truth : historical figures in contemporary fiction*, Carbondale, 1990 (Crosscurrents/modern critiques).

8. Henry Jenkins, *Convergence culture : where old and new media collide*, OCLC : ocm64594290, New York, 2006.

agissent comme un canon de la culture numérique, d'autre part.

Il est essentiel de souligner la distinction du cadre théorique entre la culture de la convergence traditionnellement étudiée par Jenkins et la convergence observée dans le contexte des creepypastas. Alors que Jenkins se concentre sur la fusion entre les industries médiatiques traditionnelles et les nouvelles formes de médias numériques, l'analyse des creepypastas révèle une convergence plus subtile entre les traditions narratives anciennes et les formes d'expression contemporaines, façonnées par l'ère numérique. Ainsi, la convergence dans le domaine des creepypastas offre un exemple unique et éclairant de l'évolution des cultures narratives à l'ère numérique.

II. Les CP : du pareil au même ?

Cette notion de parcours, de trajectoire nous permet de revenir au second élément de définition d'une CP. Nous avons mentionné jusqu'ici l'importance de la forme, du caractère terrifiant, percutant et facilement répliquable. Ce dernier élément rentre dans les faits à son tour dans une définition des CP comme une production qui est faite pour être partagé et copié. Cette définition d'un élément textuel ou plus globalement culturel qui existe dans le but d'être copié, est celle d'un même comme défini par Richard Dawkins dans les années 1970.⁹

Les CP comme même

R. Dawkins définit le même en étendant la conception du gène à un élément culturel, et non plus simplement biologique : le même à l'instar du gène, est

9. Richard Dawkins, *Le gène égoïste*, trad. par Laura Ovion, Paris, 2003 (Poches Odile Jacob, 111).

un réplicateur, qui ne se transmet non pas par les gamètes, mais par imitation d'un autre membre du groupe d'appartenance. Pour ce qui est d'une CP, il est difficile de ne pas la concevoir comme un mème tant sa raison d'être est d'être imité, comme l'origine du mot le laisse sous-entendre. Ainsi, il convient de se demander à l'instar d'un gène, comment une CP survie dans le 'pool de CP'(sic) pour reprendre l'expression de Dawkins ? Les 3 caractéristique que souligne l'auteur sont : la longévité, la fécondité, et la fidélité de copie. Parmi ces caractéristiques, R. Dawkins exclut rapidement la première, en considérant que la question de la longévité importe peu : tant qu'une copie existe qu'importe la durée de vie d'une copie ¹⁰. Néanmoins dans notre cas, la question de la longévité est cruciale : Internet n'agit pas comme une archive sans fin. Au contraire, il est assez fréquent que des productions nativement numériques disparaissent ¹¹.

Dans le cas d'une CP, on peut distinguer plusieurs formes de longévité : la longévité sur la plateforme d'origine, la longévité d'une copie sur une autre plateforme ou bien encore la longévité « absolue » c'est-à-dire son espérance de vie de la première publication, en passant par toutes ses copies. Or celle-ci est difficilement quantifiable et représente en enjeu clé : remonter à l'origine d'une CP n'est pas évident, du fait de sa nature.

Les deux autres caractéristiques sont toutes aussi importantes : est-il possible de mesurer cette fécondité ? Si la fécondité d'un mème scientifique est mesurée par son nombre de citation, est-il possible d'identifier de quoi quantifier cette fécondité pour les CP ? De même pour la fidélité, il serait intéressant de voir

10. « La longévité de n'importe quelle copie d'un mème est probablement relativement peu importante, comme c'est le cas pour n'importe quelle copie d'un gène. » *Ibid.*, voir p.218

11. C'est le cas par exemple de l'ARG *This house has people in it*, jeu de piste en ligne dont la grande majorité des ressources sont aujourd'hui indisponibles

comment une ou plusieurs CP ont pu subir transformations et changements au cours de leur existence. Enfin, toujours dans le cadre théorique du même, il serait intéressant de voir si la fécondité est liée à certaines caractéristiques textuelles. Autrement dit, il serait intéressant de voir s'il existe certains tropes ou façon de faire d'un point de vue littéraire qui agissent sur la survie ou réussite d'une CP.

Les différentes trajectoire d'une CP

Ces 3 caractéristiques constituent une première base, mais ne représente pas une fin en soi. En effet, ces caractéristiques assez générales ne permettent pas de rendre compte de la diversité des trajectoires potentielles d'un même et par extension d'une CP .

