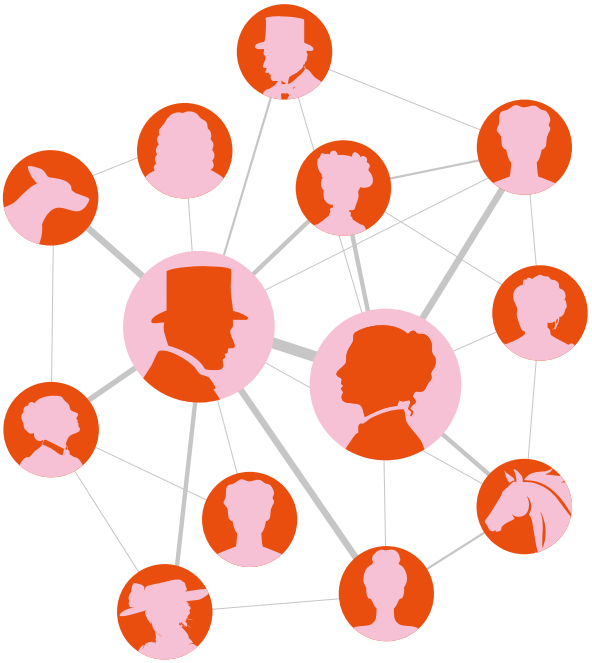
Les réseaux de personnages : nouveaux outils d'analyse

Narrations interactives, jeux vidéo et adaptations

Coline Métrailler



humanités numériques

Les réseaux de personnages sont couramment utilisés dans les études littéraires depuis le début du siècle pour cartographier les interactions entre les protagonistes d'un récit. Mais comment les adapter à des œuvres non linéaires, interactives ou cross-médiatiques ? Cet ouvrage relève le défi en proposant des outils innovants pour explorer la fiction à l'ère de ces nouvelles formes de narration.

Après avoir posé les bases narratives, mathématiques et informatiques de sa réflexion, l'autrice met en pratique les outils qu'elle a conçus à travers des études de cas concrètes. Le modèle du flux narratif permet une analyse novatrice de l'évolution des relations entre les personnages dans les jeux vidéo, et la comparaison de réseaux ouvre la voie à une lecture transversale des adaptations littéraires au cinéma.

Au cœur des humanités numériques, cet ouvrage allie clarté théorique et précision méthodologique. Il constitue un guide essentiel pour accompagner l'évolution de la recherche en sciences humaines.

Coline Métrailler est diplômée d'un master en mathématiques théoriques de l'École polytechnique fédérale de Lausanne (EPFL) et docteure ès lettres de l'Université de Lausanne (UNIL). La thèse à l'origine de cet ouvrage lui a permis d'allier ses deux passions : les jeux vidéo et la littérature. Elle est également membre du Gamelab UNIL-EPFL et participe à de nombreuses activités de médiation culturelle et scientifique.

épistémé

ISBN 978-2-88915-736-5

Les réseaux de personnages: nouveaux outils d'analyse

Les réseaux de personnages : nouveaux outils d'analyse

Narrations interactives, jeux vidéo et adaptations

Coline Métrailler

L'édition de cet ouvrage a reçu le soutien du Fonds national suisse
de la recherche scientifique.

Direction générale : Lucas Giossi
Directions éditoriale et commerciale : Sylvain Collette et May Yang
Direction de la communication : Manon Reber
Chargé de liaison éditoriale : Romain Bionda
Responsable de production : Christophe Borlat
Éditorial : Alice Micheau-Thiébaud et Jean Rime
Graphisme : Kim Nanette
Comptabilité : Daniela Castan
Graphisme de couverture : Kim Nanette

Première édition, 2026
Épistémé, Lausanne
Épistémé est une maison d'édition de la fondation des Presses
polytechniques et universitaires romandes
ISBN 978-2-88915-736-5, version imprimée
ISBN 978-2-8323-2320-5, version ebook (pdf), doi.org/10.55430/8069RDPCM

Imprimé en Tchéquie



Ce texte est sous licence Creative Commons : elle vous oblige,
si vous utilisez cet écrit, à en citer l'auteur, la source et l'éditeur original,
sans modifications du texte ou de l'extrait et sans utilisation commerciale.

Sommaire

Remerciements 7

Première partie
Introduction et bases théoriques 11

1 Introduction 13

2 Bases narratives 19

3 Bases mathématiques 23

4 Bases informatiques 37

Deuxième partie
Narrations plurielles et interactives 45

5 Approche théorique 47

6 Étude de cas : *Life is Strange* 61

Troisième partie
Œuvres plurielles et comparaisons 107

7 Approche théorique 109

8 Étude de cas : comparaison d'adaptations 131

Quatrième partie
Conclusion et ouverture 161

9 Continuer l'exploration 163

Cinquième partie
Annexes 173

A Flair ou spaCy ? 175

B Charnetto 181

Sources 199

Bibliographie 201

Table des matières 211

Remerciements

J'ai eu la chance d'être extrêmement bien entourée tout au long de mon parcours de thèse, il est donc bien naturel de consacrer quelques lignes à remercier toutes les personnes qui ont contribué à ce que cette aventure aboutisse.

Évidemment, je commence par adresser un immense merci à mon directeur de thèse, Aris Xanthos, qui a été présent pour moi à chaque étape. Merci d'avoir cru en mon projet, de m'avoir laissée beaucoup d'autonomie en répondant toujours présent lorsque j'avais besoin de tes conseils. Ces cinq ans ont été ponctués d'imprévus, à commencer par une fermeture des bureaux dans les premières semaines de collaboration… Merci d'avoir gardé le cap, d'avoir mis un point d'honneur à m'accompagner jusqu'au bout ; ta fiabilité, ta confiance, ta rigueur et ton soutien ont été déterminants pour moi.

Je pense également à François Bavaud et Fanny Barnabé qui ont formé un jury attentif, précis et généreux : merci pour votre disponibilité, votre réactivité, vos retours stimulants et pour les passionnants échanges autour de mon travail. Vous avez accompagné cette fin de thèse de la meilleure des manières.

Si je remonte à la source de ce projet, j'ai envie de remercier tout spécialement les deux premières personnes qui m'ont fait entrer en contact avec les humanités numériques, celles qui ont rendu tout le reste possible : Isaac Pante et Yannick Rochat. Merci d'avoir accepté ce café en 2019, d'avoir décelé la cohérence de mon parcours (alors que je me croyais tiraillée entre deux mondes irréconciliables), de m'avoir encouragée à sauter le pas et d'avoir été des soutiens constants depuis cette époque. Merci aussi à Johan

Berdat pour l'impulsion, les brainstorms, ton aide et ta présence dans ces premiers mois décisifs. Enfin, merci à François (encore !) qui m'a aiguillée dès le début, qui a ébauché ce projet avec moi en un temps record et qui a veillé sur moi à chaque étape.

J'adresse également un grand merci à l'ensemble de la section des Sciences du langage et de l'information pour tous les échanges passionnants, formels et informels. Merci tout particulièrement aux collègues doctorants et doctorantes, ainsi qu'aux premiers assistants et premières assistantes, qui ont amené beaucoup de joie dans les différents bureaux que j'ai occupés, qui ont dynamisé mes réflexions et mes pauses café, et sans qui ces cinq ans de contrat n'auraient pas eu la même saveur.

Dans les remerciements de groupe, je tiens à mettre en avant le Gamelab et tous ses membres : lorsque j'espérais y entrer, je ne me doutais pas de tout ce que j'y trouverais – des personnes formidables, déterminées et talentueuses, des opportunités en or, un réseau précieux et surtout, un environnement sain et stimulant basé sur l'entraide et l'esprit d'initiative. J'ai beaucoup de chance de faire partie de cette joyeuse bande. Longue vie au Gamelab !

Bien entendu, pour arriver au bout d'une tâche aussi conséquente qu'une thèse, il est précieux de pouvoir partager ses questionnements, ses doutes et ses succès avec d'autres personnes qui sont dans la même situation. En dehors des cercles déjà cités, j'aimerais remercier Pierre-Yves Houlmont, pour ses relectures de qualité et son soutien constant, ainsi que Melina Gravier et Ocyna Rudmann qui m'ont suivie au rythme de leur propre thèse, pour les repas, les cafés et les remises en question.

Merci également à mes proches, mes amis et amies d'horizons divers qui ont recueilli toutes mes anecdotes, partagé toutes mes péripéties et qui ont rendu mon quotidien plus beau. Merci à mes parents et ma sœur, mes piliers de toujours : vous avez largement contribué à ce que cette thèse voie le jour. Merci de m'avoir donné la confiance de choisir ma voie, et de m'avoir appris à travailler

pour l'atteindre. Merci de m'accompagner dans toutes mes aventures, de m'encourager et de me donner de la force au quotidien. J'ai une chance folle de vous avoir dans ma vie !

Olivier, tu as été à la fois mon coach, mon relecteur, mon bras droit, mon refuge, mon conseiller et mon confident, en plus de partager ma vie de la plus belle et douce des manières. Je n'aurais pu rêver meilleur compagnon pour traverser ces années ; je redoutais les dernières étapes de thèse, tu as réussi à les rendre si simples, agréables et enthousiasmantes. Merci pour ça, merci pour tout.

Pour l'édition de cet ouvrage, je rajoute un immense merci à l'équipe des Presses polytechniques et universitaires romandes, et en particulier à Alice Micheau-Thiébaud pour son travail remarquable et ses conseils précieux.

Et enfin, merci à celles et ceux qui liront ces lignes, et qui manifestent de l'intérêt ou de la curiosité pour ce travail.

Première partie

Introduction et bases théoriques

1 Introduction

Ceci n'est pas un ouvrage sur les personnages, les adaptations ou les jeux vidéo. Ces objets seront certes évoqués et thématisés, mais il est important d'aborder ce travail avant tout comme un pont entre les techniques computationnelles, c'est-à-dire les outils de traitement de données par ordinateur, le formalisme mathématique et la recherche en sciences humaines. En ce sens, les avancées significatives dans les domaines de la narratologie ou des *game studies*, proposées par le présent travail, ne pourront exister que par l'intermédiaire de chercheurs et chercheuses desdits domaines, s'ils ou elles choisissent de s'emparer des méthodes présentées ici. On trouvera dans cet ouvrage des méthodes d'analyse d'objets de sciences humaines, conceptualisées sur le plan mathématique, développées et implémentées grâce à l'informatique et accompagnées de cas pratiques pour illustrer leur potentiel et leur emploi. Cet ouvrage s'inscrit donc dans la tradition (plutôt récente) des *humanités numériques*.

Avant de rentrer dans le vif du sujet, il me semble important de définir la façon dont je perçois mon profil et mes motivations en tant que chercheuse. Le dialogue entre sciences computationnelles et sciences humaines est bénéfique pour les deux traditions, et chaque discipline peut nourrir l'autre de ses spécificités et de ses approches propres. Dans le cas particulier de l'emploi du numérique dans des recherches en lettres, l'informatique et

sa puissance de calcul permettent notamment d'apporter de la rapidité d'exécution et de l'automatisation pour des tâches répétitives, libérant ainsi l'attention des chercheurs et chercheuses en lettres qu'ils et elles peuvent alors consacrer au cœur de leur pratique (interprétation des données, étude des phénomènes, etc.). En outre, la capacité de traiter de gros volumes de données permet des agrégations, et ainsi l'observation de phénomènes à large échelle qui peuvent enrichir la palette d'outils à disposition de la recherche, toujours dans l'idée de proposer des outils ou des résultats dont le soin de l'interprétation est laissé aux experts et aux expertes.

Dans cette envie de mettre à la disposition de la recherche en littérature mon bagage en mathématiques et en sciences computationnelles, je m'attelle donc à :

— identifier et comprendre des besoins spécifiques à la recherche en sciences humaines,
— identifier des lieux d'automatisation ou de simplification de tâches répétitives,
— formaliser, au sens mathématique (logique, ontologique, etc.), les problèmes propres aux sciences humaines pour pouvoir imaginer des solutions informatiques,
— élaborer des solutions computationnelles permettant d'aborder ces problèmes sous un angle différent et complémentaire aux méthodes classiques,
— rendre ces solutions accessibles, ou du moins expliquer en détail leur logique pour que les résultats puissent être interprétés le plus précisément possible.

Cette démarche sous-tend les analyses développées dans cet ouvrage, qui prend pour objet de réflexion le réseau de personnages.

Le réseau de personnages est un objet mathématique qui relie les personnages d'une œuvre de fiction sur la base de leurs interactions. Dans l'esprit d'un « réseau social fictionnel », il s'agit

d'un puissant outil de cartographie narrative qui permet d'observer l'articulation d'un récit de manière macro et en s'intéressant aux liens entre les personnages qui le composent. Une telle structure n'a d'intérêt que si sa génération est adossée à des techniques computationnelles : on pourrait évidemment construire un réseau de personnages à la main, en recensant chaque occurrence de chaque personnage au fil de l'œuvre, mais la tâche est bien trop chronophage pour être généralisée. Avec la montée en puissance de l'informatique, de plus en plus de chercheurs et chercheuses ont élaboré des méthodes d'automatisation pour la création de ces réseaux, qui ont progressivement rejoint la boîte à outils des narratologues. Le réseau de personnages est aujourd'hui utilisé principalement pour analyser des œuvres individuelles, et ce, de manière « linéaire » (c'est-à-dire en partant d'un point de départ unique et en déroulant l'histoire selon l'ordre prévu jusqu'au point final). Il permet aussi bien d'étudier l'importance d'un personnage précis (est-ce que Hamlet est vraiment le personnage principal de *Hamlet* ?) que d'examiner le style d'un auteur ou d'une autrice (est-ce qu'on retrouve toujours la même structure à travers ses romans ?).

Dans cet ouvrage, je pars de ce socle théorique existant afin d'élargir le champ d'action du réseau de personnages et de le rendre encore plus modulable et utile, en le confrontant à deux notions de *pluralité* (des narrations ou des œuvres) :

1. Narrations plurielles
   Il existe différents types d'œuvres, comme les livres dont vous êtes le héros ou certains jeux vidéo, qui proposent une narration dite *interactive* : elles contiennent des choix menant à différents embranchements, offrant ainsi plusieurs expériences possibles d'une même histoire. Se pose alors la question de la construction d'un réseau de personnages qui tiendrait compte de ces choix, permettant des interactions qui peuvent avoir lieu ou non selon les options choisies.

C'est l'enjeu de la deuxième partie, qui propose un modèle théorique permettant d'adapter l'usage du réseau de personnages aux narrations interactives. Cette solution fait appel à une étape intermédiaire, soit la construction d'un « flux narratif », et permet aux méthodes traditionnelles de *distant reading* (terme présenté au chapitre 2) d'aborder des œuvres pensées de manière interactive, comme illustré à la fin de cette deuxième partie.

2. Œuvres plurielles
   La troisième partie s'intéresse à l'utilisation du réseau de personnages comme moyen de comparaison entre deux œuvres cross-médiatiques. L'idée est la suivante : si le réseau lui-même s'émancipe du média d'origine et propose une abstraction de l'œuvre basée sur les interactions de ses personnages, il représente peut-être une piste intéressante pour mettre en parallèle un livre et son adaptation sous forme de film, de pièce de théâtre ou de jeu vidéo, en écartant les contraintes inhérentes à chaque média et en établissant une base commune pour leur comparaison. Par exemple, nous verrons au chapitre 8 que la comparaison entre le réseau de personnages de la pièce *Romeo and Juliet* de Shakespeare et celui du film *Romeo + Juliet* de Luhrmann met en évidence des divergences concernant l'importance des différents personnages et leurs liens dans les deux versions de l'histoire.
   Or la comparaison de réseaux mathématiques est un problème vaste et encore en exploration. Dans le cadre de cet ouvrage, plusieurs mesures mathématiques permettant de calculer la similarité des réseaux seront sollicitées pour explorer différentes pistes. La première proposition est de faire une comparaison de mesures calculées directement d'après chaque réseau, et la deuxième est d'utiliser des mesures de similarité et de dissimilarité entre deux réseaux, qui sont

ensuite mises en perspective pour en relever les différences et les phénomènes distincts qu'elles capturent.

La mise en pratique de ces deux pistes nécessite de s'intéresser aux méthodes informatiques permettant de générer automatiquement des réseaux de personnages. Les différentes parties computationnelles de ce travail ont nécessité l'écriture de nombreuses lignes de code, qui ont été réunies en deux répertoires disponibles en ligne pour une consultation plus confortable :

— « charnetto »[1] contient le code du module permettant de construire automatiquement des réseaux de personnages et est présenté dans l'annexe B,
— « character_network »[2] contient les tableaux de données ainsi que le code des expérimentations liées aux différents exemples concrets de cet ouvrage.

Chacun de ces répertoires est accompagné d'une documentation inventoriant leur contenu avec plus de précision. Les passages faisant référence aux fichiers de l'un ou l'autre de ces répertoires seront accompagnés d'un renvoi vers leur URL.

Il semble désormais clair que cet ouvrage, profondément pluridisciplinaire, nécessite certaines bases en narratologie, en mathématiques et en informatique : les prochains chapitres de cette première partie sont justement prévus pour traverser ces prérequis et permettre une meilleure compréhension des axes de recherche annoncés : ils peuvent être pris comme une sorte de glossaire, à mobiliser lorsque certains termes nécessitent clarification si l'abondance de jargon mathématique ou technique semble trop aride ou rébarbative pour une lecture suivie.

[1] https://gitlab.com/maned_wolf/charnetto
[2] https://gitlab.com/nlp_tools/character_network

2 Bases narratives

Distant reading

On ne peut parler de *distant reading* sans mentionner Franco Moretti, qui a introduit ce terme dans son article « Conjectures on world literature » (2000) pour introduire un nouveau courant de méthodes servant à agréger et à analyser une grande quantité de données littéraires. L'idée est de mobiliser des outils statistiques, quantitatifs et computationnels pour explorer de larges jeux de données (romans complets ou métadonnées de librairies, par exemple) et voir ainsi émerger des résultats à large échelle. Le *distant reading* se positionne en complément du *close reading*, désignant l'étude détaillée d'un texte lu avec attention, porté par le mouvement du *new criticism* (en référence au livre *The New Criticism* de Ransom [1941]) qui a émergé en Angleterre et aux États-Unis dans les années 1920.

Ce courant a été accueilli avec méfiance par les sciences humaines, car son opposition avec le *close reading* laisse souvent penser qu'il vise à rendre l'approche traditionnelle obsolète et dépassée. Cette croyance a donné lieu à de nombreux articles, comme celui de Da (2019), critiquant l'arrivée du numérique dans la recherche littéraire et détaillant ses manquements. Or Birkholz et Budke (2021) présentent le débat sous un autre angle, arguant

que la plupart des recherches font en réalité un entre-deux, utilisant le numérique comme une aide à l'analyse et au traitement de données et puisant également dans les techniques de *close reading* pour construire un discours pertinent et enrichi de ces approches complémentaires. Le présent travail s'inscrit dans cette logique et dans la conviction que le dialogue entre numérique et sciences humaines nourrit les deux communautés.

Narratologie

La narratologie, science de la structure des récits, a émergé dans les années 1920 en Russie à travers l'étude de la construction des contes initiée en 1928 par Propp (1970 [1928] pour l'édition française). Cette recherche d'une grammaire du récit s'est élargie avec les années, pour prendre en compte tant les types de narrations que les supports médiatiques qui les accueillent. Ainsi, si les premières traditions se concentraient sur la littérature, les vingt dernières années ont vu naître de nombreux ouvrages reprenant les outils de narratologie pour d'autres types de médias, en particulier en lien avec le numérique et le caractère participatif, interactif et émergent de certaines œuvres modernes, comme l'expliquent Marti et Baroni en introduction de leur dossier « Les bifurcations du récit interactif : continuité ou rupture ? » (2014).

Définition du personnage

La définition du personnage s'est adaptée aux évolutions de la narratologie et des médias. D'après Jannidis (dans son chapitre « Character » du *Handbook of Narratology*, 2009), un personnage est un participant d'un monde narratif, par opposition aux individus du monde réel. Dans un contexte trans-médiatique et en particulier pour les jeux vidéo, Blom (2023) introduit la notion de personnage de jeu dynamique (*dynamic game character*) pour décrire ces personnages qui changent en fonction des choix des joueurs et joueuses. Au Japon et en partant des mangas, Ito

(2011) thématise l'omniprésence des personnages en distinguant les personnages « pleins » plongés dans des contextes narratifs (*kyarakutaa*) et les proto-personnages (*kyara*) utilisés comme mascottes ou personnalités non rattachées à un récit, mais pouvant évoluer en *kyarakutaa* dès lors qu'ils sont intégrés à une narration. Wilde (2019) cite Hello Kitty comme exemple de *kyara*, par sa nature première de produit commercial, et Bruce Wayne (Batman) comme *kyarakutaa* typique, avec un passé, une histoire et un contexte définis.

Ingrédients d'un récit

Si l'on s'intéresse aux ingrédients principaux d'un récit, la tradition d'Aristote a longtemps affirmé que l'intrigue était l'élément le plus important de la narration. Cette question fait désormais débat, et la recherche tend aujourd'hui à faire émerger le rôle des personnages comme essentiel à la création d'une « bonne histoire », comme le soutient McKee (1997). Jenkins mentionne également cette évolution dans son livre *Convergence Culture* (2006), rapportant notamment, en page 114, les propos d'un scénariste : « *When I first started, you would pitch a story because without a good story, you didn't really have a film. Later, once sequels started to take off, you pitched a character because a good character could support multiple stories. And now, you pitch a world because a world can support multiple characters and multiple stories across multiple media.* »[3]

Il existe de nombreux travaux qui présentent des manières de construire une histoire en partant de ses personnages, comme l'illustrent notamment Bamman *et al.* (2013) ou Cipresso et Riva

[3] « Quand j'ai commencé, on proposait une histoire parce que sans bonne histoire, on n'avait pas vraiment de film. Plus tard, quand les suites ont commencé à marcher, on proposait un personnage parce qu'un bon personnage pouvait soutenir plusieurs histoires. Aujourd'hui, on propose un monde parce qu'un monde peut accueillir plusieurs personnages et plusieurs histoires sur plusieurs médias. [Je traduis] »

(2016) dans leurs recherches. Au début des années 2000, Woloch introduit la notion de *character-space* (2003) et amène ainsi l'idée que les personnages évoluent dans un « environnement narratif » qui décrit leur position par rapport au reste du récit. Le *character-space* serait ainsi l'interaction entre un individu et un emplacement précis dans le récit, mettant en avant la place des personnages dans la structure générale de la narration.

Interactions entre personnages

Plus récemment encore, on a mis en lumière l'importance des interactions entre les personnages. Dans son livre *The Anatomy of Story* (2007), Truby parle de la « toile de personnages » (*character web*) pour expliquer la nécessité, lors de la création d'un nouveau personnage, de le relier aux autres et de considérer l'ensemble des personnages comme un réseau interconnecté. Prado *et al.* (2016) appliquent la théorie des réseaux à différentes œuvres de fiction pour observer des constantes dans la structure des interactions entre personnages à travers les types de récits. Enfin, Min et Park proposent un pont entre réalité et fiction à travers les personnages et leurs interactions (2016), pour citer quelques exemples de cette démarche.

3 Bases mathématiques

3.1 Graphes et réseaux

On attribue l'origine de la théorie des graphes au fameux problème des ponts de Königsbserg (voir figure 3.1), au XVIII^{e} siècle : les habitants de Königsberg s'interrogeaient sur l'existence d'un chemin passant par chacun des sept ponts de la ville et revenant au point de départ sans passer deux fois par le même pont. Euler utilise un graphe pour démontrer qu'il est impossible de construire un chemin satisfaisant ces conditions. Mais la théorie des graphes ne trouve pas seulement des applications en géographie, elle s'utilise également dans de nombreux autres domaines (biologie, chimie, sciences sociales, sciences humaines, etc.). Parmi les problèmes qui trouvent une solution grâce aux graphes, on peut notamment mentionner les sudokus ou les confections d'horaires qui peuvent tous deux être abordés grâce à la coloration de graphes ; les optimisations de transport qui découlent des problèmes de flots ; ou encore l'analyse de propagation des informations basée sur les réseaux sociaux.

La définition formelle du graphe est la suivante :

Définition 1. *Un graphe est un couple $G = (V, E)$, où V est l'ensemble des sommets, aussi appelés nœuds, et E est l'ensemble des arêtes. Une*

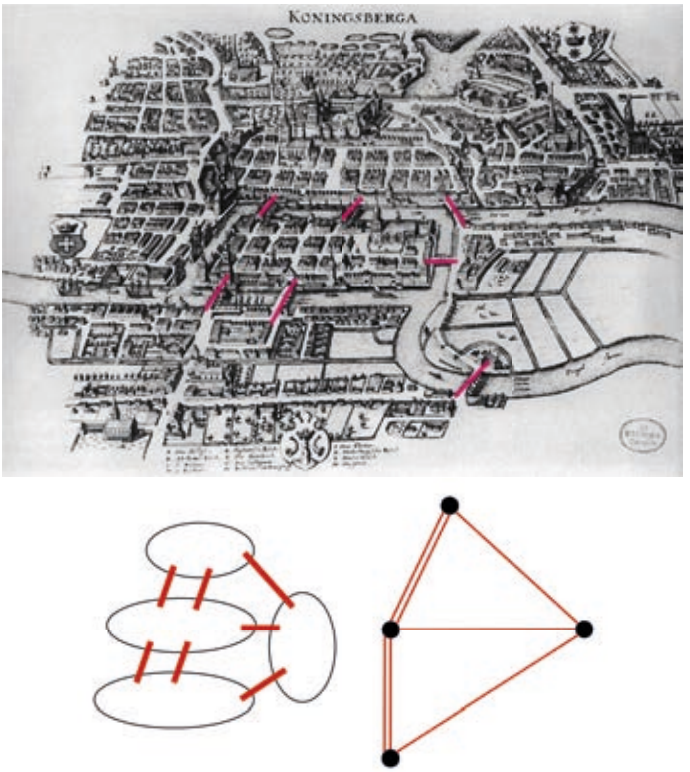


FIGURE 3.1 Les ponts de Königsberg : en haut, le plan de la ville, en bas une schématisation du problème sous forme de graphe.

arête a est déterminée par deux sommets u et v, définis comme ses extrémités, et on dit alors que a passe par u et v, qui sont considérés adjacents dans le graphe.

La différence de terminologie entre graphe et réseau n'est pas toujours claire dans la littérature, comme l'explique Barabási dans son livre *Network Science* (2016). Dans la suite de ce travail, on appelle *graphe* l'objet mathématique, et *réseau* un graphe appliqué à

un problème concret, pour lequel on attribue une interprétation et un label [4] précis aux différents nœuds et arêtes.

Définition 2. *Un chemin de longueur n dans un graphe $G = (V, E)$ est une suite de nœuds $v_0, ..., v_n$ reliés par des arêtes, de sorte que chaque arête reliant deux nœuds consécutifs dans la suite appartienne à E. Un cycle est un chemin dont l'arrivée est le même nœud que le départ, c'est-à-dire $v_0 = v_n$.*

Quelques propriétés des graphes

- Si tous les nœuds d'un graphe sont adjacents entre eux, alors le graphe est dit *complet*.
- On peut choisir d'associer un poids aux arêtes et/ou aux nœuds, valeur numérique utile pour représenter par exemple leur importance dans le réseau. Lorsque l'on attribue un poids aux arêtes, on dit que le graphe est *pondéré*.
- Il est également possible de spécifier un sens pour les arêtes, en fixant un point de départ et un point d'arrivée. Ces arêtes s'appellent alors des *arcs*, et un graphe composé de nœuds et d'arcs est un graphe *orienté*.
- Un graphe dans lequel il existe un chemin entre toutes les paires de nœuds est dit *connexe*.

On peut représenter un graphe sous la forme d'une matrice :

Définition 3. *On appelle matrice d'adjacence la matrice carrée $n \times n$ qui représente un graphe à n nœuds.*

- *Dans un graphe pondéré, la valeur a_{ij} de la matrice correspond au poids de l'arête du nœud i au nœud j, et $a_{ij} = 0$ si l'arête n'existe pas.*

[4] Dans l'ensemble de ce travail, on utilise l'anglicisme informatique *label* pour parler des étiquettes.

— *Pour un graphe non pondéré, la matrice d'adjacence est composée de 0 et de 1.*
— *Pour un graphe non orienté, la matrice d'adjacence est symétrique.*

Définition 4. *On appelle $G' = (V', E')$ un sous-graphe de $G = (V, E)$ si V' est contenu dans V et E' est contenu dans E, c'est-à-dire si G' correspond à G après lui avoir éventuellement enlevé des arêtes et des nœuds.*

Un sous-graphe G' complet de G est appelé une *clique*. Les cliques jouent un rôle important dans la théorie des réseaux sociaux, et par extension dans certaines approches d'études de réseaux de personnages, comme nous le verrons à la section 3.3.

3.2 Réseaux de personnages

Les réseaux de personnages sont des graphes appliqués à des œuvres de fiction, dans lesquels les nœuds représentent des personnages et les arêtes indiquent les interactions entre deux personnages. Il existe de nombreuses manières de définir et de représenter un réseau de personnages, en fonction de la nature de l'œuvre étudiée et des résultats que l'on désire mettre en évidence.

Nœuds : les personnages

Usuellement, chaque nœud représente un seul personnage, mais dans certaines études, on choisit de regrouper une famille ou un clan en un seul nœud par souci de clarté ou pour illustrer certaines dynamiques plus globales. Venturini *et al.* (2017) regroupent ainsi les Myrmidons en un seul nœud dans leur article sur l'*Iliade*, Falk (2016) se sert d'un nœud appelé « *the multitude* » dans son étude du *Bildungsroman*, et Liu et Albergante (2018) considèrent chaque nœud comme un représentant des différentes familles politiques dans la série *Game of Thrones*, afin d'étudier l'évolution de leurs affinités. Il est aussi possible d'ajouter d'autres types de nœuds au réseau, comme Lee et Yeung (2012) qui intègrent des nœuds de lieux pour augmenter les informations de

leurs réseaux de personnages sur les cinq premiers livres de la Bible.

Dans certaines études, on choisit de filtrer les nœuds et de ne garder qu'une liste réduite des personnages : cette décision est parfois motivée par l'angle de focalisation des recherches (par exemple si l'on veut se concentrer sur une population en particulier au sein d'une œuvre, comme c'est le cas pour Bossaert et Meidert (2013) dans leur étude des étudiants dans la saga de romans *Harry Potter*), mais le plus souvent, elle sert à éliminer les nœuds les moins fréquents pour augmenter la lisibilité des résultats, comme le font Suen *et al.* (2013) en retirant les personnages ayant moins de cinq répliques à travers les œuvres étudiées. En procédant de la sorte, avec un seuil en dessous duquel les nœuds sont exclus du réseau, on élimine une partie des erreurs provoquées par une extraction automatique des personnages (au risque toutefois, si le seuil est trop haut, d'éclipser certains personnages pertinents mais périphériques), et on concentre davantage notre attention sur les personnages les plus importants pour le récit.

On peut également associer un ou des attributs aux nœuds d'un réseau de personnages. Les attributs les plus courants sont le genre (comme chez Gleiser [2007]), l'appartenance à un clan ou à des catégories liées au récit étudié (comme Kydros *et al.* [2015] le font dans l'*Iliade* en distinguant les Grecs, les Troyens, les dieux et les autres personnages) ou des étiquettes sociologiques (comme Rochat et Triclot [2016] qui recensent notamment les personnages liés à la politique et à la science pour mesurer l'importance de ces deux sujets dans différentes œuvres de science-fiction).

Arêtes : les interactions

L'élaboration d'un réseau de personnages nécessite également de définir le concept d'*interactions*. Les choix les plus usuels sont les suivants :

— Cooccurrences : on considère que deux personnages sont en cooccurrence s'ils sont présents dans la même unité

narrative prédéfinie (groupe de paragraphes, scène de film, page, etc.). C'est le choix le plus courant et le plus facile à déterminer, mais il détache partiellement la notion d'interaction de la présence effective des deux personnages dans le même contexte de narration, puisque la taille des unités narratives ne garantit pas un découpage « organique » de l'intrigue. Il est ainsi tout à fait possible de considérer comme cooccurrents deux personnages présents sur la même page, alors même que l'on a changé de chapitre entre les occurrences de l'un et de l'autre.

— Conversations : dans un réseau conversationnel, deux personnages sont en interaction s'ils sont présents dans le même dialogue. On peut alors choisir, si l'analyse s'y prête, de construire un réseau de personnages orienté : l'arc part du personnage qui parle et pointe vers le personnage auquel il s'adresse. Ce choix d'interactions permet une interprétation plus claire du réseau, mais ignore néanmoins totalement les personnages qui ne s'exprimeraient pas verbalement au cours du récit. Un exemple de réseau conversationnel (non orienté) est à trouver dans le travail d'Elson *et al.* (2010) sur les romans britanniques du XIX^{e} siècle.
— Mentions : le réseau de mentions est en général un réseau orienté, illustrant cette fois les mentions de personnages par d'autres personnages au cours d'un dialogue. Couplé au réseau conversationnel par exemple, il peut donner un complément d'information intéressant sur les liens entre différents interlocuteurs et la fréquence à laquelle ils parlent les uns des autres, en présence ou non de la personne mentionnée. C'est ce que font notamment Deleris *et al.* (2018) à propos de la série *Friends*.
— Actions directes : dans de plus rares cas, les interactions peuvent également consister en des actions directes entre plusieurs personnages, comme se battre, s'entraider ou se serrer la main. Agarwal *et al.* (2012) recensent ce qu'ils ap-

pellent des *social events* pour définir les interactions entre les personnages dans *Alice au Pays des Merveilles*. Englober les actions directes a un sens certain pour l'interprétation des réseaux de personnages, mais la détection de ces interactions demande des efforts considérables, qu'on choisisse de les recenser à la main ou à l'aide d'algorithmes, raison pour laquelle on trouve peu d'exemples de ce type de réseaux dans la littérature.

— Affiliations : enfin, on peut choisir de représenter sous forme d'interactions les affiliations entre les personnages (liens de parenté, appartenance à un clan, etc.), comme dans les travaux de Srivastava *et al.* (2015) qui identifient automatiquement les relations entre les personnages à partir de résumés. Il s'agira souvent d'un deuxième type d'arêtes venant augmenter un réseau conversationnel ou de cooccurrences, pour fournir des informations plus détaillées sur la cartographie des liens entre les différents personnages de l'histoire.

Il est courant de fixer un seuil sur le nombre minimal d'occurrences d'une même arête pour la faire figurer dans le réseau, afin de limiter l'abondance d'informations de certaines œuvres. Certains travaux, comme celui de Iyyer *et al.* (2016), choisissent une valeur fixe (en l'occurrence, un minimum de cinq interactions pour considérer l'arête), et d'autres, notamment Park *et al.* (2013), définissent ce seuil comme un paramètre à faire varier en fonction de la densité de réseau désirée.

3.3 Mesures et interprétations

La théorie des graphes propose de nombreuses mesures qui trouvent une interprétation en sciences humaines lorsqu'on les applique à des réseaux de personnages. Outre les possibilités de visualisation qu'ils offrent, c'est dans ces mesures que réside

l'intérêt principal de la construction des réseaux de personnages, et les informations qu'elles mettent en évidence peuvent se révéler précieuses, car difficiles à extraire par des méthodes plus traditionnelles d'analyse de texte.

Les mesures présentées ici sont classées en deux catégories : d'abord celles qui s'appliquent à l'intégralité du réseau, puis celles qui se rapportent à des nœuds en particulier. Notons que par défaut, ces définitions ne tiennent pas compte de la pondération des nœuds et des arêtes, à l'exception de celles qui la mentionnent explicitement (comme la force des nœuds). Pour la plupart de ces mesures, il existe toutefois une ou des variantes pondérées, dont l'importance est thématisée à la section 7.1.

Mesures sur le réseau entier

Les mesures de distance moyenne et de diamètre reposent sur la définition de la distance entre deux nœuds d'un graphe :

Définition 5. *La distance entre deux nœuds d'un graphe est le nombre minimal d'arêtes à parcourir pour relier ces nœuds.*

La définition 5 permet d'introduire les deux premières mesures globales d'un réseau de personnages :

Diamètre et distance moyenne

Définition 6. *Le diamètre d'un graphe est défini comme la distance maximale à l'intérieur du graphe, c'est-à-dire la distance entre les deux nœuds les plus éloignés du graphe.*

Dans le contexte des réseaux de personnages, le diamètre est en général plutôt petit : c'est le cas pour Bonato *et al.* (2016) qui comparent les romans *Twilight* de Stephenie Meyer, *Harry Potter and the Goblet of Fire* de J. K. Rowling et *The Stand* de Stephen King, et trouvent des diamètres respectifs de 4, 2 et 3 (ce qui correspond à autant de poignées de main, au maximum, pour mettre

en contact deux personnages du réseau). Un diamètre plus grand peut indiquer par exemple une narration épisodique, dans laquelle le héros rencontre successivement des groupes de personnages indépendants les uns des autres.

Définition 7. *La distance moyenne est la moyenne des distances entre toutes les paires de nœuds du graphe.*

Ces deux mesures permettent d'exprimer la compacité du réseau d'interactions. Un grand diamètre peut indiquer que certains personnages sont isolés du reste de l'histoire, sans donner d'indice sur la tendance générale du réseau, alors qu'une grande distance moyenne témoigne d'un phénomène plus global de distribution des nœuds plutôt « étalée », et ainsi d'un réseau moins compact.

Densité

Définition 8. *La densité d'un graphe est la proportion d'arêtes existantes (sur toutes les arêtes possibles).*

Un réseau de densité 1 est donc un réseau dont tous les nœuds sont reliés entre eux (c'est-à-dire un réseau complet), à l'inverse d'un réseau de densité 0 qui ne contiendrait aucune arête. D'après Rochat et Triclot (2016), une grande densité est un des marqueurs des histoires en huis clos, témoignant du fait que tous les personnages sont amenés à interagir au fil du scénario. Une faible densité, quant à elle, peut indiquer des groupes sociaux très déconnectés les uns des autres.

Transitivité

Définition 9. *La transitivité d'un graphe est la proportion de triangles existants (sur tous les triangles possibles), c'est-à-dire la proportion existante de triplets de nœuds tous interconnectés.*

On peut interpréter la présence d'un triangle comme la schématisation mathématique de l'adage populaire « les amis de mes amis sont mes amis ». Une grande transitivité peut indiquer la présence de cliques (comme défini à la section 3.1), donc de groupes de personnages qui apparaissent souvent ensemble. Dans l'article de Stiller *et al.* (2003) par exemple, le coefficient de transitivité calculé sur différentes pièces de Shakespeare permet de conclure à une présence de cliques qui excède largement les observations faites sur des réseaux générés aléatoirement (voir l'article pour les détails d'implémentation), ce qui laisse entendre que les personnages de fiction de ces œuvres ont tendance à s'organiser en cliques.

Mesures sur les nœuds

Les mesures suivantes sont calculées sur chaque nœud et décrivent ses arêtes, les chemins qui le traversent ou encore la connectivité de ses voisins. Il existe tout un éventail de mesures de centralité (dont on fait varier la définition ou la pondération selon les usages) ; les deux versions présentées ici figurent parmi les plus utilisées dans l'interprétation des réseaux de personnages.

Degré et force des nœuds

Définition 10. *On définit le degré d'un nœud comme le nombre d'arêtes qui lui sont associées. Si le réseau est orienté, il est possible de spécifier le degré entrant (ou le degré sortant) d'un nœud comme le nombre d'arêtes qui pointent vers le nœud concerné (ou qui en partent).*

Définition 11. *La force d'un nœud, aussi appelée degré pondéré, est la somme des poids des arêtes rattachées à ce nœud.*

Le degré (ou la force) est la mesure la plus simple applicable à un nœud. Cette notion est très utile pour établir un classement des nœuds au degré le plus haut notamment, ce qui permet d'identifier les personnages qui interagissent beaucoup avec les autres,

qui multiplient les relations. Cela contribue à déterminer les personnages principaux, avec l'aide des mesures de centralité décrites ci-dessous.

Centralité de proximité

Définition 12. *La centralité de proximité d'un nœud est l'inverse de sa distance au reste du réseau. Plus précisément, la centralité de proximité est la réciproque de la somme des distances entre ce nœud et tous les autres. La formule mathématique associée est la suivante :*

$$C_P(x) := \frac{1}{\sum_y d(x,y)},$$

où x et y sont des nœuds, et $d(x,y)$ est la distance entre x et y.

Comme son nom l'indique, la centralité de proximité d'un personnage indique sa proximité avec le reste du réseau. Calculer la centralité de proximité de tous les nœuds d'un réseau peut permettre d'en trouver le centre, ce qui est également une approche utilisée dans la recherche pour définir un personnage principal. Dans l'étude de Masías *et al.* (2016) visant à déterminer qui de Romeo ou Juliet est le personnage principal dans la pièce de Shakespeare, la centralité de proximité figure parmi les mesures de centralité utilisées pour répondre à cette question : cette mesure tend à faire de Juliet le personnage le plus important, alors que les autres résultats mettent davantage l'accent sur Romeo.

Centralité intermédiaire

Définition 13. *La centralité intermédiaire d'un nœud est la proportion de plus courts chemins qui passent par ce nœud (sur tous les plus courts chemins entre deux nœuds du réseau). Mathématiquement, la formule est la suivante :*

$$C_I(x) := \frac{\sum \sigma_{st}(x)}{\sum \sigma_{st}},$$

où x, s et t sont des nœuds, σ_{st} est le nombre total de chemins le plus court de s à t, $\sigma_{st}(x)$ est le nombre de chemins le plus court entre s et t passant par x et les sommes $\sum$ sont calculées sur tous s, t tels que $s \neq x \neq t$.

Les nœuds ayant le plus grand score de centralité intermédiaire représentent des personnages qui servent de lien entre différentes communautés : sans eux, certaines cliques n'auraient peut-être plus de communication entre elles, c'est pourquoi on leur attribue parfois un rôle de « pont » ou de messager. Beveridge et Shan (2016) identifient ainsi Tyrion Lannister comme le personnage ayant la plus grande centralité intermédiaire à travers les huit saisons de la série *Game of Thrones*, et le décrivent comme un personnage qui tisse des connexions entre des régions distinctes du réseau.

Excentricité

Définition 14. *L'excentricité d'un nœud est sa distance maximale au reste du réseau, c'est-à-dire la longueur maximale des chemins les plus courts qui partent de ce nœud.*

Cette notion d'excentricité peut s'étendre à tout le graphe, en calculant l'excentricité moyenne de tous les nœuds, et en indiquant ainsi l'éloignement global de chaque personnage par rapport au reste du réseau.

Transitivité des nœuds

Définition 15. *La transitivité d'un nœud est la proportion des voisins de ce nœud qui sont connectés entre eux. Plus précisément, on définit la transitivité T_x du nœud x comme*

$$T_x = \frac{\lambda_G(x)}{\tau_G(x)},$$

où $\lambda_G(x)$ est le nombre de sous-graphes du graphe G contenant 3 arêtes et 3 nœuds dont x, et $\tau_G(x)$ est le nombre de sous-graphes de G contenant 2 arêtes et 3 nœuds dont x, et dont les deux arêtes sont attachées à x.

Similaire à la transitivité globale (définition 9), la transitivité d'un nœud décrit la propension de voisins d'un personnage à être voisins entre eux. Les personnages qui ont une transitivité basse font souvent partie de plusieurs cliques, ce qui en fait des personnages importants pour l'histoire parce qu'ils relient différents arcs narratifs. De manière plus générale, on remarque que dans les réseaux de personnages, la transitivité est souvent inversement proportionnelle au degré (comme chez Alberich *et al.* [2002] dans le cas de l'univers *Marvel*), ce qui signifierait que les personnages secondaires forment des petits groupes, alors que les personnages principaux sont connectés à beaucoup de groupes différents.

4 Bases informatiques

Un des enjeux de la recherche autour des réseaux de personnages est l'automatisation de leur extraction. Si plus d'une étude se base sur des réseaux construits à la main (voir par exemple les travaux de Kydros et Anastasiadis [2015] ou de Bazzan [2020]), le processus demande un temps considérable, notamment dans le recensement des interactions au fil de l'œuvre, et la plupart des travaux tendent à générer des réseaux de personnages de manière computationnelle.

Pour construire les études de cas pratiques présentées aux chapitres 6 et 8 de cet ouvrage, je me concentrerai exclusivement sur des données textuelles, en m'appuyant sur différentes techniques de traitement automatique du langage (TAL, ou NLP en anglais pour Natural Language Processing). Il existe également des méthodes d'extraction d'informations pour des données visuelles (pour la bande dessinée par exemple, ou pour des détections de visage dans des vidéos) ou sonores, mais elles font appel à des technologies plus coûteuses en ressources, plus expérimentales et moins fiables en matière de résultats.

Ce chapitre s'intéresse donc en particulier aux différentes étapes nécessaires à l'extraction automatique d'un réseau de personnages (section 4.1), ainsi qu'à la reconnaissance d'entités nommées (section 4.2), très utile pour la détection de personnages dans du texte.

En complément de ces deux grands thèmes, deux annexes permettent une plongée plus concrète dans le code produit pour ce travail. Dans l'annexe A, on évalue les forces et les faiblesses de deux grands modèles de reconnaissance d'entités nommées sur la détection de personnages dans des romans. Plus conséquente, l'annexe B présente un module en Python nommé « Charnetto » : ce module a été codé spécifiquement pour générer les différents réseaux de personnages utilisés dans ce livre, et il est disponible en libre accès. Il n'est pas nécessaire de comprendre son fonctionnement pour appréhender la suite de l'ouvrage, mais ces deux annexes pourront intéresser les esprits plus computationnels et les personnes qui désireraient approfondir ces recherches.

4.1 Extraction de réseaux

Les différentes étapes du processus d'extraction sont les suivantes (voir figure 4.1) :

1. Récupération des données
2. Identification et désambiguïsation des personnages
3. Détection des interactions
4. Génération du réseau

Une excellente schématisation de ce processus est à retrouver dans la revue de Labatut et Bost (2019). Notons que les trois premières étapes dépendent profondément de la nature de l'objet étudié, puisqu'il s'agit d'en extraire des informations, alors que la génération du réseau peut se faire de la même manière indépendamment des œuvres choisies (dès lors que l'on possède une liste de personnages et d'interactions, on peut construire les nœuds et les arêtes sans être influencé par les spécificités du média d'origine).

Il est également important d'avoir à l'esprit que les modèles d'extraction de données textuelles sont entraînés principalement sur un registre de langue plutôt informatif ou factuel (articles sur



FIGURE 4.1 Les étapes d'extraction d'un réseau de personnages.

Wikipédia, transcriptions téléphoniques, extraits de journaux, etc.). Or la prose littéraire ne suit pas toujours les mêmes règles d'écriture (voir par exemple les articles d'Elson [2012] ou Givon [2007]) : taille des textes bien plus longue, noms propres parfois inventés ou tirés de noms communs, présence de dialogues dans une structure propre à la fiction, autant de spécificités qui peuvent compliquer la tâche des outils numériques à disposition.

La première étape consiste à récupérer des données, au format textuel (comme dans cet ouvrage) ou dans d'autres formats (audio, image et/ou vidéo) selon les paramètres choisis pour sa recherche.

Pour la deuxième étape, le but est de repérer toutes les mentions de personnages au sein du texte. On peut s'appuyer par exemple sur des techniques de Named Entity Recognition (NER), présentées ci-dessous, à la section 4.2, pour capturer les occurrences des noms propres. Si l'on désire être plus précis dans le recensement des apparitions des personnages, on peut combiner cette stratégie avec des méthodes ciblées sur les pronoms et les anaphores (la difficulté étant alors de les désambiguïser).

Toujours à l'étape 2, on tente ensuite de regrouper toutes ces occurrences en une liste de personnages : dans la suite de ce travail, on appelle « alias » les différentes appellations d'un personnage, et il s'agit ici de regrouper les alias qui désignent un même personnage pour constituer cette liste. Par exemple, « Jeanne d'Arc » pourra être parfois appelée simplement « Jeanne », ou

« la jeune fille », ou « elle », et pour pouvoir correctement construire le réseau de personnages, il faut d'une part identifier les personnages, et de l'autre déterminer pour chaque occurrence le personnage qui y est associé. Cette tâche se complexifie avec l'augmentation des « types » de mentions récupérées à l'étape 1, les pronoms et les anaphores étant bien plus difficiles à désambiguïser que les noms propres (voir par exemple la revue de Sukthanker *et al*. [2020] pour un tour d'horizon du problème de résolution d'anaphores et de coréférences, encore bien présent dans la recherche actuellement). Certains s'appuient sur les codes narratifs de la fiction, comme Vala *et al*. (2015) qui établissent une série de règles pour éliminer progressivement des hypothèses (par exemple, s'il y a un dialogue, deux répliques consécutives ne peuvent pas être attribuées au même personnage). Aucune de ces méthodes n'est totalement fiable, bien entendu, et l'adaptation d'outils de TAL à des œuvres de fiction peut encore faire l'objet d'améliorations importantes.