Par définition, la CP est transmise par simple copier-coller : une fois produite sur un forum donné (par exemple *4Chan*), celle-ci se retrouve copiée puis collée sur le même forum, ou dans le cas échéant sur d'autres forums. Cette forme de transmission (forme historique du genre) était particulièrement pertinente à une époque où les forums ne pouvaient archiver indéfiniment les publications (c'est le cas de *4Chan*, berceau de bon nombre de CP devenue aujourd'hui incontournable, où les posts, après 24h d'inactivité, se voient supprimé) : le copier-coller était donc un moyen de survie pour les CP.

Les super CP

Afin d'illustrer ce principe, prenons l'exemple d'une des CP les plus emblématiques : SCP-173¹²

Parmi les grandes CP qui ont traversé le paysage internet, SCP-173 passe

12. <http://fondationscp.wikidot.com/scp-173>

difficilement inaperçue. Apparue sur le sous-forum /x/ de *4Chan*, cette CP est composée d'un court texte et d'une image. Si l'image est dérangeante, le plus important se trouve dans le corps du texte : l'entité photographiée est décrite, de telle sorte à nous laisser penser qu'un groupe, ou un organisme paragouvernemental s'en occupe, mais surtout de telle sorte à ce qu'on comprenne que cette créature n'est pas la seule retenue dans l'ombre.



FIGURE 2 – *Untitled 2004* créée par Izumi Kato. Cette photographie a été prise par Keisuke Yamamoto.

Durant plusieurs jours, SCP-173 s'est retrouvé copié puis collé (complétant ainsi, avec le caractère terrifiant son appellation de CP) sur le sous-forum /x/ puis sur la page d'accueil du site (le *board* /b/), jusqu'à ce que d'autres utilisateurs se joignent au mouvement et produisent à leur tour d'autres CP inspirées par la forme et l'univers de cette CP originelle. C'est la suite de

son cheminement qui fait passer cette histoire de simple CP à véritable canon : au fil des ans la communauté qui s'est construite autour de SCP-173 et de ce qu'on nomme désormais la *Fondation SCP* s'est autonomisée, créant un wiki dédié (une première fois grâce au CMS *EditThis* puis au CMS *Wiki-dot*) puis continuant de croître. Aujourd'hui la *Fondation SCP* est présente dans 12 langues différentes et est forte de plusieurs milliers de productions originales dans chacune de ces langues.

Cette progression et ce rapport à une création de base peut nous amener à penser les CP comme une forme de fanfiction¹³.

13. Laura Goudet, « Agentivité de l'horreur, creepypastas et jeu vidéo » (,

Ce rapprochement peut apparaître comme pertinent car, à l’instar d’un texte canonique source, les CP va être copié, annoté puis inévitablement modifié et ces modifications vont produire des textes à part entière. On note aussi la proximité en termes de moyen de production : l’écriture est collaborative en ce qu’elle produit une œuvre fragmentaire et asynchrone, chacun produisant à son rythme des éléments supplémentaires.

De plus on trouve une dynamique similaire à celle que les fanfictions entretiennent avec le canon : l’idée de fanon¹⁴, où la production qui découle de l’œuvre originale (et donc canonique) va tirer son autorité de cette massivité¹⁵. Néanmoins cette analogie ne fonctionne que dans de rares cas. Les CP sont majoritairement des textes uniques, d’effrayantes bouteilles jetées à la mer par un utilisateur, qui n’a pas vocation à s’établir comme récit fondateur.

SCP-173 est donc un des rares exemple de cette trajectoire particulière : celle d’une CP prenant tellement d’ampleur qu’elle devient elle même la base d’un canon ou d’un *fandom*.

Il n’existe que trois exemples de cette trajectoire : SCP-173 et la *Fondation SCP*, les *Backrooms* et *Slenderman*. Ces trois cas suivent une trajectoire similaire : une première publication sur un forum (respectivement *4Chan* et

2021), Artwork Size : 13 pages, pages 57-69 ISBN : 9782406125488 Medium : application/xhtml+xml,application/pdf,online,database,application/xhtml+xml; application/pdf Publisher : [object Object], 13 pages, pages 57-69, DOI : 10.48611/ISBN.978-2-406-12548-8.P.0057, voir p. 2.

14. Marion Lata, « Du canon au fanon : Sacralités multiples du canon littéraire dans la fanfiction », dans *Sacré canon : Autorité et marginalité en littérature*, dir. Anne-Catherine Baudoin, Code : Sacré canon : Autorité et marginalité en littérature, Paris, 2022 (Actes de la recherche à l’ENS), p. 109-122, URL : <http://books.openedition.org/editionsulm/4739> (visité le 11/10/2022).

15. Roy T. Cook, « Canonicity and Normativity in Massive, Serialized, Collaborative Fiction », *The Journal of Aesthetics and Art Criticism*, 71-3 (août 2013), p. 271-276, DOI : 10.1111/jaac.12021.