Ensuite, l'étape 3 permet de recenser les interactions entre ces différents personnages désormais identifiés. Si l'on reprend les types d'interactions présentés à la section 3.2, les méthodes d'extraction augmentent en complexité selon les interactions recherchées. Pour les cooccurrences, la démarche est plutôt facile : on détermine une fenêtre d'observation (paragraphe, page, nombre de lignes) et on localise chaque occurrence pour identifier celles qui sont présentes dans la même fenêtre. La question des conversations est déjà plus épineuse : de nombreux travaux tentent de délimiter correctement les dialogues dans la fiction et d'attribuer chaque réplique au bon personnage et au bon destinataire (voir par exemple les travaux de Cuesta-Lazaro *et al.* [2022] ou de O'Keefe *et al.* [2012]), mais là encore, il s'agit d'un problème ouvert avec des résultats pas toujours probants. Les mentions sont un cas un peu intermédiaire, dans lequel il faut correctement identifier les répliques et les locuteurs, mais sans avoir besoin de

détecter les destinataires, puisque l'on construit les arêtes de la personne qui parle vers la personne qui est mentionnée dans sa réplique. Et pour ce qui est des actions directes et des affiliations, leur recensement demande une compréhension fine de la langue si l'on veut éviter de travailler avec une liste exhaustive de verbes à détecter ; dans le cas des actions directes, on peut par exemple recenser des verbes d'action qui induisent en général un sujet et un objet tous deux humains : aider (Marie aide son amie), embrasser (il embrasse son frère), etc.

Enfin, l'étape 4 consiste en la génération du réseau de personnages, sur la base de toutes les informations réunies et structurées dans les étapes précédentes.

4.2 Reconnaissance d'entités nommées

L'étape d'extraction des personnages, lorsqu'elle est faite de manière automatisée, s'appuie généralement (au moins en partie) sur la reconnaissance d'entités nommées. Officialisé pour la première fois lors de la Sixth Message Understanding Conference (MUC-6) en 1995, qui était alors centrée sur l'extraction d'informations dans un texte non structuré, le terme d'entité nommée est défini comme une unité d'information, désignant un objet du monde réel et possédant un nom précis (comme des personnes, des lieux, des organisations, des monnaies, etc.). Les entités nommées sont jugées essentielles pour l'extraction d'informations et le champ de la Named Entity Recognition (NER) rejoint dès lors les nombreux centres d'intérêt de l'étude du traitement de langage naturel.

NER et textes de fiction

Si les algorithmes de NER sont aujourd'hui très performants sur des articles de journaux ou des pages de Wikipédia (voir par exemple la revue de Sun *et al.* [2018]), la fiction présente encore

certains défis que la recherche n'a pas totalement maîtrisés. Pour le cas qui nous intéresse, à savoir les noms de personnages, on peut en effet trouver des noms inventés (que l'algorithme aura du mal à catégoriser correctement entre lieu, personne ou organisation s'il ne les a jamais rencontrés dans d'autres contextes), des surnoms, adjectifs ou noms communs utilisés comme des noms propres (comme les prénoms des sept nains dans *Blanche-Neige*), autant de cas de figure rarement croisés dans la prose journalistique qui a servi de source d'entraînement aux modèles de NER et qui peuvent les mettre en difficulté.

Il n'existe pas aujourd'hui de choix par défaut qui serait unanimement reconnu comme *la* meilleure solution pour détecter les personnages d'un roman. Pour autant, tous les modèles ne sont pas équivalents, et certains fonctionnent mieux que d'autres pour ce type de données. Sur la question de la qualité des résultats, l'article de Stanislawek *et al.* (2019) illustre bien la diversité des options à disposition : pour mesurer plus en détail la précision des algorithmes de NER, les chercheurs proposent un découpage en onze catégories d'erreurs observées (qui vont des erreurs d'annotation aux usages insolites) et évaluent la performance de cinq modèles qui ont représenté des avancées significatives dans le domaine de la NER sur ces différentes catégories. Les résultats permettent ainsi d'identifier les forces et les faiblesses de ces modèles et dans notre cas, de repérer l'algorithme le plus performant sur les textes de fiction.

Comme on peut le constater dans l'article, le modèle qui semble le plus robuste aux différents types d'erreurs est celui de Flair (Akbib *et al.*, 2019), une infrastructure développée par le groupe de recherche Zalando, basée sur PyTorch (Paszke *et al.*, 2019). Les deux types d'erreurs qui font le plus obstacle à son fonctionnement sont les fautes de frappe et les incohérences générales, deux cas de figure qui ne devraient pas apparaître dans des œuvres de fiction (qui sont en général dépourvues de fautes de

frappe et conçues pour ne pas perdre le lectorat avec des appellations qui désigneraient tantôt un lieu, tantôt une personne).

Toutefois, cet article s'appuie sur le jeu de données de CoNLL 2003 (Tjong Kim Sang et De Meulder, 2003), composé d'articles de journaux, et ne teste donc pas les différents modèles sur des romans. L'annexe A s'emploie ainsi à vérifier ces résultats sur des œuvres de fiction, en comparant Flair avec spaCy, un autre modèle très populaire et absent de l'article de Stanislawek *et al.*

Deuxième partie

Narrations plurielles et interactives

5 Approche théorique

Pour reprendre la définition d'Aarseth (1997), la narration interactive (qu'il choisit d'appeler *ergodique*) est une forme de narration qui nécessite un effort non trivial afin d'être traversée. Comme présenté par Riedl et Bulitko (2013), les utilisateurs et utilisatrices de cette expérience narrative participative ont la possibilité d'influencer la progression du scénario par des choix ou des actions, s'écartant ainsi de ce que l'on appellera dans ce travail la narration « linéaire »[5]. Dans cette catégorie, on retrouve notamment les livres dont vous êtes le héros, les fictions interactives et certains jeux vidéo, selon l'approche narrative choisie par l'équipe de développement.

La narration interactive fait déjà l'objet d'un certain nombre de recherches, que ce soit sur la manière de les générer (avec l'élaboration d'outils automatisés comme Payador [Gónora *et al.*, 2024] ou Scenecraft [Kumaran *et al.*, 2023]), sur la forme des différents chemins possibles, sur l'impact des choix sur l'immersion et l'engagement des utilisateurs et utilisatrices ou encore sur l'équilibre à trouver entre liberté de choix et contrôle de la structure. Il semble toutefois que ces supports, comme les livres dont vous

5 *Linéaire* signifie ici sans bifurcations dans la narration, cela n'implique pas forcément que l'histoire est racontée dans un ordre chronologique

êtes le héros ou les jeux vidéo, soient encore peu examinés selon une perspective d'étude de la narration comme on peut le faire par exemple dans la littérature avec des approches de *close* et *distant reading* [6].

La section suivante, qui présente le modèle théorique du flux narratif, permet d'ouvrir la porte de la narration interactive à tous les outils de *distant reading* déjà disponibles pour des narrations linéaires. Si le contexte de cet ouvrage s'intéresse particulièrement aux réseaux de personnages, le flux narratif propose une perspective plus large, qui pourra intéresser également des chercheurs et chercheuses qui désirent s'approprier cet angle narratif avec des visées de stylométrie, de statistique textuelle, ou de toute autre approche liée au texte et à la narration.

5.1 Flux narratif

Pour appliquer des techniques de *distant reading* à des narrations interactives, je propose une étape intermédiaire : la construction d'un *flux narratif*, un modèle permettant de schématiser et d'organiser les informations contenues dans l'œuvre étudiée selon les différents chemins possibles de la narration. Ce modèle permet ensuite de simuler différents parcours de l'œuvre et de les aborder avec des outils d'analyse réservés jusqu'ici aux narrations linéaires.

Unité narrative

La première étape dans la construction d'un flux narratif est de segmenter l'œuvre étudiée en un certain nombre d'*unités narratives*. Une unité narrative sera donc trivialement une « portion »

[6] Pour de premières explorations dans cette direction, voir par exemple le travail de Kontopoulou *et al.* (2013) qui approchent les livres dont vous êtes le héros grâce au concept mathématique d'ergodicité et visent à regrouper les parties de l'histoire en différents concepts.

de l'œuvre, de taille et de nature variables selon le média et la granularité désirée. Si l'on travaille sur un roman, on voudra peut-être diviser ce roman en chapitres, en paragraphes ou en phrases (à la manière des « blocs » utilisés dans Charnetto dans l'annexe B), en fonction du type d'analyses que l'on souhaite effectuer, et chaque élément de ce découpage représentera une unité narrative dans le flux narratif. Dans le cas d'un film, on optera peut-être plutôt pour des scènes ou des portions de cinq minutes, par exemple. Il n'est pas nécessaire que les unités narratives d'une même œuvre soient de taille identique. Pour qu'un découpage soit correctement défini, il faut que l'union de toutes les unités narratives représente une partition de l'œuvre. Et pour le cas où ladite œuvre est une narration interactive, on va également chercher à découper les unités aux endroits qui représentent un embranchement, afin de préserver (et de schématiser) l'arborescence des possibilités de narrations [7].

Pour garder une trace des informations pertinentes contenues dans chaque portion d'œuvre ainsi agrégée, on associe à chaque unité narrative un ensemble de caractéristiques. Là encore, la nature et le nombre de ces caractéristiques peut varier selon les cas : il peut s'agir d'informations directement contenues dans l'unité narrative, comme une liste de tous les noms propres qui y sont cités, ou le temps d'apparition d'un personnage précis à l'écran, mais les caractéristiques peuvent également provenir de métadonnées comme le numéro de page d'un livre, la scène d'un film ou le niveau d'un jeu vidéo, qui peuvent servir notamment à indiquer l'emplacement de l'unité narrative dans l'œuvre d'origine.

7 Pour des ouvrages traitant de la segmentation des narrations interactives, voir les travaux de Caïra sur le chapitrage (2020) et sur les arborescences (2019).

Transitions

Lorsque toutes les unités narratives sont définies, il reste à les organiser entre elles pour les ordonner. On introduit alors des *transitions* qui ont pour point de départ une unité et pour point d'arrivée une autre unité (à moins de créer une boucle en revenant sur la même unité, ce qui peut se produire par exemple dans des jeux vidéo qui permettent d'avoir la même interaction plusieurs fois avec un personnage), et dont le sens indique l'ordre d'apparition de ces unités dans les différents parcours de l'œuvre. Dans le cas d'un roman ou d'un film classique, par exemple, les unités narratives vont être arrangées en une seule longue chaîne, selon l'ordre de lecture ou de montage des portions narratives associées. Chaque unité sera ainsi l'arrivée et le départ d'une seule transition, à l'exception des deux extrémités de la chaîne. Mais il est également possible qu'une unité serve de point de départ à plusieurs transitions sortantes, ou que plusieurs transitions entrantes arrivent sur la même unité. Ce sera notamment le cas dans les livres dont vous êtes le héros, dans lesquels le lecteur peut être amené à faire des choix (créant ainsi plusieurs transitions sortantes), qui mèneront peut-être plus tard à la même conclusion (formant plusieurs transitions entrantes sur la même unité d'arrivée).

On associe à chaque transition une probabilité d'être choisie, sous forme de score entre 0 et 1. Par défaut ou lorsque la probabilité n'est pas connue, on considère que toutes les transitions sortantes d'une unité sont équiprobables. Dans certains cas, il est possible de savoir si un choix est plus populaire que les autres, ou s'il a une probabilité plus grande d'être sélectionné. Après la sortie du film *Bandersnatch* réalisé par David Slade en 2018, Netflix a par exemple publié les statistiques pour certains des scénarios plébiscités ou non par le public, et de nombreux jeux vidéo essentiellement narratifs proposent également un écran qui situe le joueur ou la joueuse par rapport au reste de la communauté. Si

l'on souhaite tenir compte de ces tendances dans le flux narratif, on peut choisir de fixer des probabilités de transition conformes à ces données. La somme des probabilités de toutes les transitions sortantes d'une unité narrative doit toujours valoir 1 : sinon, cela voudrait dire qu'il manque des chemins ou que l'œuvre n'a été que partiellement représentée par les unités narratives (ce qui contredit leur définition de partition de l'œuvre).

Flux narratif

Le flux narratif est un graphe dont les nœuds sont des unités narratives, et dont les arêtes sont des transitions, comme dans la figure 5.1. Il s'agit donc d'un réseau orienté qui se rapproche de la logique des *réseaux de flot* en théorie des graphes : comme l'expliquent par exemple Goldberg, Tardos et Tarjan (1989), les réseaux de flot servent à modéliser des transports de matière (fluides, voitures, marchandises, etc.) en partant d'un point de départ, appelé source, pour converger vers un point d'arrivée, appelé puits. Chaque arête (que l'on nomme « arc » dans un réseau de flot) possède une caractéristique de capacité, représentant la quantité de matière qu'elle peut accueillir. Utile pour calculer notamment le flot maximal supporté par le réseau, le chemin le plus court ou d'autres problèmes plus complexes recensés notamment par Ahuja *et al.* (1993), la théorie des réseaux de flot est un champ de recherche important qui trouve des applications dans des domaines aussi variés que la chimie, l'ingénierie ou la communication.

Dans le présent contexte, on peut aisément considérer que tout flux narratif possède exactement une unité narrative de degré entrant (respectivement sortant) 0, qui correspond à la « source » (respectivement au « puits ») du réseau de flot, et donc au début (respectivement à la fin) de notre consommation de l'histoire. S'il devait y avoir plusieurs fins possibles, il suffit d'ajouter une unité narrative supplémentaire qui servira de fin globale,

ainsi qu'une transition de chacune des fins existantes vers cette nouvelle unité, qui deviendra ainsi le puits (et le raisonnement est le même pour une œuvre qui aurait plusieurs points de départ).

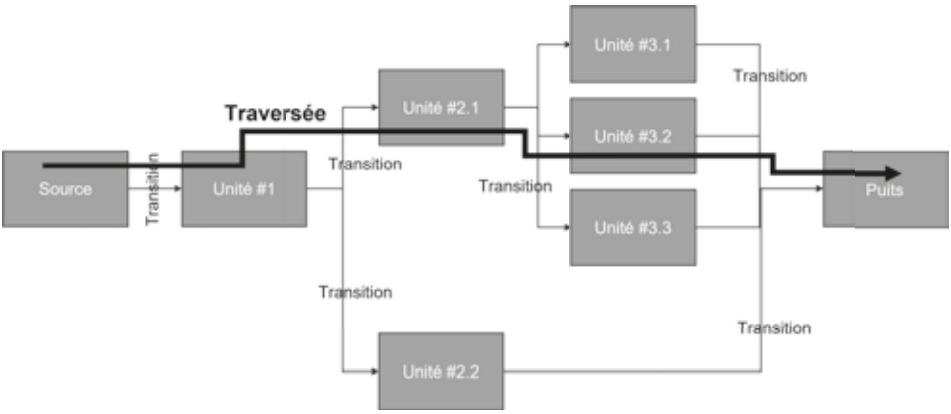


FIGURE 5.1 Exemple de flux narratif.

Traversées

Le flux narratif d'une œuvre présente sa structure globale, un genre de squelette des embranchements possibles de l'histoire et de sa complexité générale. Une fois que cette structure est en place, on peut la parcourir en générant des *traversées*, c'est-à-dire des chemins de la source jusqu'au puits. Chaque traversée correspond à un scénario, une expérience narrative parmi les possibilités offertes par l'œuvre originale.

Pour des œuvres linéaires comme des romans ou des films classiques, dont on a déjà établi que les flux narratifs ressembleraient à des chaînes, il n'existe qu'une traversée possible le long de la chaîne, donc une manière d'expérimenter l'histoire

selon l'ordre proposé [8]. Par contre, lorsque l'on étudie des narrations interactives, en préservant les embranchements induits par les différentes options de l'histoire, on obtient un flux narratif qui offre plusieurs possibilités de traversées (on pourra toutefois choisir d'ignorer certains embranchements selon la nature des analyses, par exemple s'ils se rejoignent rapidement sans provoquer de changement majeur dans la structure de l'intrigue). Plus formellement, s'il existe au moins une unité narrative dont le degré sortant est supérieur à 1, il y aura au moins deux traversées possibles différentes du flux narratif concerné. Si le flux narratif contient au moins un cycle, c'est-à-dire un chemin qui revient sur son point de départ, le nombre de traversées différentes sera infini, puisque l'on peut tourner autant de fois que l'on veut à l'intérieur du cycle. Ce cas extrême peut se produire dans les narrations interactives, par exemple dans un jeu vidéo où l'on pourrait s'adresser à un personnage non joueur autant de fois que l'on veut, et où il produirait la même réponse à chaque itération.

Fonction d'agrégation

La dernière étape de la création d'un flux narratif est l'emploi d'une fonction d'agrégation. C'est cette fonction qui, pour une traversée choisie, va collecter et rassembler toutes les caractéristiques des unités narratives visitées dans cette traversée. La définition formelle de la fonction d'agrégation dépend donc notamment du format des caractéristiques associées à chaque unité narrative. Qu'il s'agisse d'additionner des valeurs ou de faire des unions d'ensembles de mots, la démarche générale reste la réunion de toutes les informations qui concernent la traversée analysée, et donc la version de la narration sur laquelle on se penche. Cette focalisation sur une traversée, couplée à cette

[8] Évidemment, on est libre de choisir de lire un roman dans le désordre, mais ce genre de cas n'est pas traité dans ce modèle.

collecte d'informations ciblées, permet de ramener la situation au format traditionnel de narration linéaire, pour lequel il existe déjà de nombreux outils d'analyse.

Si l'on désire par exemple observer la variation de la taille du vocabulaire (en nombre de mots distincts) à travers différentes itérations d'un même livre dont vous êtes le héros, la construction du flux narratif permettra de générer une multitude de traversées, et la fonction d'agrégation se chargera ensuite de rassembler l'ensemble des mots rencontrés lors de chaque traversée. On pourra ainsi obtenir la taille du vocabulaire pour chaque traversée étudiée, et comparer ces différentes valeurs à notre convenance (en calculant la taille moyenne, l'écart-type, en identifiant la traversée qui contient le vocabulaire le plus abondant, etc.). C'est l'objet de la section 5.2 ci-dessous, qui met en pratique cet exemple. De manière générale, le flux narratif est un outil qui, au-delà de la question du réseau de personnages, permet d'appliquer toutes sortes de techniques de *distant reading* à des narrations interactives.

5.2 Illustration : taille de vocabulaire

Avant de faire le lien entre flux narratif et réseaux de personnages à la section 5.3, il peut être utile de marquer cette polyvalence du flux narratif à l'aide d'un exemple pratique qui prend un angle de recherche tout à fait différent : la taille du vocabulaire des différents chemins possibles dans un livre dont vous êtes le héros. Comme annoncé à la section précédente, l'idée est de parcourir tous les chemins possibles du livre, de recenser tous les mots uniques rencontrés dans ces chemins et de comparer les tailles de ces vocabulaires.

Le roman qui sert de base à cet exemple est *Consider the Consequences*, de Doris Webster et Mary Alden Hopkins. Publié en 1930, il s'agit du premier roman qui utilise des mécaniques de livres dont vous êtes le héros, à savoir des choix laissés aux lecteurs

et lectrices et des renvois aux chapitres à lire en fonction de ces choix. Il est divisé en trois parties consacrées à des personnages différents (Helen, Jed et Sanders) : dans le périmètre de cet exemple, seule la première partie est prise en compte. Cette partie est constituée de 33 segments de texte, dont 17 fins possibles pour un unique point de départ. Pour des études plus poussées sur ce roman et d'autres livres dont vous êtes le héros, voir par exemple les travaux de Pante (2020).

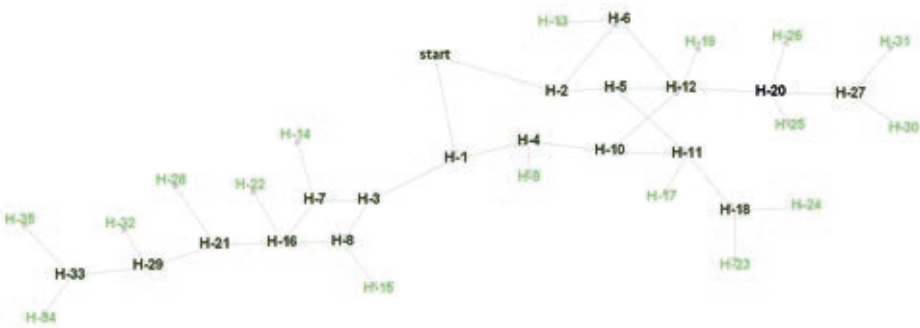


FIGURE 5.2 Flux narratif pour la première partie de *Consider the Consequences*. Les fins possibles sont en vert. Notons que pour clarifier l'illustration, le nœud de « puits » (qui sert d'arrivée à tous les nœuds finaux) a été retiré.

La construction du flux narratif associé à cette partie a été faite avec Python, en créant une unité narrative pour chaque segment, en lui associant la portion de texte correspondante et en reliant les unités par des transitions lorsque le livre propose un renvoi à un autre segment, comme illustré à la figure 5.2. L'utilisation du module `networkx` (Hagberg *et al.*, 2008) permet ensuite de générer toutes les traversées possibles de ce flux [9] : il en existe 35, qui parcourent entre 4 et 9 unités narratives différentes.

9 Le code associé est disponible sur le répertoire character_network.

Si l'on s'intéresse à la taille du vocabulaire rencontré sur chaque traversée, on s'aperçoit (en utilisant la tokenisation fournie par spaCy [Honnibal et Montani, 2017]) que l'on découvre en moyenne 715,05 mots différents en lisant cette partie du roman, avec un écart-type de 79,09.

La traversée qui propose le plus grand vocabulaire est celle qui passe par H-1, H-4, H-10, H-12, H-20, H-27 et H-31 ; avec un total de 878 mots différents, elle dépasse largement les quatre traversées les plus longues (qui contiennent entre 742 et 761 mots).

À l'inverse, celle qui a le vocabulaire le plus restreint est celle qui passe par H-1, H-4 et H-9, avec seulement 538 mots rencontrés (contre 580 pour la deuxième traversée la plus courte qui passe par H-2, H-6 et H-13).

Ces résultats sont néanmoins fortement influencés par la taille des textes : pour corriger ce problème, Jeanson *et al.* (2024) recensent différentes approches de rééchantillonnage, dont la mesure de « variété attendue de sous-échantillon » (*expected subsample variety*) qui donne la taille moyenne du vocabulaire pour des sous-échantillons de taille donnée. Pour cet exemple, la variété attendue pour des sous-échantillons de 1000 mots permet d'obtenir des tailles moyennes de vocabulaire allant de 372,07 à 401,89 : la traversée avec le plus petit vocabulaire est toujours celle qui passe par H-1, H-4 et H-9, mais la traversée offrant le vocabulaire le plus varié est cette fois celle qui passe par H-2, H-5, H-12, H-20, H-27, H-30, bien que ce ne soit pas la traversée qui propose le texte le plus long. À l'image du « *golden path* » (Bateman, 2006) des jeux vidéo, désignant le chemin qui passe le mieux à travers tous les événements considérés comme cruciaux, on pourrait se demander si le chemin offrant le plus de diversité lexicale est également celui ressenti comme le plus « riche » en matière de contenu pour les personnes qui lisent ce roman (bien que cette question dépasse le périmètre actuel de ce travail). Pour des observations plus approfondies, la liste complète des traversées, triées selon cette mesure, est à retrouver sur le répertoire.

Ces quelques résultats sont rapides à calculer une fois la structure du flux narratif établie. En passant d'une narration interactive à une collection de traversées, on peut ainsi traiter chacune de ses traversées comme un texte « classique » et produire toutes les analyses de *distant reading* désirées, avec cette possibilité de comparer les résultats entre toutes les possibilités de parcours de l'œuvre.

5.3 Réseau moyen

Dans le cas des réseaux de personnages, chaque traversée donne lieu à un réseau de personnages (différent ou non des autres), leur accumulation produisant ainsi une collection de réseaux. Une stratégie possible pour observer la tendance générale ainsi que les éventuelles fluctuations d'importance des personnages et de leurs interactions est de produire un « réseau moyen » à partir de ces différents réseaux (le réseau médian en serait une autre).

Dans ce réseau moyen, le poids d'un nœud devient donc la moyenne des poids dudit nœud dans les différents réseaux. Notons que certains nœuds pourraient être totalement absents d'une partie des réseaux, si les personnages ont une présence optionnelle pour la progression de l'histoire. Pour faire émerger ce phénomène, il est possible de fixer leur poids à 0 lorsqu'ils sont absents du réseau, ou de marquer cette absence à l'aide d'un indicateur qui permet ensuite de filtrer les nœuds toujours présents de ceux qui ont une présence variable. On peut également calculer l'écart-type, la valeur minimale et la valeur maximale de chaque nœud du réseau moyen, et ajouter ces valeurs aux attributs du nœud. En agissant de la même manière pour les arêtes, on obtient un réseau unique, chargé de toutes les informations contenues dans la collection de réseaux possibles.