SomethingAwful), une vague importante de réaction sur ce même forum, puis une "exportation" vers une plateforme dédiée. Si la trajectoire globale est la même, chacune de ces CP a des caractéristiques formelles et "trajectorielle" propre (changement de forme et/ou de média).

Les CP historiques

Alors que toutes les CP ne suivent pas la même trajectoire qui restent largement exceptionnelles, certaines productions, dès les débuts de ce genre, ont laissé une empreinte indélébile sur le lectorat. Elles se sont hissées au rang de références, d'histoires marquantes qui ont contribué à définir le genre.

Ces CP, que nous qualifierons de CP **historique**, ont suivi un début de trajectoire similaire, mais ne se sont pas hissées au rang de canon. Malgré cela elles ont connus un franc succès. Ce succès ne prend pas la forme d'histoires liées mais par des reproductions sur d'autres formes. Au delà d'un simple copier-coller, ces histoires ont été narré sur d'autres plateformes, ou bien illustré sous différentes formes (animations, dessin...). *Ben Drowned*, que nous avons mentionné plus tôt, est un exemple de cette trajectoire : en plus de sa navigation sur différents forums, de nombreux utilisateurs, sur Youtube par exemple, utilise les images du jeu présentes dans la CP originale pour illustrer tout en narrant l'histoire associée¹⁶, ou ont cherché à expliquer, approfondir l'expérience¹⁷. On peut citer, à titre d'exemple, les CP *le syndrome de Lavanville*¹⁸ qui évoque une rumeur sur la musique d'une version d'un jeu Pokémon, ou encore *Squidward's Suicide*¹⁹, histoire d'un épisode disparu du dessin-animé *Bob l'Éponge*.

16. Voir par exemple <https://www.youtube.com/watch?v=2o7lcKHjdoQ&pp=ygULYmVuIGRyb3duZWQ%3D>

17. <https://www.youtube.com/watch?v=QJlqY104B00>

18. https://creepypasta.fandom.com/wiki/Lavender_Town_Syndrome

19. https://creepypasta.fandom.com/wiki/Squidward%27s_Suicide

Le reste des productions

Aujourd'hui néanmoins les différentes plateformes ne sont plus assujetties à une telle limite : ce faisant le mode de production et de diffusion des CP a évolué. Désormais, l'écrasante majorité des CP est produite sur des forums et sites dédiés (sans que cela soit néanmoins nécessaire, comme le montre l'exemple récent des *Backrooms*, apparu la première fois sur *4Chan*). Les deux pôles principaux de productions de CP sont aujourd'hui le *subreddit* *r/nosleep*²⁰ et le fandom *Creepypasta*²¹.

Ces deux plateformes voient quotidiennement de nouvelles histoires apparaître : ainsi si la diffusion et la viralité était un moyen de survie, les histoires produites sont désormais assujetties à des règles et des méthodes bien différentes. Ce faisant une nouvelle façon d'exister, une nouvelle trajectoire, s'est développée avec la sédimentation du "genre" au cours de ces dernières années.

Il convient néanmoins de noter dès maintenant une différence cruciale dans le fonctionnement de ces deux plateformes : en plus d'accueillir du contenu produit régulièrement le fandom *Creepypasta* joue aussi un rôle d'"archive" des CP historiques.

Cette sédimentation, liée à la production massive de CP²², entraîne aussi une attente différente vis à vis de la qualité des productions : d'un genre spontané, la CP s'est construite au fur et à mesure comme un genre travaillé.²³

20. <https://www.reddit.com/r/nosleep/>

21. <https://creepypasta.fandom.com/>

22. Le Fandom compte près de 14 000 pages

23. Anastasio García Roca, « Los creepypasta como textualidades metaficcionales : oportunidades formativas para la alfabetización mediática e informacional. » *Tonos digital : revista de estudios filológicos*-40 (2021), Publisher : Servicio de Publicaciones Section : Tonos digital : revista de estudios filológicos, p. 12, URL : <https://dialnet.unirioja.es/servlet/articulo?codigo=7857341> (visité le 18/09/2023).

Les plateformes accueillant ces productions ont donc développé des standards de création devant être respectés sous peine de ne pas pouvoir publier. Avant de rentrer dans la description des règles de ces plateformes, l'existence même de celles-ci est intéressante : il existe désormais une attente liée à la littérarité des CP. Cette littérarité, comme nous l'avons laissé sous-entendre, n'est pas évidente : des productions anonymes sur des canaux de diffusion nouveaux et souffrant encore d'un manque de légitimité sont autant d'indices quant à un décalage entre les productions littéraires patrimoniales et les productions numériques.