Dans un flux narratif sans cycle, le nombre de traversées possibles est fini, il est donc théoriquement possible de générer tous les réseaux de personnages correspondants. Dans la plupart des

cas, leur nombre sera toutefois trop élevé pour espérer calculer les résultats dans des délais raisonnables, à cause de la puissance de calcul qu'un tel processus demande. Il faut donc trouver une autre manière de produire un réseau moyen représentatif.

Si l'on parvient à générer, par exemple, 100 000 traversées aléatoires du flux narratif, nos moyennes seront-elles représentatives ? On peut s'en assurer avec un calcul des moyennes et variances théoriques basé sur la théorie des chaînes de Markov. Les explications qui suivent prennent obligatoirement un angle très mathématique : elles peuvent aisément être survolées sans perte de compréhension du propos général. Les résultats seront utilisés principalement à la sous-section 6.5.1 pour contrôler la qualité des résultats obtenus dans l'étude de cas des jeux *Life is Strange*.

Espérance et variance théoriques

La variable aléatoire X_P, définie comme le nombre d'occurrences d'un personnage P dans l'œuvre dont est tiré le flux narratif, peut être exprimée comme

$$X_P = \sum_{i=1}^{n} N_i o_i^P,$$

où N_i est le nombre de passages par le nœud i et o_i^P est le nombre d'occurrences du personnage P dans le nœud i, n étant le nombre de nœuds dans le flux narratif.

Or, pour calculer l'espérance (et donc, la moyenne théorique) de cette variable aléatoire, il faut trouver l'espérance des N_i puisque par propriété de l'espérance, on a

$$\mathbb{E}(X_P) = \mathbb{E}(\sum_{i=1}^{n} N_i o_i^P) = \sum_{i=1}^{n} o_i^P \mathbb{E}(N_i).$$

De la même manière, la variance théorique de X_P peut se calculer de la manière suivante :

$$\mathrm{Var}(X_P) = \sum_{ij} \mathrm{Cov}(N_i o_i^P, N_j o_j^P) = \sum_{ij} o_i^P o_j^P \mathrm{Cov}(N_i, N_j).$$

Considérer le flux narratif comme une chaîne de Markov, c'est-à-dire un processus stochastique dont la probabilité de l'état futur ne dépend que de l'état présent, nous permet de calculer l'espérance des N_i (définis justement comme le nombre de passages en chaque nœud du flux narratif), et ainsi l'espérance de X_P (et il en va de même pour la variance). Plus précisément, on considère que :

— les états possibles sont les nœuds du flux narratif,
— les transitions d'un état à l'autre sont les arêtes du flux,
— la probabilité des transitions est la probabilité associée à chaque arête.

En partant du flux narratif, on construit donc sa matrice d'adjacence (voir définition 3, page 25). En considérant que la chaîne de Markov va de l'état s (la source du flux) à l'état t (le puits du flux), ce dernier état étant « absorbant » (c'est-à-dire qu'une fois arrivé sur t, on n'en repart plus), on a donc que la ligne de la matrice d'adjacence correspondant à l'état t sera remplie de zéros. On transforme la matrice d'adjacence en une chaîne de Markov W avec les manipulations suivantes :

— la ligne correspondant à l'état t reste à zéro,
— toutes les autres lignes sont normalisées pour que leur somme soit égale à 1.

On peut alors définir la matrice dite fondamentale $Z := (I - W)^{-1}$ (comme présenté dans le livre de Kemeny et Snell [1976 [1960]]) : cette matrice représente le nombre de fois où l'on se trouve dans l'état j si l'on a commencé à l'état i. Les explications détaillées sur la matrice fondamentale et les résultats évoqués dans cette partie figurent dans l'article de Guex *et al.* (2021) et les références citées. Le calcul des différentes valeurs est présent dans le répertoire character_network.

Par construction, chaque entrée z_{si} de la matrice Z (où s est toujours la source du flux) est égale à l'espérance de N_i. Ainsi, en construisant le vecteur $z_s := (z_{si})$ et le vecteur $o := (o_i^P)$ des

occurrences de P dans le nœud i, on obtient la formule

$$\mathbb{E}(X_P) = o^\top z_s, \tag{5.1}$$

qui nous permet d'obtenir l'espérance des occurrences du personnage P à travers le flux narratif.

Une approche similaire permet de calculer la variance théorique à l'aide de la matrice $V := (2Z - I)$ et du produit composantes par composantes $q := o \odot z_s$, en obtenant finalement la formule suivante :

$$\mathrm{Var}(X_P) = q^\top V o - \mathbb{E}(X_P)^2. \tag{5.2}$$

Ainsi, si l'on désire travailler uniquement avec les moyennes et les variances du réseau de personnages agrégé, on peut s'appuyer sur ces formules et éviter l'étape procédurale de génération d'une quantité importante de réseaux empiriques. Dans le cas où la génération de réseaux serait toutefois souhaitable, comme dans l'exemple pratique du chapitre 6 qui fait également appel aux valeurs minimales et maximales des poids, la moyenne et la variance théorique sont de bons indicateurs de qualité des résultats, comme présenté à la section 6.5. Notons également que le fait d'assimiler le flux narratif à une chaîne de Markov implique que les traversées ne soient pas contraintes par les unités précédemment visitées, ce qui peut poser un problème selon la façon dont le flux a été défini. Dans l'exemple du chapitre 6 toutefois, ces indicateurs ne servent qu'à contrôler les moyennes et variances obtenues empiriquement, et le reste des analyses ne s'appuie pas sur des propriétés de chaînes de Markov.

6 Étude de cas : *Life is Strange*

Passons maintenant à un exemple pratique, pour tester concrètement le modèle théorique développé au chapitre 5 et réfléchir à ses limites et ouvertures, toujours sous l'angle spécifique du réseau de personnages. Pour cela, j'ai choisi la série de jeux vidéo *Life is Strange*, développée par Dontnod Entertainment et Deck Nine et éditée par Square Enix, dont le premier opus est sorti en 2015 et qui sera présentée en détail à la section 6.2. Pour éviter toute confusion sur la terminologie, il est utile de préciser que l'accent principal ne porte pas directement sur *Life is Strange*, mais bien sur l'usage du flux narratif et des réseaux de personnages (appliqué ici à cette franchise). En préambule de cet exemple, la section 6.1 présente quelques recherches préliminaires sur un autre jeu vidéo narratif : *Phoenix Wright: Ace Attorney*.

6.1 Expérience préliminaire : *Phoenix Wright*

Lors de la recherche de scripts de jeux vidéo permettant la construction de réseaux de personnages, un premier jeu a émergé dans le cadre de ce travail : *Phoenix Wright: Ace Attorney*, publié par

Capcom en 2006 dans sa version européenne. Ce jeu essentiellement narratif permet d'incarner un avocat de la défense lors de plusieurs affaires criminelles, en alternant les phases d'enquête sur le terrain et les phases de procès. Un script d'excellente qualité des quatre premiers épisodes (sur cinq) du jeu est disponible sur GameFAQs[10] ; il est rédigé par Dylan Mead qui a également fourni une explication détaillée de sa syntaxe en introduction.

Ainsi, *Phoenix Wright: Ace Attorney* est apparu comme un bon candidat pour de premières expériences de narration interactive pour son côté très dirigé, son texte abondant et son script de bonne qualité (malgré l'absence du dernier épisode). L'objectif étant de se rendre compte de l'impact des options laissées aux joueurs et joueuses sur le réseau de personnages global du jeu, la recherche s'est organisée en plusieurs étapes :

1. parcourir automatiquement le script pour identifier les dialogues soumis à des choix et ceux qui ont lieu dans tous les cas,
2. construire un réseau de personnages « minimal » ne contenant que les dialogues obligatoires,
3. construire un réseau de personnages « maximal » contenant tous les dialogues,
4. comparer ces deux réseaux.

Il est important de noter qu'aucun de ces réseaux ne reflète une expérience concrète de jeu : la structure de *Phoenix Wright* ne permet pas de le terminer en évitant tous les dialogues présentant des choix, de même qu'il est impossible de les parcourir tous au sein d'une même partie. La plupart de ces dialogues à options naissent en réalité de questions posées en boucle jusqu'à ce que la bonne réponse soit donnée (et découlant sur de petites conversations différentes selon les réponses), mais certains d'entre eux

[10] https://gamefaqs.gamespot.com/ds/925589-phoenix-wright-ace-attorney/faqs/42767, consulté le 20.09.2022.

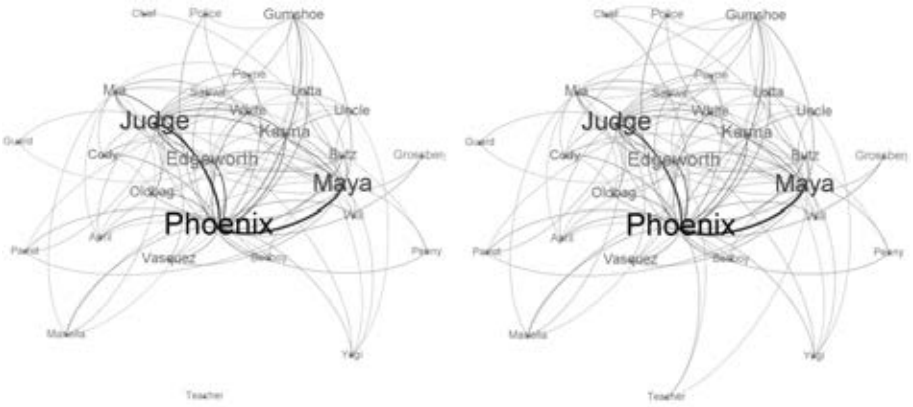


FIGURE 6.1 Réseaux de personnages de *Phoenix Wright: Ace Attorney* : à gauche en prenant uniquement les dialogues obligatoires, à droite en prenant la totalité des dialogues.

ne se présentent qu'une fois et influencent les interactions entre les personnages.

Dans la figure 6.1 qui présente les réseaux ainsi produits [11], on peut immédiatement remarquer que les différences entre ces deux réseaux sont minimes : à part quelques arêtes qui apparaissent dans la version de droite (comme celles qui relient Teacher au reste du réseau, traduisant ainsi le fait que ce personnage n'est présent que dans des dialogues présentant des options), les réseaux sont essentiellement identiques.

Les cooccurrences sont considérées scène par scène : la variation quasi inexistante des poids des arêtes s'explique par le fait que les dialogues contenant des options sont toujours contenus dans une scène plus large, qui semble faire interagir les mêmes personnages en dehors des choix disponibles. Le poids des nœuds est ici défini comme la somme des poids de ses arêtes, ce qui

[11] Le code de ces différentes étapes d'exploration est à retrouver sur le répertoire character_network.

explique également sa très faible variabilité (Phoenix passe par exemple d'un poids de 24 à un poids de 26 lorsque l'on inclut tous les dialogues). Choisir un poids des nœuds basé sur le nombre de répliques de chaque personnage aurait certainement induit une différence plus grande entre les deux réseaux, mais cette étude préliminaire s'est concentrée sur les interactions.

Sachant que les réseaux de personnages associés à des expériences concrètes du jeu se situeront forcément « entre » les deux réseaux générés ici, il semble évident que leur variation sera trop faible pour produire des résultats intéressants sur le plan de l'existence de liens entre les personnages. Les joueurs et joueuses ont bien une influence sur le cours de la partie, mais notre recherche nécessite un jeu qui laisse une liberté non seulement sur le déroulement de l'histoire, mais aussi sur le développement de liens entre les personnages (en permettant de rencontrer ou non certains personnages, ou de multiplier les interactions avec certains d'entre eux) pour produire des réseaux de personnages aux différences plus marquées.

Cette expérience a fourni des pistes de réflexion menant au choix de *Life is Strange* pour les analyses de la suite de ce chapitre, comme le présente la section suivante.

6.2 Choix de *Life is Strange*

Le choix de l'objet d'étude approprié pour cet exemple pratique repose sur différents critères :

— *Un grand nombre de personnages et d'interactions*

Sans un nombre suffisamment important de personnages et d'interactions, même si la narration est interactive, les réseaux de personnages générés n'auront que peu d'intérêt. Pour cette raison, il a fallu éliminer d'emblée les livres dont vous êtes le héros, pourtant attirants au premier abord

par leur nature de source textuelle et leur structure claire. S'il existe des livres dont vous êtes le héros comportant plusieurs personnages et des interactions, la grande majorité d'entre eux fait apparaître plutôt un héros seul, faisant face à des obstacles, des combats ou des choix d'itinéraire, mais rarement à d'autres personnages qui interagissent avec lui, et encore moins entre eux.

— *La disponibilité d'une source textuelle traduisant la narration*

Si les développeurs et développeuses de jeux vidéo utilisent plusieurs types de documents pour décrire les dialogues (script des scénaristes pour la structure de l'histoire, ensemble des répliques d'un même personnage pour les doubleurs et doubleuses, etc.), dans la grande majorité des cas, ils ne sont pas disponibles en libre accès. Une exception notable est le « puzzle document » du jeu vidéo *Grim Fandango*, écrit par son designer Tim Schafer (1996) et utilisé comme support pour proposer son idée à LucasArts. Outre la description des énigmes, ce document raconte tout le scénario initialement prévu et regorge d'informations précieuses sur la genèse de ce projet.

Dans leur article sur la science-fiction (2016), Rochat et Triclot s'interrogent sur les types de ressources textuelles permettant l'étude du contenu narratif des jeux vidéo et s'attardent notamment sur l'hypothèse des textes de solutions en ligne. Malheureusement, il est rare que les dialogues ou cinématiques [12] soient racontés dans ces textes, ce qui en fait là aussi de mauvais candidats pour une étude spécifique sur les réseaux de personnages.

[12] Les cinématiques sont de courtes séquences vidéo, souvent utilisées dans les jeux vidéo pour faire progresser l'histoire.

En revanche, certains jeux très populaires comme *The Witcher 3* ou *The Last of Us* possèdent des scripts intégraux, rédigés par des fans et disponibles gratuitement sur internet (sur les Wiki des jeux, ou dans des forums de solutions comme GameFAQs [13] notamment). Ces ressources textuelles sont rares mais infiniment précieuses, puisqu'elles tiennent compte des différentes options laissées aux joueurs et joueuses et des conséquences portées sur le scénario. L'absence de règles d'uniformisation pour la génération de scripts de jeux pèse sur la capacité à récolter des données en masse, mais s'il s'agit de se pencher sur un jeu en particulier, la logique semi-structurée du texte facilite grandement la génération d'un flux narratif.

— *Un accent mis sur l'importance des choix et sur les liens entre personnages*

Dans l'optique d'analyser une variation au sein des interactions entre les personnages, il semble utile de se concentrer sur des jeux qui revendiquent un scénario s'adaptant aux choix des joueurs et joueuses. D'une part, cela garantit un scénario qui cherche à offrir des expériences diverses et, d'autre part, cela permet une approche quantitative quant à l'impact réel des choix des joueurs et joueuses sur la progression de l'intrigue. En outre, les premières explorations sur *Phoenix Wright: Ace Attorney* (section 6.1) ont montré qu'il était nécessaire que ces choix soient tournés vers les relations entre les personnages, et pas uniquement sur le déroulement de l'histoire, dans le but de produire des réseaux de personnages variés et intéressants à comparer.

À l'issue de cette réflexion et d'un premier inventaire des scripts disponibles, notre choix s'est arrêté sur la franchise de

[13] https://gamefaqs.gamespot.com/, consulté le 19.06.2024.

jeux d'aventure épisodiques *Life is Strange*, pour laquelle j'ai récupéré les scripts des opus suivants [14] :

— *Life is Strange* (2015, 5 épisodes)
— *Life is Strange : Before the Storm* (2017, 4 épisodes)
— *Life is Strange 2* (2019, 5 épisodes)
— *Life is Strange : True Colors* (2021, 5 épisodes)

Le premier opus s'engageait à révolutionner les codes du jeu basé sur les choix et les conséquences en autorisant les joueurs et joueuses à remonter le temps pour agir sur le passé, le présent et le futur. On y contrôle en effet une jeune fille du nom de Max, qui découvre qu'elle a le pouvoir de remonter le temps et qui enquête avec sa meilleure amie Chloe sur la disparition d'une autre adolescente de leur école.

Life is Strange : Before the Storm présente les événements survenus avant l'arrivée de Max dans la région du premier opus, en permettant cette fois de contrôler Chloe. Ce jeu est plus court que les autres titres et n'exploite pas à proprement parler un pouvoir surnaturel. Il se positionne également comme centré sur les choix des joueurs et joueuses.

Life is Strange 2 reprend des codes du premier jeu en changeant le décor et les personnages. On contrôle cette fois Sean, accompagné de son petit frère Daniel, qui s'est découvert un pouvoir de télékinésie. Forcés de fuir leur foyer, les deux protagonistes traversent les États-Unis en direction du Mexique, là aussi dans une narration qui se dit fortement influencée par des choix.

Dans *Life is Strange : True Colors*, on découvre encore d'autres personnages en accompagnant Alex qui a la capacité de manipuler les émotions des autres. À la mort suspecte de son frère, elle

[14] Manquent à cette liste le très court jeu *The Awesome Adventures of Captain Spirit* qui sert d'introduction à *Life is Strange 2* et dont le script n'est pas complet, ainsi que *Life is Strange : Double Exposure*, sorti le 29 octobre 2024 après la phase exploratoire de ce travail.

décide d'utiliser ce pouvoir jusqu'alors réprimé pour découvrir la vérité.

Tous ces jeux ont en commun une structure narrative centrée autour d'un personnage principal (dirigé par le joueur ou la joueuse), parfois accompagné d'un ou une acolyte, et un accent mis sur la liberté laissée aux joueurs et joueuses qui leur permet d'avoir une influence sur le récit.

6.3 Récolte des données

La page Fandom [15] de la franchise *Life is Strange* contient les scripts des jeux précités, accessibles sous forme de site interactif dans lequel les choix de dialogues sont représentés par des boutons cliquables, qui révèlent les portions de dialogues entraînées par ces choix. La syntaxe des scripts est construite sur une forme de wikitexte [16], comme présenté à la figure 6.2.

```
252 ==== '''Dana's Room''' ====
253 ''Max enters Dana's room through the open door. Dana has the lights on and is sitting on her couch, looking distressed.''
254
255 '''<u>Optional Conversation with Dana</u>'''
256
257 '''Max:''' ''(sympathetically)'' Hey, Dana... How are you doing?
258
259 <tabber>{Saved Kate)=<h5 style="display:none">(Saved Kate)</h5><blockquote>'''Dana:''' ''(sadly)'' Better than Kate... I just
    can't believe she would even attempt suicide...
260
261 {{#tag:tabber|We're all responsible.=<h5 style="display:none">We're all responsible.</h5><blockquote>'''Max:''' I think we're
    all responsible for what happened...
262
263 '''Dana:''' True...but you're the only one who went up to that roof with Kate.
264
265 {{#tag:tabber|I was lucky.=<h5 style="display:none">I was lucky.</h5><blockquote>'''Max:''' I was lucky, that's all.
```

FIGURE 6.2 Extrait du script de *Life is Strange*.

[15] Site internet fondé en 2004 pour rassembler les fans de différentes licences de pop culture, à la manière d'un Wikipédia pour les univers fictionnels, https://life-is-strange.fandom.com/wiki/Life_is_Strange_Wiki, consulté le 13.08.2024.

[16] Langage de balisage dit « léger », qui propose une syntaxe facilement accessible et est notamment utilisée pour écrire les pages de Wikipédia.

Une fois les 19 scripts d'épisodes récoltés et réunis, ils ont été organisés en données structurées, avec deux objectifs en tête :

1. construire de manière automatisée les flux narratifs correspondant à chaque épisode,
2. parcourir automatiquement ces flux narratifs pour générer des réseaux de personnages.

Il est donc important de recenser à la fois les éléments du texte qui vont servir à construire les flux narratifs et ceux qui seront utiles à l'élaboration des réseaux de personnages (en l'occurrence, les noms des personnages et le découpage en dialogues).

De manière générale, la difficulté d'une construction automatisée de flux narratifs tient dans la compréhension de la structure narrative, et donc dans l'interprétation (informatisée) de la syntaxe utilisée pour expliciter cette structure. Dans cet exemple comme pour n'importe quelle source textuelle, la première grande étape est de parcourir attentivement les textes et d'identifier les marqueurs de l'arborescence des dialogues (souvent sous forme de balises ou de symboles typographiques qui indiquent un début d'option, un embranchement ou une fin de conversation). Si la transcription est de bonne qualité, la syntaxe choisie sera stable sur l'ensemble du texte, et l'on peut alors élaborer un code qui s'appuie sur ces marqueurs pour construire progressivement les nœuds et les arêtes du flux narratif associé. À noter que dans ce cas, une part conséquente des nœuds servira à indiquer ces marqueurs (et ainsi à faciliter la création des arêtes, comme expliqué à la section 6.4 plus bas), ce qui augmente artificiellement la taille globale du flux et donc le temps nécessaire à le parcourir pour générer ensuite les réseaux de personnages (raison pour laquelle une simplification est ensuite proposée à la fin de la même section).

Dans le cas présent, un examen détaillé des scripts nous permet d'identifier 22 marqueurs nécessaires à la construction du

flux (appelés « labels » et expliqués en détail ci-dessous), pour lesquels il faut renseigner 6 valeurs. Le tout donne lieu à des tableaux nettoyés et mis à disposition sur le répertoire character_network, et dont un extrait est fourni dans le tableau 6.1 pour plus de clarté. Notons que dans un esprit de synthèse, le mot « choix » désigne dorénavant l'ensemble des options et sous-options de dialogue qui fait basculer le script d'un scénario linéaire à un scénario pluriel, et le mot « option » désigne les différentes articulations de ce choix.

TABLE 6.1 Extrait du tableau regroupant les informations tirées des scripts de *Life is Strange*.

label	text	start_pos	end_pos	block	episode	conv
subchapter	Main Campus	35383	35407	30	1	107
choices_start	<tabber>	35408	35416	30	1	108
firstchoice_action	(Reported Nathan)	35409	35435	30	1	109
name	Principal Wells	35495	35517	30	1	109
conv_end	</blockquote>	35608	35621	30	1	110
choices_end	</tabber>	35623	35632	30	1	111
choices_start	<tabber>	35634	35642	30	1	112
firstchoice_action	(Max didn't look at missing posters)	35635	35680	30	1	113
name	Max	35859	35869	30	1	113
conv_end	</blockquote>	35974	35987	30	1	114
choices_end	</tabber>	35989	35998	30	1	115
conversation_opt	Optional Conversation with [[Michelle Grant|Ms. Grant]]	36000	36069	31	1	116
name	Max	36071	36081	31	1	116
name	Ms. Grant	36098	36114	31	1	116

Les 6 valeurs associées aux 22 labels sont les suivantes :

— `text` : le titre de la ligne (nom du personnage, nom du lieu ou du sous-chapitre, texte de l'option, ou encore texte associé aux balises),
— `start_pos` : la position du premier caractère de l'information capturée dans le script initial, afin de trier ensuite toutes les lignes du tableau par ordre chronologique,
— `end_pos` : la position du dernier caractère de l'information capturée dans le script initial,

— `episode` : le numéro de l'épisode dans le jeu (de 1 à 5 pour les trois premiers jeux, de 1 à 4 pour *Life is Strange : True Colors*),
— `block` : le numéro de la scène, pour délimiter les contextes utiles au calcul des cooccurrences dans les réseaux de personnages finaux. Ce numéro s'incrémente à chaque changement de lieu, de sous-chapitre ou de conversation,
— `conv` : le numéro du fragment de conversation, pour délimiter les unités narratives dans la construction du flux narratif. Ce numéro s'incrémente à chaque changement de structure du script (à chaque nouvel emplacement, nouvelle balise ou nouvelle option).

Les lignes du tableau contiennent les informations récupérées dans le script et organisées en 22 labels, qui peuvent être rassemblés en quatre grandes catégories : 5 labels d'emplacement, 4 labels de balises, 12 labels d'options et 1 label de nom.

Emplacement

La catégorie de l'emplacement regroupe les informations qui concernent le chapitre, le lieu et le type de conversation. Cinq labels lui sont associés.