La question de la littérarité ("c'est-à-dire ce qui fait d'une œuvre donnée une œuvre littéraire"²⁴) est, d'un point de vue quantitatif (et donc du point de vue de la forme et non du fond), définissable sous différentes formes : le but est de mettre en avant une complexité textuelle caractéristique des œuvres littéraires, des caractéristiques qui différencieraient les textes de simple publication sur un forum. De précédentes recherches ont montré à la fois la capacité à modéliser la littérarité²⁵ et quels éléments du texte pouvaient servir à l'évaluation de la littérarité²⁶. Dans le cadre de notre étude, nous commencerons par voir si des indices plus simples et plus facilement obtenables suffisent à caractériser le caractère littéraire des productions, sans exclure la possibilité future d'appliquer les dites méthodes.

La première hypothèse qu'il est possible d'émettre tiens à la sédimentation

24. T. Aron et Groupe de recherches en linguistique et sémiotique (France), *Littérature et littérarité : Un essai de mise au point*, 1984 (Annales littéraires de l'université de Franche-Comté), URL : https://books.google.fr/books?id=CL24GEWr_SEC.

25. Corina Koolen, Karina van Dalen-Oskam, Andreas van Cranenburgh et Erica Nagelhout, « Literary quality in the eye of the Dutch reader : The National Reader Survey », *Poetics*, 79 (avr. 2020), p. 101439, DOI : 10.1016/j.poetic.2020.101439.

26. A. van Cranenburgh et C. Koolen, « Identifying Literary Texts with Bigrams », dans *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, dir. Anna Feldman, Anna Kazantseva, Stan Szpakowicz et Corina Koolen, Denver, Colorado, USA, 2015, p. 58-67, DOI : 10.3115/v1/W15-0707.

du genre : avec l'apparition d'attentes liées à la qualité, on peut conjecturer l'apparition ou la croissance de la littérarité des textes avec le temps.

Ainsi une première étape dans la perspective d'une analyse génétique se dessine : peut-on voir se dessiner des tendances dans la littérarité des productions littéraires ?

Pour ce faire, il est possible de mobiliser et de comparer des métriques comme la longueur des textes, la taille des phrases, la complexité du vocabulaire ou des indices de lisibilité (*readability*). En plus de l'étude de l'évolution dans le temps, il peut être intéressant de déterminer si ces différents indices de littérarité évoluent ou sont présents/absents de la même manière sur les différentes plateformes : cette analyse permettra de dresser la présence ou l'absence de similarité dans la forme au sein des différentes plateformes, et donc de déterminer une forme d'autorité propre à chaque plateforme.²⁷ Enfin, une dernière comparaison apparaît comme féconde : à défaut d'avoir une indication quantitative du succès des publications sur le fandom, une donnée qualitative (le caractère "historique") va nous permettre d'isoler un sous-corpus. L'existence de ce corpus nous permettra de tester les hypothèses d'évolutions plus en profondeur : est-ce que ces CP ont émergé au sein d'une tendance, ou est-ce qu'au contraire, ce sont elles qui ont permis l'émergence d'une tendance ?

27. Ariane Mayer et Nicolas Sauret, « L'autorité dans Anarchy. Les constructions de l'autorité et de l'auctorialité dans un dispositif de production littéraire collaborative : le cas de l'expérience transmédia Anarchy.fr », *Quaderni*-93 (mai 2017), Publisher : Éditions de la Maison des Sciences de l'Homme, p. 63-73, DOI : 10.4000/quaderni.1078.

Récupération des données

La récupération des données a nécessité une méthode différente pour chacune des plateformes. En effet, le moissonnage des données présentes sur le web est dépendant à la fois de la forme de la plateforme et des données qui y sont présentes tout comme de la potentielle politique d'utilisation et de récupération de celle-ci.

Pour ce qui est du fandom CP, la récupération des données a été relativement facile : le site n'est pas structuré sous la forme d'un arbre, comme on pourrait s'y attendre. Au contraire, une grande partie du site, dont les histoires à récupérer font parties, n'est pas hiérarchisée. Ce faisant, il n'est pas possible d'utiliser l'organisation du site pour repérer et récupérer les différentes publications. Cette structure ne permet pas une méthode "brutale" qui consisterait à récupérer l'ensemble du site puis filtrer les pages qui nous intéressent en fonction du lien.

A défaut de hiérarchisation, le site utilise une structure par hyperlien, où les pages sont organisées autour d'autres pages centrales, qui permettent par exemple de référencer certaines catégories. Ainsi pour accéder à une partie du site il faut : la chercher explicitement ,ou dans la plupart des cas, y accéder en appuyant sur un lien présent sur un page. Ainsi, afin de récupérer les publications, il est nécessaire de trouver une page qui renverrait vers toutes les autres publications. Par chance, une telle page existe sur le site : cette page a néanmoins la particularité de ne pas être accessible depuis la page d'accueil. A partir de cette page la méthode de récupération est la suivante : dans un premier temps on récupère l'ensemble des liens des pages, puis on récupère les données au sein des pages qui nous intéresse. Pour ce faire, les

bibliothèques Python *Selenium* et *BeautifulSoup* ont été mobilisé : ces deux bibliothèques permettent respectivement de naviguer au sein des pages en simulant le comportement d'un utilisateur, et de récupérer les informations dans le code des pages. Afin de naviguer à travers la page de référence, il suffit de "cliquer" sur le bouton page suivante (action réalisée à l'aide de *Selenium*). Une fois toutes les pages collectées, il faut récupérer la partie qui nous intéresse : le corps du texte. Nous avons utilisé la fonctionnalité d'édition du site à notre avantage : cette fonctionnalité permet d'afficher le texte brut, sans mise en page. De plus, cette représentation place le texte dans une zone mieux délimitée dans le code source, tout en donnant accès à des informations moins accessibles : les catégories par exemple, mais aussi et surtout l'auteur s'il est mentionné. [insérer illustrations]

Une fois la récupération de donnée effectuée, les textes bruts et les métadonnées (date, catégories, auteur) sont structuré puis stocké dans un fichier à part, au format `.csv` .

Pour ce qui est du subreddit, la marche à suivre pour récupérer les textes et les données associées est sensiblement différente. D'une part, *reddit* dans son entiereté est un site à l'affichage dynamique : autrement dit, les données apparaissent au fur et à mesure que l'utilisateur progresse sur la page, et ne sont pas à une place défini. A cela s'ajoute une limite d'affichage arbitraire définie par le site (en l'occurence, il n'est pas possible d'afficher plus de 1000 publications à la fois). Outre la difficulté de récupération des données, ces limites ne permettent pas de saisir la taille réelle du subreddit²⁸ : la seule donnée disponible est le nombre de membre du subreddit. Si cette

28. Les évolutions récentes de la politique de gestion des données de la plateforme a rendu presque impossible la collecte automatique de données et de statistiques. Ainsi des sites spécialisés comme <https://subredditstats.com> ne sont plus à jour

donnée est importante, elle n'est qu'une indication indirecte sur le nombre de publications (chaque membre ne publie pas nécessairement). Ce moissonnage nous a permis de récupérer l'ensemble des productions du site, soit près de 13000 publications (13128 pour être exact).

Pour le premier moissonnage, nous avons décidé de récolter les publications au nombre de vote le plus élevé, indépendamment de la date. En effet, contrairement à la plateforme *Fandom*, reddit dispose d'un indicateur quantitatif quant à la réception d'une publication : cette information nous permettra par la suite d'explorer les inférences entre réussite et forme. De cette manière, nous avons récupéré un millier de pages.

S'il est possible de considérer que les presque 14000 productions récupérées sont suffisantes, le déséquilibre important entre les corpus risque de biaiser les résultats. De plus, le fait d'avoir sélectionné les productions les mieux notées de la plateforme reddit est intéressant, mais risque là encore de mener à un biais important quant à la caractérisation des corpus : on peut supposer que les caractéristiques textuelles des meilleures productions ne reflètent pas nécessairement les caractéristiques de toutes la plateforme.

Une recherche plus approfondie nous a permis de découvrir la base de données Pushshift, base de données recensant les productions de l'ensemble des subreddit de 2005 à 2019²⁹. Ainsi de 1000 publications, nous avons été capable de récupérer pas loin de 190000 publications³⁰, nous permettant par la même de déterminer le nombre de production totale sur cette période.

La base de données se présente en premier lieu sous la forme d'une base

29. Je tiens à remercier particulièrement M. Armin Pournaki pour m'avoir aiguillé dans mes recherches

30. le total de publication s'élève dans les faits à 385668 productions. Néanmoins plus la moitié a été supprimé ou retiré de la plateforme.

SQL, que nous avons converti en un format tabulaire (csv à nouveau) pour une utilisation plus facile.