— `subchapter` : sous-chapitre dans lequel se déroule la suite de l'histoire, lorsqu'il est mentionné,
— `location` : lieu dans lequel se déroule la scène, lorsqu'il est mentionné,
— `conversation` : conversation actuelle [17],
— `conversation_opt` : conversation optionnelle, tel qu'indiqué parfois dans le nom de la conversation pour signifier qu'on peut passer à côté de cet échange sans être coincé dans la progression du jeu,

[17] Le fait de déclencher un dialogue dans le jeu est indiqué dans le script avec un titre comme « Conversation avec Warren ».

— `side_conversation` : conversation qui n'inclut pas le personnage contrôlé par le joueur ou la joueuse, qui n'est alors qu'un témoin passif.

Balises

Pour comprendre la structure en arborescence des différents choix, il faut identifier les marqueurs qui génèrent les boutons correspondants sur la page du script. La catégorie des balises regroupe ainsi quatre labels.

— `choices_start` : début d'un choix, indiqué par la balise `<tabber>`. L'introduction de chaque option dans ce choix se fait de différentes manières, toutes décrites dans la catégorie suivante,
— `option_end` : fin d'une option démarrée grâce à un des labels d'options,
— `choices_end` : fin d'un choix, et retour à un scénario linéaire,
— `conv_end` : fin d'une conversation ou d'une réplique. C'est une balise qui se révèle utile pour segmenter correctement les portions de dialogues, et pour construire le flux narratif de manière automatisée (comme présenté à la section 6.4).

Options

La syntaxe de génération des options est la suivante : une balise principale pour le début d'un choix, une syntaxe particulière pour les différentes options d'un choix de premier niveau (syntaxe dans laquelle la première option est exprimée d'une certaine manière, et toutes les autres options sont identiquement présentées), et une autre syntaxe pour les options imbriquées dans une option de premier niveau (à noter qu'il n'y a pas de différence syntaxique entre une option de deuxième ou de troisième niveau). Pour tenir compte de ces différences, les labels d'options ont été pensés pour refléter d'une part la position de l'option (selon qu'il

s'agit ou non de la première), et d'autre part son niveau (premier niveau ou non).

— `firstchoice` : début d'une première option de premier niveau,
— `alternchoice` : début d'une autre option de premier niveau,
— `hiddenfirstchoice` : début d'une première option de niveau 2 ou supérieur,
— `alternhiddenchoice` : début d'une autre option de niveau 2 ou supérieur.

Il existe en outre deux autres types d'options rencontrés dans les différents scripts, qui sont traduits par des suffixes ajoutés aux quatre labels précédemment cités :

— les options importantes, écrites en lettres capitales et qui ont un effet majeur sur le scénario (suffixées `_big`),
— les options qui découlent d'actions précédemment réalisées dans la partie, qui ne sont donc pas présentées directement aux joueurs et joueuses et qui sont indiquées entre parenthèses (suffixées `_action`).

La combinaison des quatre labels et des deux suffixes donne ainsi lieu à 8 labels supplémentaires distincts pour décrire les options de dialogues.

Nom

Enfin, le label `name` récupère le nom des personnages qui prennent la parole tout au long du script. Il serait également possible de garder une trace de leurs répliques, mais cette donnée n'étant pas exploitée pour construire les réseaux de personnages dans cette étude de cas, elle ne figure pas dans les tableaux. Toutefois, chaque ligne du tableau (qui correspond toujours à l'un des 22 labels) est localisée dans le texte initial, ce qui permet de reconstruire l'information a posteriori si nécessaire.

Ces données ainsi structurées permettent d'identifier chaque occurrence d'un personnage, chaque embranchement d'un choix

et chaque changement de scène, autant d'informations nécessaires à l'élaboration du flux narratif, puis des réseaux de personnages associés à ce flux.

6.4 Construction du flux

La construction automatisée du flux narratif se fait en deux étapes : d'abord la création des nœuds qui représentent les unités narratives, puis, plus délicate, la création des arêtes (c'est-à-dire des transitions). Pour éviter toute confusion avec le vocabulaire des réseaux de personnages (qui seront ensuite construits en parcourant les flux), je parlerai dans le reste de ce chapitre de nœuds et d'arêtes du réseau, ainsi que d'unités narratives et de transitions du flux.

La figure 6.3 présente un extrait de flux narratif théorique de l'épisode 5 de *Life is Strange*, pour présenter les spécificités de cette franchise. Pour les besoins de clarté de l'illustration, les unités narratives représentent ici soit des conversations complètes sans embranchement (sous l'intitulé « Conv »), soit les répliques correspondant aux choix laissés aux joueurs et joueuses (par exemple, on commence par choisir entre « I started the storm » et « He killed C »). Il faut donc bien avoir à l'esprit que les flux construits automatiquement dans cette section seront bien plus complexes et enrichis de nombreuses unités narratives purement structurelles.

La source (START) et le puits (END) ont été ajoutés au script d'origine pour répondre à la définition du flux narratif. Les autres unités narratives en majuscule, à la droite de la figure (KISS, HUG et LEAVE) correspondent à des options importantes (comme décrit plus haut), qui peuvent avoir un impact plus large sur la suite du récit. Le jeu propose, à la fin de chaque épisode, un récapitulatif qui situe notre expérience par rapport aux choix globaux de la communauté pour ces options importantes : on peut ainsi voir que 67 % des joueurs et joueuses ont choisi « KISS », alors

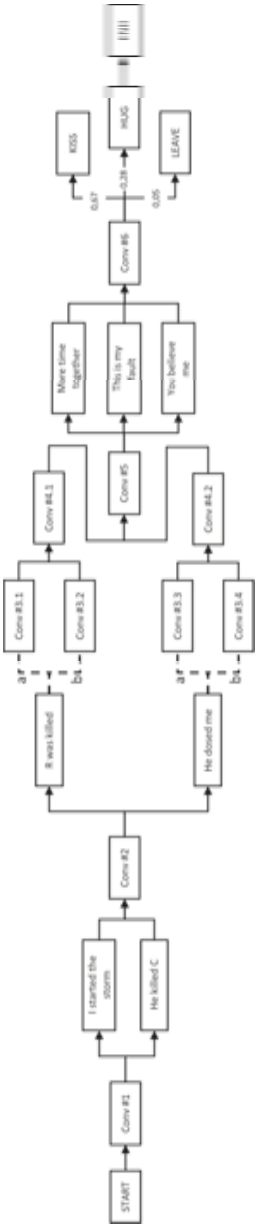


FIGURE 6.3 Flux narratif tiré d'une scène de l'épisode 5 de *Life is Strange*.

que seulement 5 % ont pris l'option « LEAVE ». Ces statistiques ont été ajoutées sur les arêtes de la figure 6.3 pour illustrer les probabilités qu'une traversée aléatoire (et représentative des parties déjà jouées) passe par l'une ou l'autre de ces options importantes (par défaut et en l'absence de données, on considère que toutes les autres options sont équiprobables). Toutefois, dans le flux narratif créé automatiquement pour l'exemple pratique, ces probabilités ne sont pas prises en compte pour des contraintes de temps et l'on considère que tous les choix sont équiprobables.

Les transitions en pointillés qu'on peut observer au centre de la figure, qui sortent des unités narratives « R was killed » et « He dosed me », indiquent des choix implicites, qui sont en réalité la résultante de choix faits précédemment. En prenant l'option de dialogue « R was killed », si l'on avait décidé précédemment de faire l'action « a », alors on va passer à la conversation 3.1 sans avoir conscience d'avoir pris un nouvel embranchement. On peut voir que dans cet extrait de flux précis, les choix implicites n'ont pas un impact majeur sur la progression de l'intrigue, puisqu'on débouche dans tous les cas sur la conversation 4.1. Il aurait tout de même été souhaitable de trouver une solution pour implémenter automatiquement ces choix implicites, et empêcher la traversée de passer par la transition « b » si l'on n'a pas pris l'option « b » dans le choix d'origine. Malheureusement, la syntaxe des scripts rend la tâche extrêmement fastidieuse : ces choix sont simplement indiqués par un énoncé entre parenthèses, disant par exemple « (Max a dénoncé Nathan) » ou « (Max a caché la vérité) » . Leur présence peut donc être détectée automatiquement (grâce aux parenthèses), mais le renvoi au choix d'origine manque de précision : en l'absence de référence structurelle (par exemple, à l'aide d'un identifiant unique), il faudrait pouvoir interpréter le contenu des parenthèses et remonter dans le jeu jusqu'à trouver un embranchement qui semble proposer un choix correspondant afin de garder en mémoire l'option sélectionnée pour ledit choix. Cette tâche nécessite un modèle poussé de com-

préhension de la langue, qui est en outre rendue impossible par le fait que certains de ces choix implicites reposent sur des actions de jeu, et non sur des options de dialogue (par exemple, Max se dirige d'elle-même vers Daniel/il engage la conversation alors qu'elle s'éloigne). Dans le périmètre de ce projet, il a donc été décidé de ne pas traiter ces choix différemment des autres, créant ainsi des traversées en réalité impossibles qui seront thématisées dans les résultats, à la section 6.5.

Création des unités narratives

Comme indiqué à la section 6.3, pour faciliter la logique de définition des transitions, il a paru judicieux de rendre les différentes articulations de la structure narrative les plus explicites possibles lors de la création des unités narratives. C'est la raison pour laquelle une part conséquente des unités narratives de ce flux ne contiennent pas de personnages et d'interactions, mais servent de marqueurs pour indiquer le début d'un choix, la fin d'une conversation ou un changement de lieu.

Ainsi, chaque valeur de la colonne `conv` représente une unité narrative, chargée de plusieurs attributs utiles pour la situer (chapitre, sous-chapitre, lieu, numéro de la scène) et l'identifier (titre et label associé), et évidemment pour garder une trace des personnages rencontrés dans cette unité. Cette dernière information est organisée en dictionnaire (au sens informatique du terme), dans lequel les clés sont les noms des personnages cités dans la portion de script correspondante et les valeurs sont le nombre de répliques associées aux personnages dans cette même portion. Il est important de récolter ce nombre de répliques si l'on souhaite construire par la suite un réseau de personnages dont le poids des nœuds est la somme de leurs apparitions dans l'œuvre, ainsi que Charnetto le permet.

Toutes ces unités narratives sont ensuite rassemblées dans une liste, dont voici un extrait :

```
[{'id': 1,
  'title': '',
  'location': '',
  'chapter': '4',
  'subchapter': "Chloe's Room",
  'block': 1,
  'type': '',
  'characters': {'Chloe': 4, 'Max': 2}},
 {'id': 2,
  'title': '<tabber>',
  'location': '',
  'chapter': '4',
  'subchapter': "Chloe's Room",
  'block': 1,
  'type': 'choices_start',
  'characters': {}},
 {'id': 3,
  'title': "You're insane.",
  'location': '',
  'chapter': '4',
  'subchapter': "Chloe's Room",
  'block': 1,


  'type': 'firstchoice',
  'characters': {'Max': 2, 'Chloe': 1}},
 {'id': 4,
  'title': 'end_node',
  'location': '',
  'chapter': '4',
  'subchapter': "Chloe's Room",
  'block': 1,
  'type': 'conv_end',
  'characters': {}},
  ...]
```

Comme on peut le voir ici, les unités narratives 2 et 4 sont des marqueurs de structure (respectivement le début d'un choix et la fin d'un segment de conversation), alors que les unités narratives 1 et 3 représentent des dialogues : un échange direct faisant intervenir 4 fois Chloe et 2 fois Max, et la première option d'un choix, « You're insane » , débouchant sur un dialogue entre Chloe et Max.

Création des transitions

Il s'agit à présent d'articuler cette liste d'unités narratives en un graphe orienté, avec un seul point de départ (la source) et un seul point d'arrivée (le puits). Le code intégral est à trouver sur le répertoire character_network et repose sur l'idée de parcourir une à une toutes les unités de la liste, en créant des transitions vers l'arrière (c'est-à-dire en considérant l'unité observée comme l'arrivée d'une transition). Pour accompagner ces explications, la figure 6.4 représente un exemple fictif de flux narratif généré automatiquement avec les différents types d'unités narratives présentés ci-dessous.

La complexité vient principalement de la présence de choix imbriqués dans d'autres choix, qui nécessitent une bonne vision d'ensemble sur l'emplacement de l'unité narrative actuelle dans le flux pour comprendre quelle partie des options est encore active, quels choix ont été intégralement assemblés, etc. On ouvre donc au préalable trois marqueurs d'observation, qui devraient être vides à chaque fois qu'un choix de premier niveau est construit entièrement :

— `sources` : l'ensemble des unités narratives qui débouchent sur une ou plusieurs options qui n'ont pas encore été « refermées » (en vert dans la figure 6.4),
— `convs` : un dictionnaire qui réunit les fragments de conversation actuellement ouverts, associés à leur source (en bleu dans la figure 6.4),

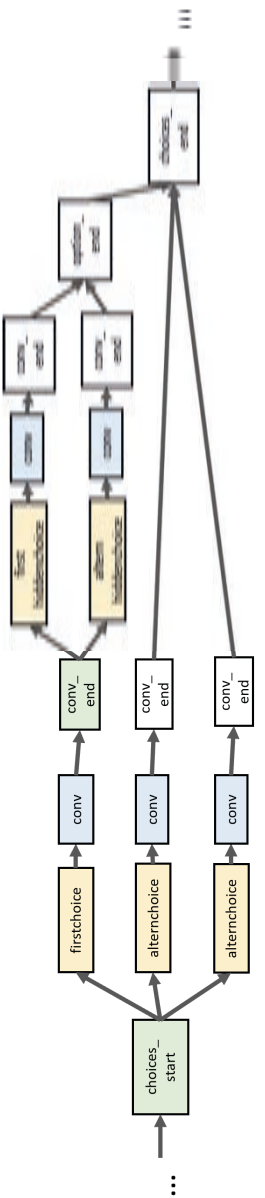


FIGURE 6.4 Exemple de flux narratif généré automatiquement : les unités de type `sources` sont en vert, les `choices` en jaune et les `convs` en bleu.

— `choices` : un dictionnaire qui réunit les options actuellement ouvertes, associées à leur source (en jaune dans la figure 6.4).

La stratégie de création des transitions dépend de chaque type d'unité narrative, selon la logique suivante (voir la figure 6.4) :

— `choices_start` : on crée une transition partant de l'unité précédente pour arriver sur l'unité actuelle,
— `firstchoice` : on ajoute l'unité précédente à `sources` et on crée une transition partant de la source la plus récente et arrivant sur l'unité actuelle,
— `alternchoice` : on crée une transition partant de cette source et arrivant sur l'unité actuelle,
— `conv` : on ajoute l'unité actuelle à `convs` en y associant la source la plus récente et on crée la transition qui part de l'unité précédente pour arriver sur l'unité actuelle,
— `conv_end` : on retire la dernière conversation de `convs`, on ajoute l'unité à `choices` (pour garder une trace du fait qu'il faudra refermer ce chemin plus tard) et on crée une transition partant de l'unité précédente vers l'unité actuelle,
— `option_end` : ce type d'unité narrative est le plus délicat à traiter, car il déclenche la fermeture de toutes les options encore ouvertes pour la source concernée. On retire d'abord de `convs` les unités rattachées à la source la plus récente et on les ajoute à `choices`. Ensuite, pour toutes les unités dans `choices` rattachées à cette même source, on crée une transition entre ces unités et l'unité actuelle. Enfin, on retire ces unités de `choices` et on retire la source de `sources` puisque ce choix est désormais clos. Dans le cas particulier où il n'y a qu'une seule option (ce qui se produit parfois pour des choix passifs, pour lesquels soit on a fait une action précise précédemment qui découle sur une nouvelle portion de scénario, soit on ne l'a pas faite et rien ne se passe), il faut créer une autre arête de la source la plus récente à l'unité

actuelle pour construire cette deuxième option qui illustre le fait que rien ne se passe. La dernière étape est de rajouter l'unité actuelle à `choices` pour pouvoir la relier ultérieurement à l'unité `choices_end` qui marquera la fin de toute l'arborescence de ce choix,

— `choices_end` : dans le même esprit que `option_end` mais à une échelle plus large, cette unité déclenche la fermeture de toute l'arborescence du choix actuel. On déplace tout le contenu de `convs` dans `choices`. À ce stade, il n'y a plus qu'une unité narrative dans `sources` et l'on crée donc une transition partant de chaque unité de `choices` vers l'unité actuelle, ainsi que la deuxième transition de la source vers l'unité actuelle si l'on est dans le cas d'un choix à une seule option, comme expliqué dans `option_end`. Enfin, on vide `convs` pour repartir de marqueurs vierges pour la suite du script.

Si le script du jeu est correctement écrit de bout en bout, on se retrouve donc avec un flux narratif comportant un seul point de départ et un seul point d'arrivée, auquel on peut ajouter des attributs de probabilité sur chaque transition, selon des données externes ou en calculant des probabilités équiprobables pour chaque transition sortante d'une unité narrative. Dans les faits, il a fallu corriger quelques points de syntaxe sur les différents épisodes pour construire un graphe cohérent [18].

Simplification du flux

Comme expliqué plus haut, la stratégie d'automatisation de la création de ce flux narratif passe malheureusement par une multiplication des unités narratives qui certes, facilitent son implémentation, mais sont pour certaines superflues et vides de sens

[18] Toutes les données nettoyées sont accessibles sur le répertoire character_network pour de futures études.

quant aux analyses narratives qu'on cherche à mettre en place. Le résultat est un flux narratif de plusieurs centaines d'unités et de transitions, avec un nombre total de traversées bien trop important pour espérer les générer toutes dans un temps raisonnable (ce qui n'est pas grave si l'on souhaite en extraire un réseau de personnages moyen avec moyenne et variance calculables théoriquement, comme exposé à la section 5.3, mais empêche d'autres types de résultats comme le minimum ou le maximum pour chaque poids de nœud et d'arête).

Pour adoucir cet effet d'explosion du nombre de traversées, une étape de simplification du flux narratif a été ajoutée. L'idée est que si l'on isole un choix (c'est-à-dire un groupe allant de `choices_start` à `choices_end`, comme dans la figure 6.4), qu'on en parcourt toutes les possibilités de traversées et qu'on se retrouve systématiquement avec les mêmes personnages et le même nombre d'interventions à la sortie, on peut considérer que tous les chemins de ce choix sont *équivalents*. Il est alors possible de remplacer l'intégralité de ce groupe d'unités et de transitions par une seule unité narrative : si elle réunit toutes les informations récoltées le long d'un de ces chemins équivalents, passer par cette unité reviendra à avoir traversé le choix initial par l'un ou l'autre des chemins possibles, sans perdre d'éléments importants pour les analyses ultérieures.

Par exemple, si l'on souhaite recenser les personnages rencontrés dans chaque dialogue et leur nombre de répliques, et que tous les chemins contenus dans le choix présenté à la figure 6.4 font intervenir exactement Max 3 fois et Chloe 2 fois, ces chemins sont considérés équivalents et on peut remplacer tout le choix par une seule unité narrative contenant 3 mentions de Max et 2 mentions de Chloe, en simplifiant sensiblement le flux et en préservant toutes les informations désirées.

Cette simplification a des effets non négligeables sur la taille totale des flux narratifs. À titre d'exemple, l'épisode 1 de *Life is Strange* comporte à la base 856 unités narratives et 74 choix. En

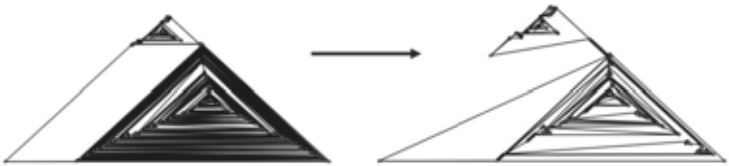


FIGURE 6.5 Simplification du flux narratif, en réunissant les chemins équivalents (à gauche, le flux à 856 unités narratives, à droite le même flux avec seulement 529 unités narratives). La visualisation en triangle est simplement une manière d'éviter les chevauchements des arêtes pour une meilleure lisibilité de l'image.

parcourant chacun de ces choix et en simplifiant ceux dont les chemins sont équivalents, on se retrouve avec un réseau de 529 unités narratives, comme l'illustre la figure 6.5. Un algorithme de comptage de tous les chemins possibles d'un flux [19] nous permet ainsi de constater que cette étape réduit le nombre de traversées possibles de 10^{63} à 10^{28} sur cet épisode.

6.5 Résultats et analyses

Pour garder l'exemple du premier *Life is Strange*, qui est découpé en cinq épisodes, on se retrouve ainsi avec cinq flux narratifs simplifiés qu'il faudrait parcourir successivement pour générer une traversée complète du jeu. Pour des raisons de temps de calcul, j'ai choisi de procéder de la manière suivante :

— générer 100 000 traversées aléatoires de chaque épisode du jeu et les réseaux de personnages associés, de sorte à estimer les valeurs minimales et maximales en plus des moyennes et écarts-types qui peuvent être calculés théoriquement (comme exposé à la section 5.3),

[19] L'implémentation de cet algorithme est présente sur le répertoire character_network.

— calculer le réseau moyen de chaque épisode sur la base des 100 000 réseaux (avec pour chaque arête la moyenne, l'écart-type, le minimum et le maximum),
— en déduire le réseau moyen de l'ensemble du jeu.

Pour la dernière étape, on s'appuie sur les propriétés mathématiques de la variance et de l'écart-type sur des variables indépendantes (voir section 5.3). Le poids moyen de chaque arête (et de chaque nœud) est la somme des poids moyens de cette même arête (et de ce même nœud) à travers les cinq épisodes, et la variance totale est également la somme des variances pour chaque épisode (en travaillant avec les écarts-types, il suffit donc de les mettre au carré pour obtenir les variances, de les sommer, puis de refaire une racine carrée qui nous donne l'écart-type final). Le maximum et le minimum se calculent de la même manière, en sommant les valeurs récoltées.

Nous avons maintenant, pour chacun des quatre jeux de la franchise, un réseau par épisode, un réseau global et une série de mesures réunies dans des tableaux[20]. Ces éléments rassemblés me permettent une série d'observations, destinées à ouvrir des pistes de recherche pour les spécialistes du domaine et organisées en différentes catégories dans les sections suivantes. En préambule de ces analyses, la sous-section 6.5.1 contrôle la qualité des réseaux obtenus à l'aide des notions théoriques présentées à la section 5.3.

6.5.1 Vérification des valeurs empiriques

Comme introduit à la section 5.3, il est possible de vérifier que les 100 000 réseaux générés aléatoirement produisent un réseau moyen représentatif en comparant les poids moyens empiriques (ainsi que leur écart-type) avec les valeurs théoriques attendues. Ce contrôle, intégré au code du répertoire character_network, a

[20] Ces tableaux sont accessibles sur le répertoire character_network.

été effectué pour toutes les valeurs de moyenne et d'écart-type des nœuds et arêtes des différents réseaux moyens. À des fins d'exemple, les nœuds du réseau moyen relatif au premier épisode de LIS1 sont reportés dans le tableau 6.2 : on peut constater la grande proximité entre les valeurs attendues (en gras) et les valeurs obtenues. Cette proximité se retrouve de manière similaire dans les autres réseaux moyens ainsi générés.

TABLE 6.2 Comparaison des valeurs théoriques (en gras) et empiriques pour les nœuds du réseau moyen du premier épisode de LIS1, sur la base des 100 000 traversées aléatoires du flux narratif.

	poids moyen	moyenne théorique	écart-type	écart-type théorique
Max	266,66821	**266,667**	5,37	**5,375**
Mr. Jefferson	43,0	**43,0**	0,0	**0,0**
Victoria	27,00139	**27,0**	1,73	**1,732**
Kate	13,49992	**13,5**	0,5	**0,5**
Juliet	19,66302	**19,667**	1,174	**1,179**
Zachary	8,0	**8,0**	0,0	**0,0**
Courtney	7,0	**7,0**	0,0	**0,0**
Taylor	7,0	**7,0**	0,0	**0,0**
Nathan	27,0	**27,0**	0,0	**0,0**
Chloe	126,97486	**126,938**	4,783	**4,78**
Justin	11,99918	**12,0**	0,815	**0,816**
Dana	14,41645	**14,417**	0,996	**0,997**
Logan	5,0	**5,0**	0,0	**0,0**
Friend	3,0	**3,0**	0,0	**0,0**
Daniel	5,6274	**5,625**	0,698	**0,696**
Brooke	4,0	**4,0**	0,0	**0,0**
Samuel	6,0	**6,0**	0,0	**0,0**
David	25,78229	**25,75**	5,097	**5,101**
Principal Wells	7,99579	**8,0**	1,222	**1,225**
Ms. Grant	4,50775	**4,5**	2,5	**2,5**
Hayden	4,24954	**4,25**	1,091	**1,09**
Stella	5,66466	**5,667**	0,472	**0,471**
Evan	5,24759	**5,25**	0,432	**0,433**
Luke	2,86905	**2,875**	1,964	**1,965**
Alyssa	8,0	**8,0**	0,0	**0,0**
Warren	21,50288	**21,5**	1,5	**1,5**

On peut également se demander s'il est possible d'obtenir des résultats tout aussi satisfaisants en produisant moins de traversées, pour limiter les coûts de production de ces réseaux moyens. Pour esquisser une réponse à cette question, d'autres réseaux

moyens ont été générés, toujours sur le premier épisode de LIS1, avec 5, 10, 100, 1000, et 10 000 traversées. Les deux graphes de la figure 6.6 représentent l'évolution de la moyenne et de l'écart-type du poids du nœud « Max » (ayant à la fois le plus grand poids moyen et le plus grand écart-type sur cet épisode) : comme on peut le constater, un millier de traversées aurait suffi à produire des résultats très proches des valeurs théoriques attendues dans ce cas précis. Dans le cadre de ce projet, la question du nombre optimal de traversées nécessaires n'a pas fait l'objet d'études plus poussées (le matériel à disposition permettant aisément de produire les 100 000 réseaux), mais de telles études pourraient aider

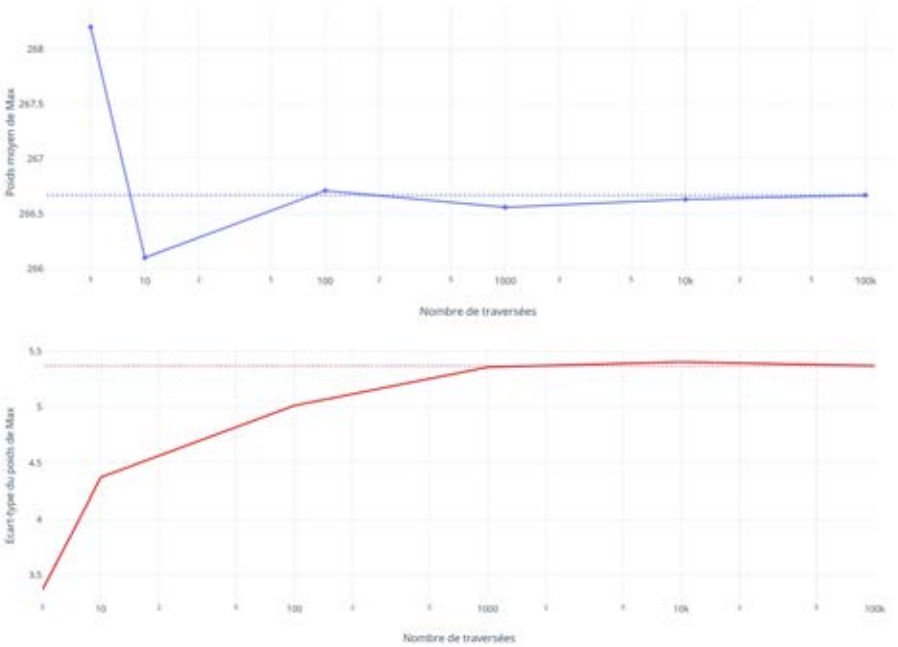


FIGURE 6.6 Variation de la moyenne et de l'écart-type du poids du nœud « Max » selon le nombre de traversées du flux narratif du premier épisode de LIS1. Les droites représentent les valeurs théoriques attendues.

à établir un ordre de grandeur lié à la taille du flux narratif et à la variation attendue, afin de calibrer plus finement l'effort à fournir pour produire des résultats satisfaisants.

6.5.2 Poids moyen des nœuds et des arêtes

Commençons par nous intéresser à la répartition des poids des nœuds au sein des différents jeux. Chaque jeu présente un personnage principal qui est contrôlé par le joueur ou la joueuse et qui participe à la grande majorité des échanges avec les autres personnages : Max pour *Life is Strange* (LIS1), Sean pour *Life is Strange 2* (LIS2), Chloe pour *Life is Strange : Before the Storm* (LISBEF) et Alex pour *Life is Strange : True Colors* (LISTC). Sans surprise, on peut ainsi constater que dans les quatre jeux, c'est ce personnage qui obtient le poids moyen le plus haut (voir figure 6.7), le poids étant défini ici comme le nombre de répliques associées à ce personnage au cours de la partie. On verra dans la sous-section 6.5.4 que ce phénomène se répète avec les autres mesures, qui ancrent fermement ce personnage comme indispensable en matière de centralité dans le récit.

En ce qui concerne la répartition des répliques des personnages, voici la proportion de personnages dont le poids est supérieur à 1 :

— LIS1 : 54/68, soit 79,41 %
— LIS2 : 60/63, soit 95,23 %
— LISBEF : 38/44, soit 68,36 %
— LISTC : 58/90, soit 64,44 %

Largement plus de la moitié des personnages conçus dans ces jeux ont plus d'une réplique. On peut remarquer que si LISTC est le jeu comportant le plus de personnages, c'est également celui qui présente la proportion la plus basse de personnages ayant plusieurs répliques.

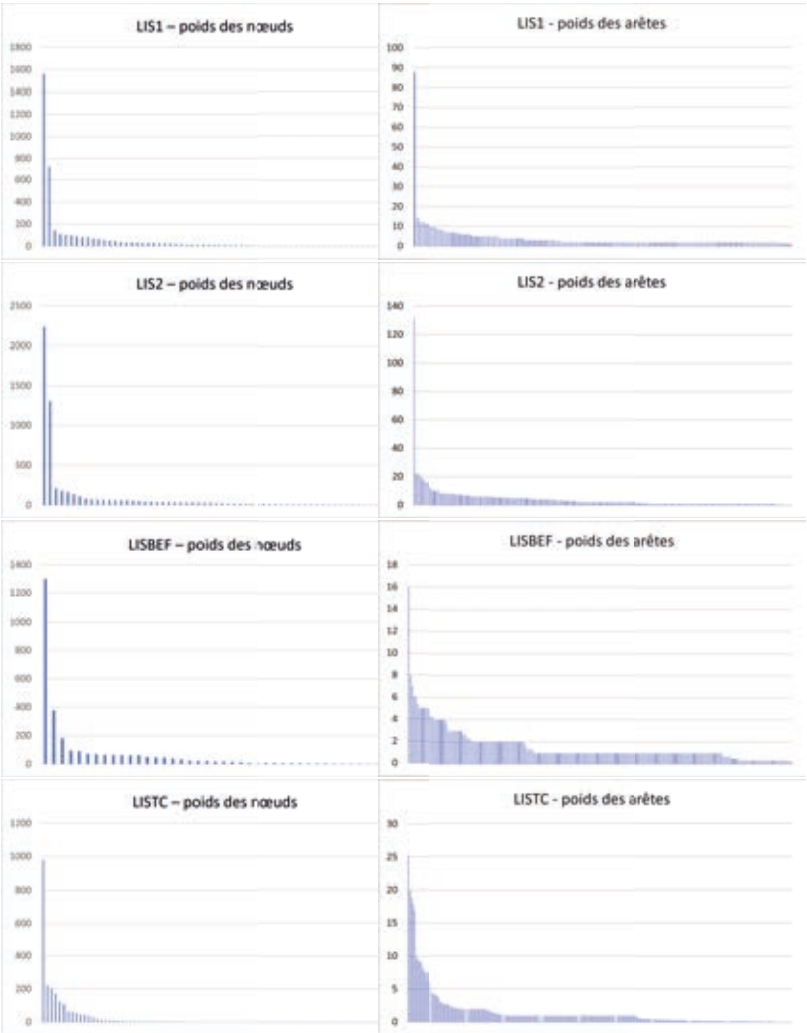


FIGURE 6.7 Distribution des poids des personnages (colonne de gauche) et des interactions (colonne de droite) pour les quatre jeux. Les moyennes des poids sont calculées sur les 100 000 réseaux : lorsqu'un nœud ou une arête n'apparaît pas dans un réseau, on considère que son poids est de 0.

Quant à la répartition des poids des personnages, on peut constater avec la figure 6.7 un phénomène de décroissance exponentielle entre les personnages les plus fréquents et, bien plus nombreux, ceux qui n'apparaissent que quelques fois au cours du jeu.

Si l'on fait ce même travail sur les arêtes, voici la proportion d'arêtes dont le poids est supérieur à 1 :

— LIS1 : 133/474, soit 28,06 %
— LIS2 : 137/218, soit 62,84 %
— LISBEF : 69/211, soit 32,70 %
— LISTC : 65/250, soit 26,00 %

Pour LIS1, LISBEF et LISTC, on a donc entre un quart et un tiers des arêtes seulement qui représentent plus d'une interaction. Par contre, dans le cas de LIS2, on s'approche des deux tiers : à l'image des proportions des personnages, on constate que LIS2 est un jeu beaucoup plus fourni que les trois autres, en ce qui concerne l'importance de chaque personnage et la quantité d'interactions entre eux.

La distribution des interactions à la figure 6.7 nous montre un phénomène intéressant pour LISBEF : l'amplitude des valeurs de poids des arêtes est bien moins grande que dans les autres jeux (ce qui s'explique par la durée plus courte de cet opus) et, surtout, on a 38 arêtes (soit 18 %) dont le poids moyen est compris entre 0 et 1 et qui représentent donc des interactions qui peuvent exister ou disparaître complètement selon les choix faits par le joueur ou la joueuse.

Ce dernier point mérite d'être creusé. Si l'on regarde de plus près les nœuds et arêtes qui ont parfois un poids de 0 à travers les épisodes des quatre jeux, on obtient les résultats suivants (voir le tableau 6.3) :

— Dans LIS1, tous les personnages apparaissent dans les 100 000 réseaux générés automatiquement. Il en est de même pour la grande majorité des arêtes : celles qui ont une

TABLE 6.3 Présence des nœuds et arêtes à travers les différents réseaux des quatre jeux. Une présence de 100 % signifie que tous les objets de ce type sont représentés dans l'ensemble des 100 000 réseaux générés automatiquement.

	épisode	nœuds	arêtes
LIS1	1	**100 %**	**100 %**
	2	**100 %**	97,28 %
	3	**100 %**	95,92 %
	4	**100 %**	95,45 %
	5	**100 %**	**100 %**
LIS2	1	**100 %**	83,72 %
	2	85,71 %	94,28 %
	3	**100 %**	**100 %**
	4	90 %	93,33 %
	5	**100 %**	93,22 %
LISBEF	1	**100 %**	89,09 %
	2	96,77 %	88,81 %
	3	81,48 %	68,04 %
	4	**100 %**	**100 %**
LISTC	1	69,23 %	50,72 %
	2	41,.6 %	58,22 %
	3	48 %	50 %
	4	43,75 %	43,18 %
	5	96 %	84,76 %

présence partielle ont un poids de 1 et représentent donc des interactions très marginales. Le jeu semble donc proposer tous ses personnages et la majeure partie de ses interactions indépendamment des choix laissés aux joueurs et joueuses : si liberté il y a, elle intervient donc surtout dans la nature des échanges, et non dans leur existence.

— LIS2 et LISBEF offrent un profil relativement similaire : quelques épisodes dans lesquels nœuds et arêtes sont toujours représentés, d'autres qui offrent une forme relative de

variété (allant jusqu'à presque un tiers des arêtes qui disparaissent selon les choix, à l'épisode 3 de LISBEF). Dans les deux jeux, les nœuds et arêtes qui disparaissent de certains réseaux concernent des personnages marginaux et des interactions uniques.

— LISTC obtient des scores bien plus marqués qui rejoignent les résultats précédents : la présence importante de nœuds et d'arêtes dont le poids moyen ne dépasse pas 1 est cohérente avec le fait que beaucoup de personnages et d'interactions ne sont rencontrés que dans certaines traversées du flux narratif. Le jeu paraît ainsi plus généreux en contenu optionnel, et si la grande majorité de ces objets à présence variable a un poids faible, on y trouve quelques personnages qui peuvent avoir entre 0 et 15 répliques, ce qui représente une amplitude bien plus nette que les jeux précédents.

— Dans les différents jeux, on retrouve un phénomène similaire au niveau des épisodes : les épisodes de milieu de jeu proposent davantage de contenu facultatif que ceux de début et de fin, qui proposent un démarrage plus linéaire et rassemblent les trames narratives pour offrir une fin cohérente.

Pour ce qui est de la variété des personnages rencontrés et des interactions qui peuvent exister ou non entre ces personnages, LISTC se démarque clairement des trois précédents opus en proposant une palette large de personnages dont la présence est optionnelle et d'interactions non systématiques. Malgré tout, aucun des jeux étudiés ne propose de personnage qui pourrait être soit très présent, soit totalement absent ; les efforts à fournir pour développer des jeux à embranchements multiples ont certainement une influence sur ce phénomène, puisqu'il semble moins coûteux d'offrir de la liberté dans la nature des échanges que dans le développement de trames narratives totalement différentes.

6.5.3 Écart-type absolu et relatif

Une des questions initiales au moment du choix de la franchise *Life is Strange* pour cette étude était l'effet concret des choix des joueurs et joueuses sur le cours du jeu, en matière d'interactions et d'importance des personnages. Pour mesurer cette variation, on peut s'intéresser à l'écart-type des poids des nœuds et des arêtes, c'est-à-dire à l'amplitude de l'écart entre les différentes valeurs et leur moyenne. La proportion des nœuds dont l'écart-type est plus grand que zéro (et dont le poids varie donc entre les différents réseaux possibles) est la suivante :

— LIS1 : 37/68, soit 54,41 %
— LIS2 : 52/63, soit 82,53 %
— LISBEF : 31/44, soit 70,45 %
— LISTC : 43/90, soit 47,77 %

Ces chiffres indiquent que la liberté laissée aux joueurs et joueuses est appréciable, puisque plus de la moitié (ou presque dans le cas de LISTC) des personnages voient une variation dans le nombre de leurs répliques en fonction des décisions prises. Comme dans le cas des poids, c'est LIS2 qui obtient les scores les plus élevés, avec plus de quatre personnages sur cinq présentant une variation de leur poids. La situation est toutefois bien différente si on regarde du côté des arêtes dont l'écart-type est plus grand que zéro :

— LIS1 : 16/474, soit 3,37 %
— LIS2 : 66/218, soit 30,27 %
— LISBEF : 7/211, soit 3,32 %
— LISTC : 27/250, soit 10,80 %

Ces valeurs plutôt basses peuvent surprendre : dans le cas de LIS1 et LISBEF, cela signifie que la quasi-totalité des interactions ont lieu peu importe les choix laissés aux joueurs et joueuses. En réalité, ce n'est pas si étonnant, puisque la plupart des décisions concernent des choix de dialogue au sein d'une même interaction, et on peut donc s'attendre à ce que la plupart des échanges

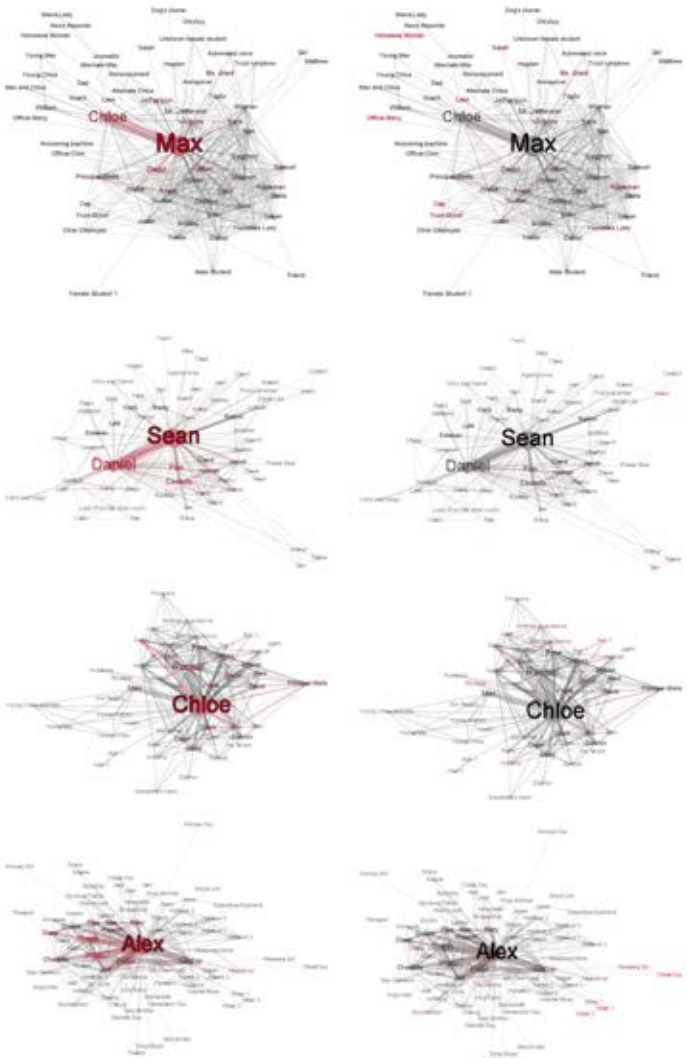


FIGURE 6.8 Réseaux moyens globaux pour, de haut en bas : LIS1, LIS2, LISBEF, LISTC ; à gauche avec écart-type global, à droite avec écart-type relatif (tous deux indiqués par la couleur, le rouge indiquant un écart-type élevé).

existent, même si leur nature et le nombre de répliques sont amenés à varier. Notons tout de même, comme dans les trois précédentes statistiques, le score bien supérieur de LIS2, dont presque le tiers des interactions présente une variation.

Outre la question de la *présence* de variations dans les poids des arêtes et des nœuds, on peut évidemment s'intéresser à leur importance et donc, aux valeurs de ces écarts-types. Dans la colonne de gauche de la figure 6.8, l'écart-type est représenté par la couleur des nœuds et des arêtes : plus ils sont rouges, plus l'écart-type est grand. On voit immédiatement que les nœuds et arêtes les plus importants sont très rouges, ce qui donne l'impression que les choix des joueurs et joueuses ont beaucoup d'effet sur la présence et les dialogues des personnages principaux. Toutefois, cette impression est à mettre en perspective avec les valeurs moyennes de ces objets, c'est pourquoi on regarde également l'écart-type *relatif* (aussi appelé coefficient de variation), défini comme l'écart-type divisé par la moyenne. Les nouveaux résultats, représentés dans la colonne de droite, donnent une impression tout autre : relativement aux poids moyens, les nœuds et arêtes principaux varient en réalité très peu, et les zones les plus rouges des différents graphes sont concentrées sur des personnages très marginaux, pouvant apparaître entre 0 et 1 fois par exemple, selon les décisions prises. Attention toutefois à ne pas se baser sur ces graphes pour conclure à une fausse promesse (et ainsi une absence de liberté) de la part des développeurs du jeu : je ne me concentre ici que sur l'existence d'échanges et de répliques, pas sur leur nature (comme développé plus bas dans la section 6.6).

6.5.4 Mesures et interprétations

On peut également s'intéresser aux différentes mesures sur les réseaux moyens, comme on le ferait pour un réseau de personnages unique. Le tableau 6.4 présente une série de mesures

TABLE 6.4 Mesures sur les réseaux moyens des quatre jeux.

	LIS1	**LIS2**	**LISBEF**	**LISTC**
# nœuds	62	60	40	77
diamètre	3	2	2	4
distance moyenne	2,5645	1,9833	1,9750	2,9870
densité	0,2501	0,1232	0,2679	0,0848
fermeture triadique	0,7498	0,3005	0,5801	0,2766
chemin moyen	1,8112	1,8768	1,7321	1,9925
taille max. clique	26	11	11	10
degré des nœuds	Max 57 Justin 31 Woman 31 Man 31 Trevor 31	Sean 59 Daniel 48 Jacob 15 Cassidy 13 Finn 12	Chloe 39 Rachel 22 Skip 20 James 19 Mr. Keaton 17	Alex 72 Jed 28 Ethan 27 Gabe 24 Ryan 18
force des nœuds	Max 316,90 Chloe 146,50 Victoria 68,50 Alyssa 55,0 Warren 55,0	Sean 423,93 Daniel 324,28 Cassidy 109,87 Finn 88,34 Hannah 76,80	Chloe 130,74 Rachel 58,25 James 31,33 Frank 29,0 Rose 28,33	Alex 200,80 Jed 60,30 Ryan 59,42 Steph 54,23 Gabe 48,64
poids des arêtes	Max - Chloe 88,00 Max - Victoria 14,25 Max - Kate 12,33 Max - Alyssa 12,0 Max - David 11,29	Daniel - Sean 131,64 Sean - Jacob 22,58 Sean - Cassidy 21,83 Sean - Karen 21,0 Daniel - Cassidy 19,19	Chloe - Rachel 16,0 Chloe - Max 8,0 Chloe - James 7,0 Chloe - Frank 6,0 Chloe - Steph 6,0	Alex - Ethan 25,20 Alex - Ryan 19,95 Alex - Steph 18,93 Alex - Jed 17,86 Alex - Gabe 16,94
centralité de proximité	Max 0,9385 Justin 0,6703 Woman 0,6703 Man 0,6703 Trevor 0,6703	Sean 1,0 Daniel 0,8429 Jacob 0,5728 Cassidy 0,5620 Finn 0,5566	Chloe 1,0 Rachel 0,6964 Skip 0,6724 James 0,6610 Mr. Keaton 0,6393	Alex 0,95 Jed 0,6129 Ethan 0,5984 Gabe 0,5846 Ryan 0,5672
centralité intermédiaire	Max 0,4832 Chloe 0,0615 Woman 0,0429 Justin 0,04147 Man 0,0362	Sean 0,5984 Daniel 0,2768 Jacob 0,0104 Karen 0,0035 Cassidy 0,0033	Chloe 0,5570 Rachel 0,0337 Skip 0,0251 James 0,0204 Mr. Keaton 0,0120	Alex 0,7604 Ethan 0,0824 Jed 0,0439 Gabe 0,0316 Duckie 0,028

(non pondérées), calculées sur les quatre jeux (et définies à la section 3.3). Afin de pouvoir obtenir des scores pour certaines mesures qui nécessitent un réseau connecté (comme la distance moyenne ou le chemin moyen le plus court), nous avons restreint cette étude à la plus grande composante connexe de chacun des réseaux (c'est-à-dire que l'on a exclu les nœuds qui ne sont pas connectés au reste du réseau). Les cinq dernières lignes du tableau présentent uniquement les cinq valeurs les plus hautes pour les différentes mesures sur les nœuds et arêtes, dans un souci de concision.

Le calcul de ces scores nous permet d'observer différents phénomènes : si l'on cherche à repérer un personnage principal, on

peut voir très rapidement que le personnage ayant le poids le plus élevé pour chaque jeu à la sous-section 6.5.2 (c'est-à-dire dans l'ordre Max, Sean, Chloe et Alex) obtient également les scores les plus hauts, pour les quatre jeux, en ce qui concerne le degré, la force, la centralité de proximité et la centralité intermédiaire. On peut ainsi prendre la mesure de l'importance du personnage dirigé par le joueur ou la joueuse, puisqu'il prend toutes les casquettes (personnage le plus exposé, qui a le plus d'échanges avec les autres, qui est associé aux personnes les plus influentes du réseau et qui sert de meilleur intermédiaire entre les autres personnages).

On peut également s'interroger sur la très grande fermeture triadique de LIS1, cohérente avec sa plus grande clique de 26 personnages et sa densité élevée. Le réseau semble ainsi très interconnecté, ce qui pourrait venir du fait que ce jeu se déroule dans une école (qui peut prendre certains aspects d'un huis clos, avec de nombreux personnages qui se côtoient), de même que LISBEF qui a des scores assez similaires (proportionnellement à sa taille totale plus petite). À l'inverse, LIS2 et LISTC sont bien moins connectés et bien moins denses : dans le cas de LIS2, cela peut faire écho à son scénario moins « communautaire » et plus centré autour d'un noyau dur qui est en fuite (et qui peut donc rencontrer des personnages successivement sans pour autant qu'ils se connaissent entre eux) ; pour LISTC, outre le fait que le scénario se déroule également dans un cadre plus large (tout un village), les valeurs sont certainement influencées par l'abondance de personnages dont le poids est très faible (parfois même plus bas que 1, comme présenté dans la sous-section 6.5.2), et qui ne sont d'ailleurs pas toujours nommés (Parent 3, Anxious Girl, etc.).

Pour les différentes valeurs par nœud ou par degré, on peut par exemple s'étonner du degré très haut des personnages « Man » et « Woman » dans LIS1. Une vérification dans les scripts et dans les réseaux par épisode nous apprend que ces personnages apparaissent principalement dans le dernier épisode, lors de la visite

de la galerie. En réalité, il s'agit d'un regroupement de plusieurs anonymes, désignés à chaque fois comme « Man » ou « Woman » et prenant ainsi une importance disproportionnée par rapport aux rôles qu'ils représentent (et donc d'une question d'encodage, qui aurait pu être résolue dans le script en leur donnant des numéros pour les distinguer). On remarque d'ailleurs que si leur degré est important, ils ne sont pas dans le classement des forces les plus grandes : ils sont présents dans certaines scènes qui les mettent en connexion avec beaucoup de personnages différents, mais ils n'ont que peu d'interactions au total, ce qui relativise leur impact concret.