Premiers résultats

Exploration et visualisation

TABLE 1 – Comparaison des moyennes et écarts types de la taille moyenne des phrases dans les différents corpus

	Corpus_CP	Reddit	CP_fandom	Littérature Anglaise	Web
Moyenne	14.55	14.75	14.28	20.63	19.27
Écart type	3.53	3.58	3.46	5.56	8.66
Min	4.50	4.50	4.53	11.01	2.60
Max	25.18	25.18	25.09	39.02	215.00

Topic Modelling

Richesse Lexicale et indice de lisibilité

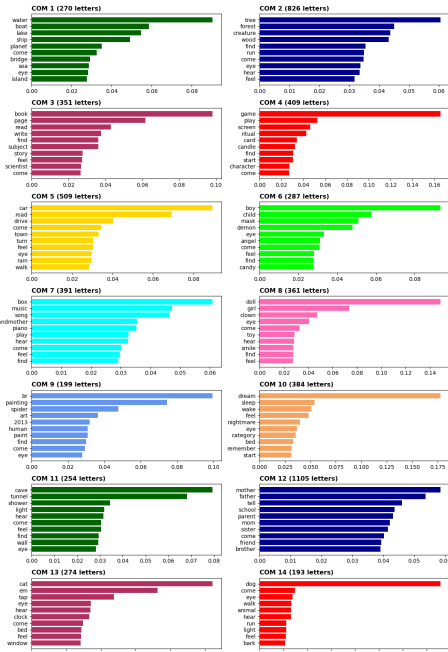


FIGURE 3 – Topic les plus fréquents du *Fandom Creepypasta*

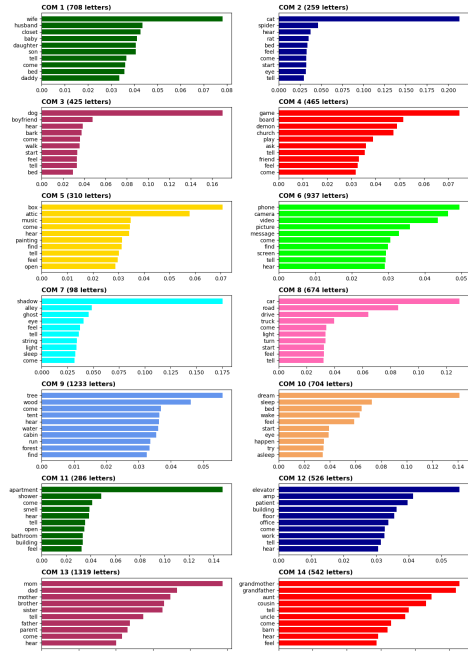


FIGURE 4 – Topic les plus fréquents du subreddit */nosleep*

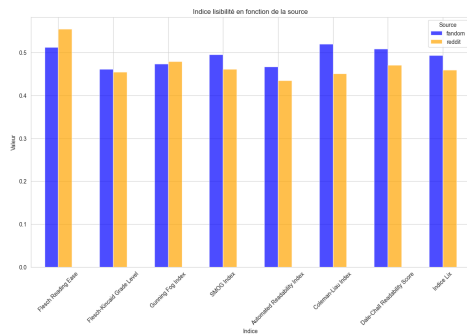


FIGURE 5 – Comparaison des indices de lisibilité en fonction de la plateforme

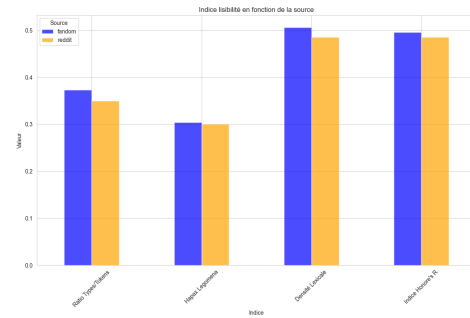


FIGURE 6 – Comparaison des indices de richesse lexicale en fonction de la plateforme

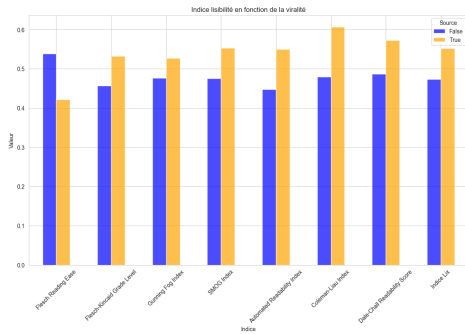


FIGURE 7 – Comparaison des indices de lisibilité en fonction de la viralité

O

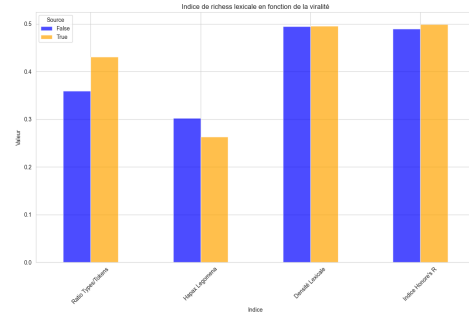


FIGURE 8 – Comparaison des indices de richesse lexicale en fonction de la viralité