Toujours dans LIS1, on peut s'étonner que Chloe, pourtant perçue comme un des protagonistes principaux de l'histoire, ne soit pas plus visible dans le classement des degrés des nœuds (où elle occupe la huitième position avec un degré de 30) et de la centralité de proximité (où elle figure en dixième position avec un score de 0,597). Ces résultats laissent à penser que si elle est souvent présente aux côtés de Max (ce dont témoigne la force du nœud, et le poids de l'arête qui relie Chloe et Max), elle n'est pas pour autant en contact avec tout le reste du réseau, et Max est régulièrement seule pour interagir avec les autres personnages.

Il est intéressant d'identifier le personnage qui apparaît le plus souvent en deuxième position, à travers les quatre jeux. Pour LIS1, LIS2 et LISBEF, un personnage se détache clairement (notamment dans la force des nœuds) : on a respectivement Chloe, la meilleure amie de Max qui la suit dans son enquête, Daniel, le frère de Sean qui l'accompagne dans sa fuite, et Rachel, meilleure amie de Chloe dans le préquel. Dans LISTC toutefois, Jed et Ethan ont des valeurs très proches dans les différents scores : Jed est en réalité l'antagoniste du récit, alors qu'Ethan est un des amis d'Alex qui lui propose également son aide.

Enfin, on peut observer la présence d'un trio dans LIS2, seul jeu parmi les quatre dont les cinq arêtes aux degrés les plus hauts ne sont pas toutes reliées au personnage principal. En effet, en

cinquième position, on trouve les interactions entre Daniel et Cassidy, qui bouclent un triangle d'interactions très fortes puisque Daniel et Sean, et Sean et Cassidy sont également présents dans les cinq premières places du classement. Cassidy est un des personnages que Sean et Daniel rencontrent lors de leur voyage ; les différentes mesures témoignent de son rôle important au sein du réseau et du récit lui-même.

6.5.5 Observations par épisode

En ce qui concerne le découpage par épisode, on peut observer les phénomènes en faisant par exemple un réseau dynamique, comme l'illustre la figure 6.9 pour LIS1. Il est alors facile de voir l'évolution (dans sa présence et dans ses interactions) du personnage de Chloe, par exemple, qui est plus discret dans les épisodes 2 et 5 alors que les épisodes 3 et 4, plus intimistes, semblent

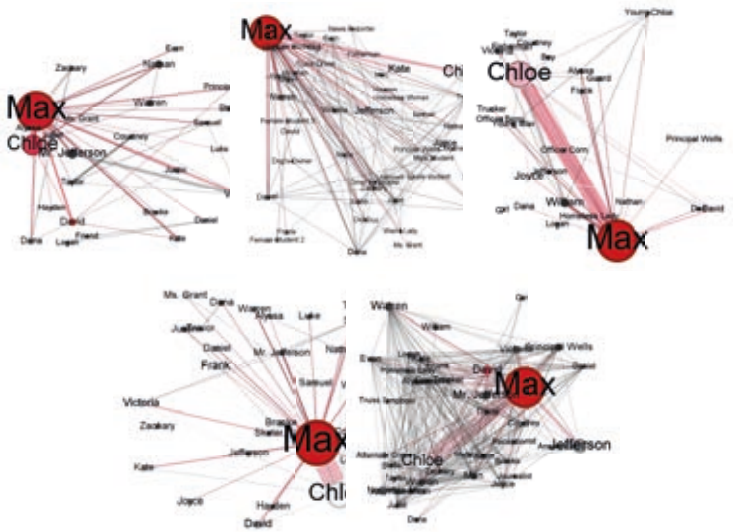


FIGURE 6.9 Réseau dynamique par épisode pour LIS1.

TABLE 6.5 Poids, écart-type, écart-type relatif, valeur minimale et valeur maximale pour le personnage de Kate dans LIS1.

épisode	poids	std	rsd	min	max
1	13,50	0,50	3,70	13	14
2	57,98	2,54	4,39	49	67
4	12,63	1,58	12,47	11	16
5	6	0	0	6	6

se concentrer beaucoup sur la relation entre Max et Chloe. L'épisode 5 ressemble à un grand rassemblement de tous les personnages croisés au cours du jeu, ce qui correspond à son rôle de conclusion du récit.

Dans LIS1, les choix laissés aux joueurs et joueuses peuvent notamment mener au décès ou à la survie du personnage de Kate, à l'épisode 2. Pour voir si ce phénomène est visible dans les données, on regarde l'évolution du nœud de Kate à travers les cinq épisodes, dont le résultat est reporté dans le tableau 6.5. On rencontre Kate dans l'épisode 1, avec une présence qui peut varier selon les choix et un écart-type relatif de 3,70. Ensuite, l'épisode 2 met en scène le long dialogue qui va déterminer du sort de ce personnage : c'est l'épisode dans lequel elle est le plus présente (comme en témoigne le poids), avec un écart-type relatif proche de celui du premier épisode (4,39). Kate n'est pas présente dans l'épisode 3, peu importe l'issue de cette discussion. En revanche, les interactions avec Kate à l'épisode 4 n'ont lieu que si elle a survécu : on peut alors s'étonner que sa valeur minimale soit à 11 (et non à 0), mais il faut avoir à l'esprit que la création automatisée du flux narratif des différents jeux de *Life is Strange* n'a pas permis de tenir compte des choix dépendant de choix précédents. Les traversées sont donc autorisées à passer par des chemins qui leur seraient normalement inaccessibles selon les décisions prises plus tôt dans le jeu, et certaines parties simulées sont donc en réalité impossibles. Malgré ce manque de précision dans

le modèle, on peut tout de même constater que dans cet épisode, l'écart-type relatif explose et monte à 12,47, ce qui donne tout de même un indice sur la présence plus variable de Kate qu'au début du jeu. Enfin, l'épisode 5 mentionne Kate par des souvenirs, qui sont stables dans tous les scénarios, d'où un écart-type nul dans cette dernière partie.

6.6 Ouverture sur la suite

Les résultats présentés dans la section 6.5 se basent sur une approche simplifiée de la narration de *Life is Strange* : on considère (à tort) tous les chemins comme possibles, indépendamment des choix faits dans les épisodes précédents, on part de l'idée que tous les embranchements sont équiprobables et on compte le nombre de répliques et d'échanges, sans tenir compte de leur nature. Si ces choix mettent déjà en évidence des résultats intéressants et cohérents par rapport au scénario des différents jeux, il va sans dire que l'on pourrait envisager des études bien plus approfondies de ces jeux.

En assouplissant les contraintes techniques (soit par une passe manuelle, soit en affinant l'algorithme de détection des embranchements), on pourrait construire un flux narratif plus fidèle à la construction du jeu, qui supprime les transitions impossibles pour ne générer que des scénarios existants, et ainsi obtenir des résultats encore plus fiables et marqués : pour reprendre l'exemple de Kate à la section 6.5, on verrait nettement qu'elle disparaît parfois complètement de l'épisode 4, mettant bien plus clairement en évidence le fait que sa survie dépend des choix laissés aux joueurs et joueuses.

On pourrait également ajouter manuellement au flux narratif les informations fournies par le jeu sur la proportion de joueurs et joueuses qui optent pour chaque option d'un choix important (comme présenté en section 6.4), de sorte à générer les traversées les plus classiques (qui passent par les transitions à la probabilité

la plus haute) ou les moins plébiscitées (qui passent par les transitions à la probabilité la plus basse), pour voir ce qui les différencie et ce que tout ceci permet de déduire quant au public qui joue à ces jeux.

Une autre piste intéressante, liée à la présence quasi totale du personnage dirigé par les joueurs et joueuses dans l'ensemble du jeu, serait de produire des réseaux de personnages avec et sans ce personnage, afin de mieux ressentir son impact sur le reste du réseau. Cela pourrait par exemple permettre une analyse plus fine du rôle de l'acolyte de ce personnage, en mettant en lumière la manière dont il peut ou non « exister » socialement en l'absence du héros.

Pour ouvrir des questions encore plus larges, on pourrait également s'aventurer à faire du *clustering* de réseaux (à l'image d'Arinik *et al.* [2020] sur des réseaux de votes au parlement), c'est-à-dire de générer une grande quantité de réseaux de personnages basés sur des traversées aléatoires, puis de les regrouper en différentes catégories de réseaux relativement similaires pour compter le nombre de « groupes » de réseaux possibles sur le jeu. Cette méthode pourrait permettre d'identifier de grandes tendances, sujettes à quelques petites variations, mais potentiellement rattachées à de grands arcs narratifs différents selon les décisions prises.

Enfin, pour répondre à la question de départ : ces jeux qui se revendiquent centrés sur les choix laissent-ils réellement une liberté aux joueurs et joueuses ?, il serait passionnant de se pencher non pas sur le nombre d'interactions, mais sur leur nature. Au vu du travail titanesque déployé pour développer un jeu à embranchements multiples (puisqu'il s'agit de coder l'équivalent de plusieurs jeux, alors que chaque joueur et joueuse n'en expérimentera qu'un seul s'il ou elle ne choisit pas de lancer plusieurs parties), on pouvait en effet anticiper en partie les résultats obtenus sur les différents *Life is Strange* en matière de variation très relative des réseaux si le poids des nœuds et arêtes ne témoigne que

de la présence des personnages et interactions. En revanche, il est vraisemblable que la pluralité des expériences du jeu se traduise davantage par une évolution variable des liens entre ces personnages, que l'on puisse se faire des amis ou des ennemis en fonction des choix qu'on fait. En récupérant les répliques sur la même base que le code utilisé ici pour capturer les occurrences des personnages, et en utilisant par exemple un algorithme de *sentiment analysis* (qui est justement entraîné pour donner un score aux phrases selon différentes émotions véhiculées) pour identifier automatiquement les phrases amicales ou hostiles, on pourrait construire un réseau de personnages basé sur les affinités, à la manière de Lee et Jung (2019), et observer la variation du poids de ses arêtes au cours des traversées générées.

En étendant la recherche à d'autres types de jeux, le flux narratif peut également se confronter à d'autres questions et amener d'autres éclairages au média vidéoludique. On peut par exemple se demander comment découper un monde ouvert en flux narratif, comment en définir les unités narratives et les transitions. A priori, un jeu totalement ouvert nécessiterait un flux narratif très « permissif », sans séquentialité forcée des unités à traverser. Mais selon les analyses désirées, on pourrait tout de même se baser sur les expériences des joueurs et joueuses ou sur l'envie d'observer une progression « encouragée » par le jeu pour construire des transitions en ce sens. Dans *Outer Wilds* par exemple, qui offre un accès immédiat à toutes les zones de la carte, mais s'appuie sur des connaissances à débloquer progressivement pour comprendre les mécaniques du jeu, on pourrait construire le flux narratif en partant du principe que les joueurs et joueuses doivent acquérir ces connaissances (et en laissant de côté les profils qui tentent de terminer le jeu le plus rapidement possible ou qui ont consulté des solutions préalablement) et ainsi retrouver une forme de séquentialité.

On peut également s'interroger sur la déclinaison des notions de flot maximal et de capacités, tirées de la théorie des graphes,

au flux narratif. Pour chaque arête d'un graphe, la capacité est définie comme la quantité maximale de flot qui peut passer par cette arête. Le problème du flot maximal est ainsi la recherche de la plus grande quantité possible de flot qui puisse traverser le graphe depuis la source jusqu'au puits, en respectant la capacité de chaque arête. Dans le cas du flux narratif adapté à des jeux vidéo, on peut par exemple imaginer le cas d'un jeu de type *dating simulation*, c'est-à-dire un jeu qui propose de développer des affinités avec différents personnages, en général dans un temps limité. L'usage des capacités pourrait permettre de modéliser le nombre maximal d'interactions que l'on peut créer avec un même personnage, et la somme des capacités des transitions menant au puits pourrait indiquer le nombre maximal d'unités de temps à disposition pour activer ces interactions. Pour aborder l'aspect de la réputation de son personnage (autre mécanique récurrente des *dating simulations* qui consiste à interdire l'accès à certains dialogues ou personnages tant que l'on n'a pas atteint un certain degré de notoriété en multipliant les interactions fructueuses), on pourrait s'appuyer sur la contrainte des « choix implicites » mentionnée à la section 6.4, qui permet de débloquer l'accès à certaines transitions sous des conditions précises (dans la section 6.4, il s'agit d'unités narratives précédemment visitées ou non, dans cet exemple-ci on pourrait fixer des seuils minimaux d'un score de réputation qui augmente en passant par certaines transitions).

Dans le cas d'une étude sur la pratique du *speedrun*, discipline consistant à atteindre le plus rapidement possible un objectif donné (souvent la fin d'un jeu) en exploitant autant les raccourcis prévus que les *bugs*, on pourrait imaginer transformer ces capacités en résistances (illustrant ainsi l'effort à fournir pour passer par une transition), afin d'observer la facilité d'exécution des chemins les plus courts pour passer de la source au puits.

De manière générale, si le présent ouvrage a pour objectif d'ouvrir une porte vers une méthode d'approche des narrations interactives, les possibilités d'approfondissement sont multiples et les résultats potentiels sont, quant à eux, remplis de promesses.

Troisième partie

Œuvres plurielles et comparaisons

7 Approche théorique

La comparaison de réseaux est un problème largement rencontré dans l'analyse de données et il existe de nombreuses manières de l'aborder. Comme l'exposent Wills et Meyer dans leur article sur les mesures (2019), la plupart des méthodes de comparaison présentent des temps de calcul prohibitifs et chaque situation nécessite une approche individualisée, selon la taille et la nature des réseaux à comparer ainsi que le type d'information que l'on cherche à mettre en évidence.

Dans le périmètre de cet ouvrage, le souhait est de comparer des réseaux de personnages tirés de livres et de leur adaptation cinématographique, dans le but de déterminer si ladite adaptation est « fidèle » à l'original ou à quel point elle s'en écarte. Après avoir constitué un jeu de données (qui sera présenté à la section 8.1) composé d'une dizaine de livres et de leur adaptation en films, l'idée sera donc de sélectionner des outils pertinents pour mesurer la « distance » entre chaque paire d'œuvres, afin de voir notamment si certaines adaptations sortent du lot, par une proximité spécialement forte ou faible.

Les questions de fidélité et de théorie des adaptations sont traitées au chapitre suivant. Je vais d'abord me concentrer sur le

choix et l'élaboration de mesures adaptées à ces objets. Ce chapitre est articulé en deux grandes parties : en se basant sur l'existence de mesures (définies au chapitre 3) calculées sur les différents réseaux et utilisées pour des interprétations en sciences humaines, j'examinerai d'abord une première stratégie consistant à comparer ces mesures à travers les réseaux du livre et du film. Ensuite, je choisirai trois mesures de comparaison [21], avec plusieurs variations de leurs définitions, avant de tester ces différentes méthodes sur des données au chapitre 8.

On aboutit ainsi à deux approches complémentaires :

— la comparaison de mesures (section 7.1) : approche monadique, qui applique les mesures sur une seule œuvre à la fois (et qui met en perspective ces valeurs avec celles des autres œuvres étudiées),
— les mesures de comparaison (section 7.2) : approche dyadique, qui prend une paire d'œuvres et calcule des scores de similarité ou de dissimilarité entre ces œuvres.

7.1 Comparaison de mesures

Une première idée pour aborder la comparaison de réseaux est celle de la comparaison de leurs mesures : comme dans l'annexe B.6, on se concentre ici non pas sur la visualisation ou la cartographie des réseaux, mais sur les mesures effectuées sur chaque réseau et introduites à la section 3.3. Du fait que ces mesures sont, dans les cas favorables, interprétables en sciences humaines, on imagine ainsi pouvoir observer des différences ou des similitudes entre les adaptations. Par exemple, un score de densité plus faible dans le roman, couplé à un nombre de person-

[21] Dans la suite de cet ouvrage et par souci de simplification, le terme « mesure de comparaison » est choisi pour désigner à la fois les mesures de similarité et les mesures de dissimilarité sur les réseaux.

nages plus important, indique peut-être un réseau plus vaste, moins interconnecté, et une histoire plus resserrée autour d'un noyau dur de personnages dans le film. Sur un grand nombre de données, ces scores peuvent également permettre d'identifier des tendances, si elles existent, pour vérifier par exemple la croyance intuitive qu'il y a moins de personnages dans un film que dans le livre dont il est adapté.

Aparté sur la pondération

À ce stade, il peut être intéressant d'aborder la question de la pondération dans les mesures de réseaux et de son effet sur l'interprétation des résultats. Par défaut, la plupart de ces mesures sont définies de manière non pondérée, c'est-à-dire que l'on considère les arêtes et les nœuds comme absents ou présents, en leur assignant un poids de 0 ou 1. Or, dans certains contextes, on préfère utiliser des variantes pondérées des mesures, pour tenir compte de l'intégralité des informations du réseau : ces poids peuvent alors s'interpréter comme des capacités ou comme des résistances, et différentes définitions sont proposées pour intégrer cette pondération aux mesures binaires, selon que l'on choisit d'agréger toutes les informations (et de donner la même force à un nœud possédant 10 arêtes de poids 1, ou une seule arête de poids 10, par exemple) ou de donner une importance différente au *nombre* d'arêtes et au *poids* cumulé de ces arêtes (voir par exemple l'article d'Opsahl *et al.* [2010] et ses références).

Dans le cas des réseaux de personnages, il est rare de tomber sur des recherches qui choisissent explicitement d'utiliser des mesures pondérées, ou qui justifient le choix d'une démarche binaire. Et pourtant, ce choix a un effet non négligeable sur l'interprétation des résultats. Prenons un exemple très simple pour illustrer ce phénomène. Imaginons un récit construit autour d'un personnage central (P), en contact rapproché avec deux grandes familles ennemies (A et B), qui ne se parlent jamais. Admettons

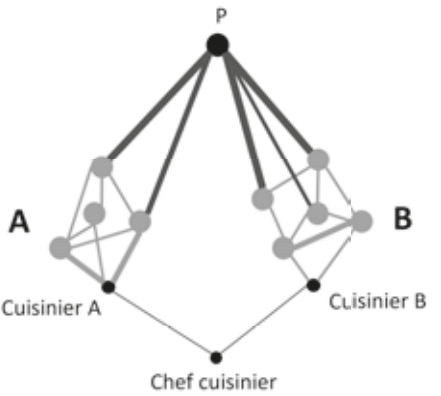


FIGURE 7.1 Exemple de réseau de personnages lourdement influencé par la pondération des mesures.

que chaque famille emploie un cuisinier et que, à une seule occasion dans l'histoire, apparaisse un chef cuisinier qui communique tour à tour avec chacun des deux cuisiniers. Le réseau de personnages associé à cette histoire ressemblerait donc à la figure 7.1. Si l'on veut observer la centralité intermédiaire (voir définition 13, page 33) dans sa version non pondérée, on constatera que le chef cuisinier a une mesure de centralité intermédiaire non nulle, car il se trouve sur le chemin le plus court entre les deux cuisiniers (qui se relient en deux arêtes en passant par lui). Il en sera de même pour plusieurs autres personnages de l'histoire (P bien sûr, mais aussi d'autres personnages qui peuvent servir d'intermédiaires, en raison de la nature de ce réseau).

Or, on pourrait également se dire que le chemin le plus efficace pour faire passer une information du cuisinier A au cuisinier B, plutôt que d'attendre la seule fois du récit où le chef cuisinier fait son apparition, est de la faire transiter par P au travers des familles, puisque leurs contacts sont bien plus fréquents. Pour refléter ce phénomène, on peut donc choisir de considérer une mesure pondérée de la centralité intermédiaire, dans laquelle les

poids des arêtes sont vus comme des capacités (les poids plus élevés représentant des chemins « plus efficaces » pour véhiculer de l'information). Avec cette approche, le chef cuisinier risque bien de ne plus être l'intermédiaire privilégié de qui que ce soit (en matière de rapidité de transfert d'information), et donc sa centralité intermédiaire tombe à 0.

En réalité, les deux approches sont pertinentes, selon les phénomènes que l'on cherche à observer, le tout étant de les choisir en toute conscience pour interpréter correctement les valeurs obtenues. L'attitude que je privilégie tout au long de ce travail, et qui constitue une troisième manière d'aborder la question, est de fonctionner avec les deux variantes de mesures en parallèle : en calculant pour chaque nœud la centralité intermédiaire pondérée *et* non pondérée, on a la possibilité de comparer ces deux scores et d'aller observer de plus près les endroits qui présentent une forte variation, en particulier les nœuds qui ont une centralité nulle dans un seul des deux scénarios. Pour reprendre l'exemple de notre chef cuisinier :

— le score non pondéré de centralité intermédiaire lui donne une valeur non nulle, potentiellement haute, mais pas nécessairement extrême (ce qui risque de le noyer au milieu d'un tableau trié par ordre de grandeur),
— sa centralité intermédiaire pondérée est de 0, comme plusieurs autres personnages, ce qui ne le distingue pas nécessairement des autres non plus,
— la combinaison des deux, par contre, peut le faire émerger comme un des rares personnages à passer d'une centralité importante à une centralité nulle, incitant ainsi à aller observer plus en détail son rôle dans l'histoire globale.

On peut aborder de la même manière beaucoup de mesures sur les réseaux, avec toujours l'idée qu'une meilleure compréhension en profondeur des outils numériques permet des observations plus fines et précises sur les œuvres étudiées.

7.2 Mesures de comparaison

La sélection de mesures intéressantes pour la question de la comparaison de réseaux de personnages n'est pas triviale. Beaucoup d'articles proposant des mesures de comparaison de graphes ont pour but de comparer de nombreux graphes afin d'en dégager des tendances, ce qui n'est pas notre cas (l'idée étant ici de comparer les œuvres deux par deux). Lorsque l'on souhaite se prononcer sur la différence entre deux graphes, avec l'objectif d'observer des modifications précises et locales en plus d'un phénomène global, il est nécessaire de combiner plusieurs approches et de les sélectionner avec soin.

Comme on peut le voir par exemple dans l'article de Wills et Meyer (2019) qui passe en revue différentes mesures en donnant des indications précises sur leurs forces et leurs faiblesses, les mesures permettant de différencier des graphes sur la base d'informations locales (n'impliquant que les relations entre voisins immédiats) sont peu nombreuses et peinent à tenir compte de l'importance des différences observées (l'emplacement dans le réseau, au centre ou en périphérie, n'influe pas sur le score associé au changement).

Donnat et Holmes (2018) classent également les mesures disponibles en deux catégories générales et plutôt mutuellement exclusives : les « distances structurelles » (basées directement sur la structure des réseaux), qui capturent des changements locaux, et les « distances spectrales » (basées généralement sur les valeurs propres des matrices d'adjacence) qui se concentrent sur les phénomènes globaux. Cette division ne semble toutefois pas convenir pour toutes les mesures (le transport optimal présenté plus bas, par exemple, ne rentre pas de manière évidente dans l'une ou l'autre de ces catégories), mais il est intéressant d'approcher les mesures par l'angle de la précision et de l'échelle des informations qu'elles mettent en évidence.

En suivant cette logique, trois mesures seront présentées dans ce chapitre et utilisées dans le suivant :

— la distance d'édition de graphes, à la sous-section 7.2.1. Cette distance structurelle est un classique de la comparaison de réseaux qui devrait permettre de capturer des changements locaux, mais sa grande simplicité ne laisse que peu d'espoir quant à des interprétations poussées de ses résultats,
— le transport optimal, à la sous-section 7.2.2. Issue d'une autre tradition mathématique tournée vers la transformation d'une distribution en une autre, cette distance semble pertinente pour la question des adaptations et permet en outre une transformation concrète d'un réseau en l'autre, faisant l'objet d'expérimentations additionnelles à la section 8.5,
— le coefficient RV, présenté à la sous-section 7.2.3. Mesure classique de similarité dans le domaine de l'analyse de données, cette distance capturera les changements globaux et offrira ainsi des informations complémentaires aux deux autres mesures.

7.2.1 Distance d'édition de graphes

La distance d'édition de graphes ou *graph edit distance* (GED) est un dérivé de la distance d'édition pour le texte, qui repose usuellement sur la distance de Levenshtein. L'idée centrale est de compter le nombre d'opérations nécessaires pour transformer un objet en un autre, par ajout ou retrait d'éléments.