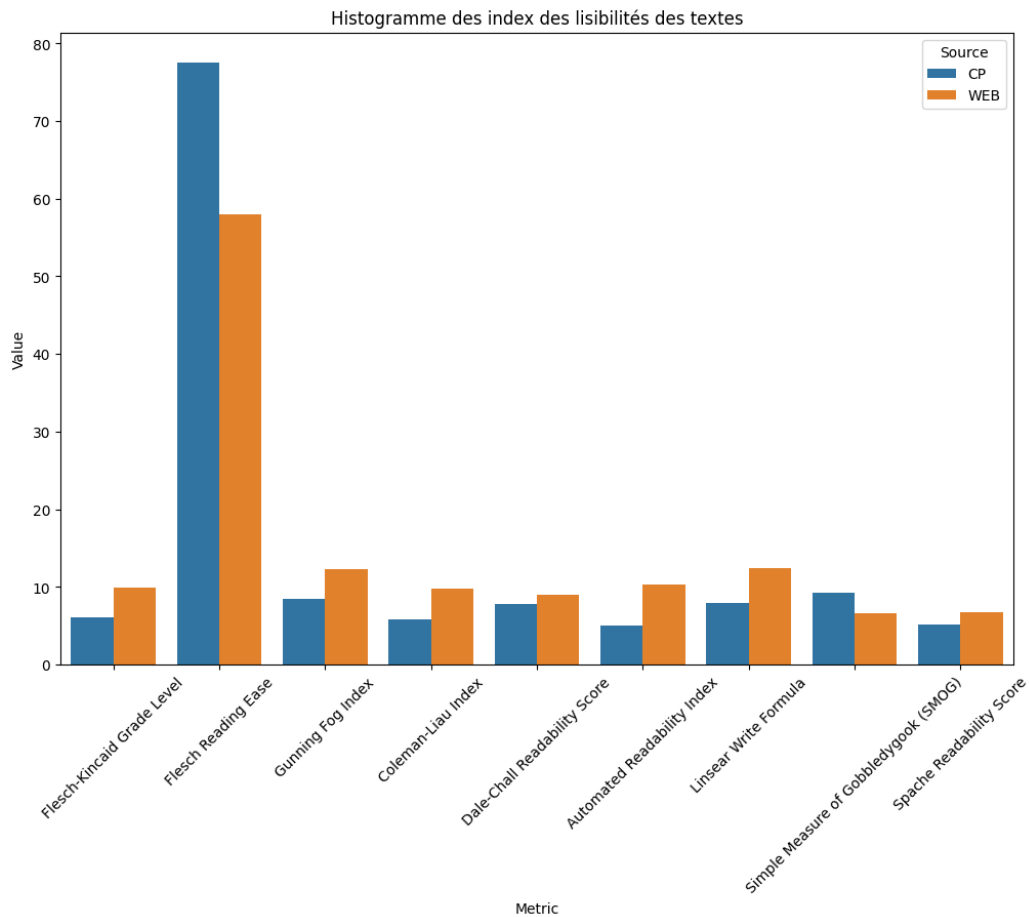


FIGURE 9 – Comparaison des indices de lisibilité entre les CP et un corpus Web

1 Annexe

1.0.1 Les indices de lisibilité

- Flesch Reading Ease : Évalue la facilité de lecture en se basant sur le nombre de mots par phrase et de syllabes par mot. Un score élevé (de 0 à 100) indique une lecture plus facile, tandis qu'un score bas indique un texte plus difficile.
- Flesch-Kincaid Grade Level : Exprime le niveau de lecture en termes de grade scolaire aux États-Unis. Plus le score est bas, plus le texte est facile à lire.
- Gunning Fog Index : Mesure la difficulté d'un texte en fonction du nombre de mots par phrase et du pourcentage de mots polysyllabiques. Il fournit un niveau de lecture approximatif.
- SMOG Index : Évalue la complexité d'un texte en comptant le nombre de mots polysyllabiques dans un échantillon de texte. Le score indique le niveau de lecture estimé.
- Automated Readability Index : Calculé en fonction du nombre moyen de lettres par mot et du nombre moyen de mots par phrase. Plus le score est bas, plus le texte est facile à lire.
- Coleman-Liau Index : Détermine la facilité de lecture en évaluant le nombre de lettres par mot et le nombre de phrases par 100 mots. Il fournit un niveau de lecture estimé.
- Dale-Chall Readability Score : Évalue la lisibilité en fonction du nombre de mots difficiles à comprendre dans un texte. Il est adapté pour les textes destinés à des lecteurs peu expérimentés.
- Indice Lix : Mesure la complexité d'un texte en évaluant le nombre de mots longs par phrase et le pourcentage de mots courts. Plus le score

est élevé, plus le texte est difficile à lire.

1.0.2 Les indices de richesse lexicale

- Ratio Types/Tokens : Évalue la richesse lexicale en comparant le nombre total de mots distincts au nombre total de mots. Un ratio élevé indique une plus grande variété de mots utilisés.
- Hapax Legomena : Nombre de mots n'apparaissant qu'une seule fois dans un texte, indiquant la variété du vocabulaire et la rareté des mots utilisés.
- Densité Lexicale : Mesure la proportion de mots uniques dans un texte par rapport au nombre total de mots. Une densité plus élevée indique une plus grande variété lexicale.
- Indice Honore's R : Calculé en divisant le nombre de mots uniques par la racine carrée du nombre total de mots, pour évaluer la richesse lexicale ajustée à la longueur du texte.

Références

- ARON (T.) et (FRANCE) (Groupe de recherches en linguistique et sémiotique), *Littérature et littérarité : Un essai de mise au point*, 1984 (Annales littéraires de l'université de Franche-Comté), URL : https://books.google.fr/books?id=CL24GEWr_SEC.
- BALANZATEGUI (Jessica), « Creepypasta, 'Candle Cove', and the digital gothic », *Journal of Visual Culture*, 18–2 (août 2019), p. 187-208, DOI : 10.1177/1470412919841018.
- Trevor J. Blank et Lynne S. McNeill (éd.), *Slender Man is coming : creepypasta and contemporary legends on the Internet*, Logan, 2018.
- COOK (Roy T.), « Canonicity and Normativity in Massive, Serialized, Collaborative Fiction », *The Journal of Aesthetics and Art Criticism*, 71–3 (août 2013), p. 271-276, DOI : 10.1111/jaac.12021.
- CRANENBURGH (Andreas van) et KOOLEN (Corina), « Identifying Literary Texts with Bigrams », dans *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, dir. Anna Feldman, et al., Denver, Colorado, USA, 2015, p. 58-67, DOI : 10.3115/v1/W15-0707.
- DAWKINS (Richard), *Le gène égoïste*, trad. par Laura Ovion, Paris, 2003 (Poches Odile Jacob, 111).
- GARCÍA ROCA (Anastasio), « Los creepypasta como textualidades metafictionales : oportunidades formativas para la alfabetización mediática e informacional. » *Tonos digital : revista de estudios filológicos*–40 (2021), Publisher : Servicio de Publicaciones Section : Tonos digital : revista de estudios filológicos, p. 12, URL : <https://dialnet.unirioja.es/servlet/articulo?codigo=7857341> (visité le 18/09/2023).

- GOUDET (Laura), « Agentivité de l'horreur, creepypastas et jeu vidéo » (, 2021), Artwork Size : 13 pages, pages 57-69 ISBN : 9782406125488 Medium : application/xhtml+xml,application/pdf,online,database,application/xhtml+xml ; application/pdf Publisher : [object Object], 13 pages, pages 57-69, DOI : 10.48611/ISBN.978-2-406-12548-8.P.0057.
- JACOBS (Naomi), *The character of truth : historical figures in contemporary fiction*, Carbondale, 1990 (Crosscurrents/modern critiques).
- JENKINS (Henry), *Convergence culture : where old and new media collide*, OCLC : ocm64594290, New York, 2006.
- KOOLEN (Corina), DALEN-OSKAM (Karina van), CRANENBURGH (Andreas van) et NAGELHOUT (Erica), « Literary quality in the eye of the Dutch reader : The National Reader Survey », *Poetics*, 79 (avr. 2020), p. 101439, DOI : 10.1016/j.poetic.2020.101439.
- LATA (Marion), « Du canon au fanon : Sacralités multiples du canon littéraire dans la fanfiction », dans *Sacré canon : Autorité et marginalité en littérature*, dir. Anne-Catherine Baudoin, Code : Sacré canon : Autorité et marginalité en littérature, Paris, 2022 (Actes de la recherche à l'ENS), p. 109-122, URL : <http://books.openedition.org/editionsulm/4739> (visité le 11/10/2022).
- MAYER (Ariane) et SAURET (Nicolas), « L'autorité dans Anarchy. Les constructions de l'autorité et de l'auctorialité dans un dispositif de production littéraire collaborative : le cas de l'expérience transmédia Anarchy.fr », *Quaderni*-93 (mai 2017), Publisher : Éditions de la Maison des Sciences de l'Homme, p. 63-73, DOI : 10.4000/quaderni.1078.
- ONDRAK (Joe), « Spectres des Monstres : Post-postmodernisms, hauntology and creepypasta narratives as digital fiction », *Horror Studies*, 9-2 (oct. 2018), p. 161-178, DOI : 10.1386/host.9.2.161_1.