De la même manière, la distance d'édition de graphes est le coût minimal de la séquence d'opérations nécessaires pour transformer un graphe en un autre. Il existe différentes manières de définir cette distance et ce coût, selon la façon dont les graphes sont construits (avec ou sans labels, par exemple) et les objectifs à atteindre. La GED est notamment souvent utilisée dans le cas de reconnaissance de motifs pour sa tolérance au bruit et à la distorsion (voir les travaux de Gao *et al.* [2010]).

Dans notre cas, les nœuds de nos réseaux possèdent des labels, mais les arêtes n'en ont pas. Les opérations de transformation d'un graphe à l'autre sont donc les suivantes (comme défini par exemple chez Sorlin et Solnon [2005]) :

— ajout/retrait d'un nœud,
— ajout/retrait d'une arête,
— changement de label d'un nœud.

Par défaut, chaque opération ne dépend pas du poids des nœuds et arêtes transformés, comme présenté dans la revue de Akoglu *et al.* (2014) qui font ainsi la distinction avec la distance de Showbridge *et al.* (« Error Correcting Graph Matching Distance », 1999). On peut alors choisir une fonction très simple, qui attribue à chaque opération un coût de 1, ou construire une fonction plus complexe pour pondérer différemment l'importance de certaines opérations par rapport aux autres. Le temps de computation des algorithmes de GED est en général cubique [22] ou carré [23] (par rapport au nombre de nœuds maximal des graphes) selon les implémentations.

7.2.1.1 Exemple

Dans l'exemple de la figure 7.2, on constate que pour transformer le réseau A en réseau B le plus simplement possible, il faut réaliser les opérations suivantes :

— retrait de l'arête Alice-Paul,
— changement de label du nœud Lea en Stan,
— ajout du nœud Fiona,
— ajout de l'arête Alice-Fiona,

ce qui, pour une fonction de coût définie comme le nombre d'opérations effectuées, correspond à une distance d'édition de graphe de 4.

[22] Comme chez Riesen (2016).
[23] Comme chez Donnat et Holmes (2018).

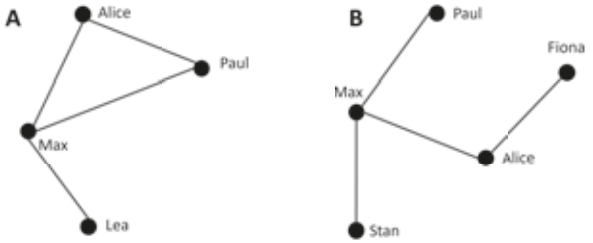


FIGURE 7.2 À gauche le réseau initial (*A*), à droite le réseau transformé (*B*).

Interprétation

Comme l'exemple peut déjà le laisser suggérer, le problème principal de cette distance, dans notre cas, tient dans la difficulté d'interprétation de son résultat. La valeur n'étant pas bornée, elle est sensible autant à la taille initiale de chacun des deux réseaux comparés qu'à l'écart de taille entre les deux. Si l'emploi de la GED pour une collection de réseaux peut tout à fait permettre de les regrouper en opérant des comparaisons par paires et en identifiant des sous-ensembles de réseaux relativement similaires (par exemple, dans le cas de réseaux dynamiques qui varient dans le temps et qui décrivent les mêmes objets, comme les réseaux de personnages des différents chapitres d'un même roman), la valeur d'une seule distance entre deux graphes ne permet pas de dire grand-chose des similarités et des différences entre lesdits graphes, comme le concluent notamment Donnat et Holmes (2018).

7.2.2 Transport optimal

Comme l'expose pour la première fois Monge dans son mémoire sur les déblais et les remblais (1781), la présentation informelle du problème du transport optimal peut s'aborder de la manière suivante : on souhaite déplacer différents tas de terre (déblais)

dans des espaces prévus pour les accueillir (remblais). En cherchant à minimiser le coût de ce transport (en matière de distance ou d'effort, par exemple), on remarque que toutes les solutions ne sont pas égales et qu'il en existe une qui permet une minimisation de ce coût. Sur un plan mathématique (voir par exemple le livre de Villani [2008]), le problème du transport optimal vise ainsi à minimiser une fonction entre deux espaces métriques X et Y, « transportant » une mesure de probabilité μ sur X vers une mesure de probabilité ν sur Y avec un coût de transformation préalablement défini.

Cas général discret

La situation qui nous intéresse concerne des données discrètes de taille finie (en l'occurrence, des listes de personnages de livres et de films), nous pouvons donc nous concentrer sur cette version spécifique du problème du transport optimal, qui comporte évidemment un nombre important de variations. On peut trouver par exemple chez Solomon (2018) une formalisation de ce type de problème. Si l'on considère m origines (les déblais de l'exemple de Monge, ou ici les personnages d'un livre) et n destinations (les remblais, ou ici les personnages d'un film adapté du livre), sachant que la quantité totale de « matière » (la terre, ou la présence totale des personnages qu'on transpose du livre au film) transportée est identique dans X et Y, on peut définir les proportions de matière $f_1, \ldots, f_m$ présentes aux m origines et $g_1, \ldots, g_n$ attendues aux n destinations, de sorte que

$$\sum_{i=1}^{m} f_i = \sum_{j=1}^{n} g_j = 1. \tag{7.1}$$

On introduit également la notion de couplage, purement mathématique dans sa définition et qui détermine la manière dont on transporte le premier espace métrique vers le deuxième, ou ici la façon dont on transforme les personnages du livre en personnages du film :

Définition 16. *Soient (X, μ) et (Y, ν) deux espaces métriques. Alors le couplage de μ et ν est la construction d'une mesure π définie sur $X \times Y$ telle que pour tous les ensembles mesurables $A \subset X$, $B \subset Y$, on a que $\pi(A \times Y) = \mu(A)$ et $\pi(X \times B) = \nu(B)$.*

On peut dire de manière équivalente que π admet une marge de μ sur X et de ν sur Y. Il est également possible de considérer π_{ij} comme la probabilité jointe de sélectionner l'origine i et la destination j.

Dans notre cas, la mesure de couplage $\pi \in \mathbb{R}^{n \times m}$ doit satisfaire :

$$\sum_{j=1}^{n} \pi_{ij} = f_i \text{ et } \sum_{i=1}^{m} \pi_{ij} = g_j.$$

En ajoutant à cela une fonction de coût c sur $X \times Y$, pour laquelle $c_{ij} \geq 0$ représente le coût de transport de l'origine i à la destination j, on peut écrire le coût total de transport comme

$$\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} \pi_{ij}, \tag{7.2}$$

qu'on cherche donc à minimiser sur les π. En d'autres termes, le coût total de transformation des personnages du livre en personnages du film est la combinaison du couplage choisi (c'est-à-dire, la façon dont on a choisi de répartir les personnages du livre sur les personnages du film) et des coûts cumulés de chaque transformation selon ce couplage. Minimiser ce coût total de transformation revient à choisir le couplage qui coûtera le moins cher, c'est-à-dire la table de correspondances entre personnages du livre et du film qui nécessitera le moins d'effort pour chaque passage d'un personnage à l'autre. Avant de pouvoir résoudre cette équation, il faut choisir la définition du coût, pour exprimer l'effort à consentir pour transformer ces personnages : est-ce basé sur le nombre de répliques ou sur la place dans le réseau de personnages, par exemple ? De nombreuses définitions sont possibles, et la sous-section suivante en propose deux.

Application aux personnages et propositions de coût

L'application de ce problème à des réseaux de personnages peut se faire de différentes manières. Dans tous les cas, il faut d'abord passer par une étape de normalisation des distributions [24] : il faut s'attendre à ce que la somme totale des poids des nœuds ne soit pas identique dans le livre et dans le film ; pour satisfaire les prérequis du transport optimal, les valeurs des poids sont donc transformées en pourcentages de la somme totale (pour chaque œuvre), ce qui permet bien d'avoir une somme des poids à 1 (et donc à 100 % d'apparitions ou d'interactions réparties sur les différents personnages) pour le livre et pour le film.

Considérons une distribution $f_1, \ldots, f_m$ des m personnages d'un livre (où f_1 correspond à la proportion d'apparitions ou d'interactions du premier personnage de la liste), et la distribution $g_1, \ldots, g_n$ des n personnages du film adapté de ce livre. Selon la formule 7.2, on définit la dissimilarité de ces deux distributions comme

$$d(f, g) := \min_{\pi} \sum_{i}^{m} \sum_{j}^{n} c_{ij} \pi_{ij}. \tag{7.3}$$

Reste à définir les coûts c_{ij} de transformation entre les deux distributions de personnages. Supposant que certains personnages sont communs aux deux œuvres, je propose deux approches pour calculer les coûts, afin de pouvoir les comparer dans la suite (à la sous-section 8.3.1). La démarche la plus directe est de favoriser les transports qui respectent la correspondance des personnages, et de pénaliser les transformations d'un personnage vers un autre. On peut le faire à l'aide d'un simple coût discret, défini

[24] En mathématiques, la « distribution » est une forme généralisée de fonction. Ici, il faut y voir la manière dont les personnages sont distribués dans l'œuvre, c'est-à-dire leur nombre d'apparitions ou d'interactions (selon la version choisie) sur l'ensemble de l'œuvre.

de la manière suivante :

$$c_{ij} := \begin{cases} 0 & \text{si } i = j \\ 1 & \text{si } i \neq j \end{cases} \tag{7.4}$$

Cette fonction de coût considère donc de la même manière tous les personnages qui ne seraient pas identiques, sans tenir compte de leur position dans le réseau de personnages associé à chaque œuvre, ou à une quelconque notion de proximité dans le récit : si deux personnages sont identiques dans le livre et dans le film, leur coût de transformation est nul, s'ils sont différents, leur transformation coûte 1.

Avec cette fonction de coût, la formule 7.3 de dissimilarité entre deux distributions devient donc

$$d(f, g) = \frac{1}{2} \sum_{i=1}^{n} |f_i - g_i|. \tag{7.5}$$

En effet, pour trouver le couplage idéal qui minimise le coût, il faut transporter autant que possible les personnages sur eux-mêmes : pour un personnage i, π_{ii} prend ainsi la valeur minimale entre f_i et g_i, c'est-à-dire que si i est plus fréquent dans le livre que dans le film ($f_i > g_i$), on le transporte sur l'intégralité de la destination ($\pi_{ii} = g_i$), et si i est plus fréquent dans le film ($g_i > f_i$), alors on transporte l'intégralité de l'origine sur la destination ($\pi_{ii} = f_i$). Par la construction de c_{ij}, les valeurs de π_{ii} sont multipliées par 0, et les valeurs restantes dans la somme sont composées de l'écart restant entre f_i et g_i. Le facteur $\frac{1}{2}$ vient quant à lui de la formule de la distance en variation totale (qui dit notamment que $|x| = 2x_+ - x$, où x_+ est la partie positive de x).

Dans une idée de contraste, et pour essayer de récupérer une notion de distance entre les personnages d'origine et les personnages de destination dans la fonction de coût, on propose également une seconde approche, qui sera à mettre en perspective avec la première dans la sous-section 8.3.1.

Pour élaborer la notion de distance qui nous servira de coût, on s'inspire de la « distance euclidienne carrée itérée »[25], qui propose pour une configuration (f, D) (où f est la distribution normalisée à 1, et $D = (d_{ij})$ une matrice de dissimilarité entre chaque paire de personnages au sein d'une même œuvre) la relation suivante :

$$4\, d_{ij}^{(2)} := \sum_k f_k (d_{ik} - d_{jk})^2 - \sum_{kl} f_k f_l (d_{ik} - d_{jk})(d_{il} - d_{jl}). \quad (7.6)$$

Il nous faut les ingrédients suivants pour décliner cette formule dans notre situation :

— des poids étendus à un même ensemble de personnages,
— des dissimilarités entre personnages dans une même œuvre (livre ou film).

On commence par redéfinir les distributions de chaque œuvre pour les étendre à tous les $N \leq n+m$ personnages distincts issus de la fusion entre les deux listes de personnages. Ainsi, $f_1, \ldots, f_N$ et $g_1, \ldots, g_N$ sont désormais deux listes de distributions à N éléments, et on fixe $f_i := 0$ si le personnage i n'est pas présent dans le livre (et de la même manière, $g_i := 0$ si le personnage i n'est pas présent dans le film).

On peut alors construire des *poids de compromis* combinant l'importance de chaque personnage dans le livre et dans le film, et normalisés pour préserver la somme des poids à 1 comme dans le cas général discret 7.1 :

$$h_i := \frac{\sqrt{f_i g_i}}{Z}, \quad (7.7)$$

où $Z := \sum_{i=1}^{N} \sqrt{f_i g_i}$. Par définition Z vaut 1 lorsque tous les personnages sont identiques, et 0 lorsque tous les personnages sont

[25] L'excellente définition de cette formule provient du professeur François Bavaud, expert en statistiques et virtuose du formalisme mathématique. Je le remercie vivement de m'avoir permis de l'utiliser ici.

différents. Notons que les poids h_i sont nuls pour les personnages présents dans une seule œuvre, et strictement positifs sinon. Par la normalisation, on a bien que $\sum_{i=1}^{N} h_i = 1$.

Quant aux mesures de dissimilarité au sein d'une œuvre, on peut se baser sur la matrice d'adjacence (voir la définition 3, page 25) pondérée (par le poids des arêtes), et normalisée sur les lignes, du réseau de personnages pour représenter une distance entre paires de personnages dans l'œuvre.

Avec ces différents éléments, on peut exprimer le coût de transformation du livre en film comme

$$d_{ij} = \sum_{k=1}^{N} h_k (d_{ik}^A - d_{jk}^B)^2 - \sum_{k,l=1}^{N} h_k h_l (d_{ik}^A - d_{jk}^B)(d_{il}^A - d_{jl}^B), \quad (7.8)$$

où (d_{ij}^A) et (d_{ij}^B) sont les mesures de dissimilarité au sein du livre et du film, respectivement. On remarque que ce coût n'est finalement influencé que par les personnages communs aux deux œuvres, pour lesquels le coefficient h est non nul. Attention toutefois à ne pas confondre le poids de compromis (utilisé pour construire le coût de transformation basé sur la distance carrée itérée) et le poids qui sera attribué aux nœuds dans le calcul de transport optimal (qui pourra découler du nombre total d'occurrences du personnage, ou de la somme de ses interactions). La distinction sera rappelée et détaillée à la sous-section 8.3.1 lors du calcul des résultats.

Interprétation

L'approche du transport optimal apporte différents types de résultats, qui peuvent être intéressants à analyser. D'une part, on obtient un score pour chaque couple d'œuvres, représentant la distance minimale de leurs distributions en fonction du coût choisi. Bien que difficile à interpréter en soi (comme la distance d'édition 7.2.1), ce score peut servir de point de comparaison entre différents couples d'œuvres, ajouté aux autres mesures collectées.

D'autre part, pour le couplage considéré comme optimal, on peut aussi observer la façon dont les personnages du livre sont répartis sur les personnages du film (en quelle proportion et sur quelle[s] destination[s]), ce qui peut donner lieu à des interprétations peut-être insolites, mais pas dénuées de sens. Une petite étude de cet ordre est menée à la section 8.5 pour proposer des pistes de réflexion sur cette méthode, avec le triplet d'œuvres autour de *Romeo and Juliet*.

Dans mon étude de cas pratique (en sous-section 8.3.1), j'ai choisi de comparer deux choix de distributions (f_i) pour chaque œuvre :

— une distribution basée sur les poids des nœuds du réseau, définis ici comme la somme des occurrences de chaque personnage dans l'œuvre (et donc indépendants du nombre d'interactions avec les autres personnages), que l'on appellera désormais *poids d'apparitions*,
— une distribution basée sur la somme des poids des arêtes du réseau, représentant donc la somme des cooccurrences de chaque personnage avec tous les autres dans le réseau, que l'on appellera *poids d'interactions*,

ce qui permet de donner une importance différente aux personnages en fonction de ce que l'on veut mettre en priorité (leur importance « absolue », ou leur lien avec le reste du réseau). Dans le cas des poids d'interactions, il faut garder à l'esprit que pour une fenêtre d'observation donnée, le nombre de mentions d'un même personnage n'influence pas son poids : la présence de deux personnages dans une même fenêtre crée une seule cooccurrence entre eux, indépendamment de leur nombre de mentions.

Dans le cas de la deuxième approche de coût (équation 7.8), il est possible que la diagonale de la matrice correspondant aux distances euclidiennes carrées itérées ne soit pas nulle. J'ai choisi d'observer là aussi deux phénomènes en parallèle : le transport optimal avec les matrices préservées, et le transport optimal lorsque l'on force les diagonales à 0 (et que l'on donne donc une

indication explicite au couplage sur la façon minimale de relier les personnages identiques), que l'on appellera transport optimal *rectifié*. Cette rectification rejoint en partie la stratégie du coût discret (équation 7.4) en encourageant fortement le couplage entre personnages identiques, mais elle préserve toutefois les distances entre deux personnages différents, permettant ainsi une nuance plus fine dans l'élaboration du transport optimal.

Toutes ces approches combinées donnent une série de cas de figure (deux pour le premier coût, et quatre pour le deuxième) qui seront commentées en détail dans la partie pratique en sous-section 8.3.1.

7.2.3 Coefficient RV

Introduit par Escoufier en 1973 dans son article « Traitement des variables vectorielles », puis dans un article dédié avec Robert (1976), le coefficient RV (« rhô-vectoriel ») est une célèbre mesure de similarité entre des matrices carrées symétriques dont les valeurs propres sont nulles ou positives (c'est-à-dire des matrices « semi-définies positives »). La définition originale du coefficient RV concerne des matrices de covariance entre variables (et fait intervenir des matrices transposées). Une définition duale, mais équivalente, porte sur les matrices symétriques d'affinité entre les individus (ce qui simplifie la notation puisque par construction, ces matrices sont égales à leur transposée), et s'écrit de la manière suivante :

Définition 17. *Soient A, B deux matrices semi-définies positives de dimension $n \times n$, représentant des configurations d'un même ensemble de n individus. Alors*

$$RV_{AB} := \frac{\text{Trace}(AB)}{\sqrt{\text{Trace}(A^2) \cdot \text{Trace}(B^2)}}, \tag{7.9}$$

où $\text{Trace}(A)$ *est la somme des valeurs de la diagonale de* A.

De nombreuses études utilisent le coefficient RV comme indicateur de similarité pour diverses analyses de données en statistiques multivariées, et nombre de travaux évaluent également la pertinence et la précision de cet indicateur (voir par exemple ceux de Josse *et al.* [2008] ou ceux de Abdi [2010]). Dans notre cas, le choix du coefficient RV est motivé par plusieurs raisons :

— par définition, il est compris entre 0 et 1, contrairement aux deux mesures précédemment présentées, ce qui devrait permettre une meilleure intuition de proximité ou de distance entre deux œuvres, dès lors que le score de similarité se rapproche de 1 ou de 0,
— il est possible de décomposer le coefficient RV en cinq termes identifiables, comme le présentent Bavaud et Métrailler (2023), ce qui augmente les possibilités d'interprétation de cet indicateur de similarité.

Application aux personnages

La définition originale du coefficient RV (voir la définition 17, page 125) ne tient pas compte d'une éventuelle pondération des n individus. Une généralisation du coefficient RV a été proposée par Bavaud (2023), permettant d'intégrer un poids aux n individus pour autant qu'il soit identique dans les deux configurations. Or, notre cas est encore différent : un personnage n'aura en général pas le même poids dans le livre ou dans le film, il faut donc trouver une solution pour étendre encore la définition 17 à des configurations admettant des poids *différents* pour les n individus. La démarche est décrite en détail dans l'article de Bavaud et Métrailler (2023) et s'appuie sur des techniques de statistiques et d'analyse multivariée. Elle se présente comme suit :

1. On transforme le réseau pondéré de chaque œuvre en une *configuration euclidienne carrée* (f, D), où f représente la distribution normalisée à 1 comme dans la sous-section 7.2.2 et où D est une matrice de distance euclidienne carrée (c'est-

à-dire que chaque entrée d_{ij}, représentant la distance entre le personnage i et le personnage j de l'œuvre, est de la forme $\|x_i - x_j\|^2$, où x_i est le vecteur qui représente le nœud i). Cette étape est purement formelle et peut être effectuée de différentes manières, tant la question de l'extraction de dissimilarités euclidiennes carrées sur un réseau pondéré est vaste.

2. On s'appuie sur le MDS (*Multi-Dimensional Scaling*), un ensemble de méthodes permettant de projeter des ensembles de points dans un espace de dimension réduite en préservant au mieux leurs distances, pour visualiser nos deux configurations euclidiennes carrées et permettre une comparaison entre elles. Pour ce faire, on passe par la construction d'un *kernel*, ou *matrice de produit scalaire pondéré* : la méthode du kernel est, là aussi, un vaste champ d'études, exposé par exemple dans le livre de Shawe-Taylor et Cristianini (2004), qui permet de résoudre en temps linéaire et avec une efficacité satisfaisante des problèmes en haute dimension en se basant sur des fonctions « d'apprentissage instantané ».
3. Nos deux kernels symétriques K_A et K_B construits sur les œuvres A et B sont ensuite insérés dans le calcul du coefficient RV, qui devient alors

$$\mathrm{RV}_{AB} = \frac{\mathrm{Trace}(K_A K_B)}{\sqrt{\mathrm{Trace}(K_A^2) \cdot \mathrm{Trace}(K_B^2)}}. \tag{7.10}$$

Ce coefficient tient ainsi compte de la distribution pondérée des deux œuvres et offre un score de similarité entre 0 et 1, comme le coefficient RV original.

Le code et les détails d'implémentation sont à retrouver en intégralité sur le répertoire character_network.

Décomposition et interprétation

Les deux distributions f et g représentant les poids des personnages peuvent varier énormément entre les deux œuvres et l'on s'attend donc à ce que le coefficient RV soit vulnérable à ces variations. Il peut être intéressant de faire un parallèle entre ce score et un coefficient RV « de compromis », que l'on aurait calculé sur la base des poids de compromis définis en 7.7 et mettant en avant les personnages communs aux deux œuvres uniquement. C'est ce que l'on fait dans le cadre de la décomposition du coefficient en 5 parties annoncé plus haut et expliqué en détail dans l'article de Bavaud et Métrailler (2023). En prenant comme point de départ non pas la similarité $\texttt{RV}_{AB}$ mais la dissimilarité associée $-\ln(\texttt{RV}_{AB})$ (pour transformer les produits en somme et améliorer la lisibilité de la formule), on peut obtenir par développement mathématique la décomposition suivante :

$$\underbrace{-\ln(\texttt{RV}_{AB})}_{\text{dissimilarité combinée}} = \underbrace{-\ln(\texttt{RVh}_{AB})}_{\text{dissimilarité de compromis}} \quad \underbrace{-2\ln(Z)}_{\substack{\text{dissimilarité entre les}\\ \text{poids des personnages}}} \quad \underbrace{-\frac{1}{2}\ln(\Gamma_A)}_{\text{dispersion relative dans A}}$$

$$\underbrace{-\frac{1}{2}\ln(\Gamma_B)}_{\text{dispersion relative dans B}} \quad \underbrace{-\ln(1+\epsilon)}_{\text{correction du centroïde}} . \tag{7.11}$$

Notons que :

— le premier terme est composé de ce coefficient RV de compromis $\texttt{RVh}_{AB}$, dont les kernels K_{hA} et K_{hB} sont construits sur les poids de compromis h_i. Il représente ainsi une dissimilarité basée uniquement sur les personnages communs des deux œuvres,

— le deuxième terme représente l'écart entre les deux distributions f (de l'œuvre A) et g (de l'œuvre B) : Z (défini à la formule 7.7) est une mesure de similarité qui vaut 0 s'il n'y a aucun personnage commun et 1 si $f = g$,

— le troisième terme concerne la dispersion relative au sein de l'œuvre A. La fonction $\Gamma_A := \frac{\text{Trace}(K_{hA}^2)}{\text{Trace}(K_A^2)}$ mesure le rap-

port des dispersions de la matrice de dissimilarité D_A par rapport aux poids h_i (en haut) et par rapport aux poids f_i (en bas),
— le quatrième terme est construit de la même manière que le précédent, par rapport à l'œuvre B. Si le terme est positif, cela laisse supposer que l'importance relative des personnages est plus grande dans la version de départ que dans la version ajustée avec les poids de compromis (et donc que les personnages qui ne sont pas communs aux deux œuvres augmentent la dispersion générale du réseau),
— le dernier terme est une petite quantité résiduelle résultant de la transformation vers les poids de compromis et du déplacement des centroïdes entre les distributions basées sur f et h.

En guise d'observation finale, on peut faire un rapprochement entre le Z présent dans cette décomposition et la distance d liée au transport optimal dans le cas du coût discret (voir la formule 7.5). Par construction, on peut définir Z comme $1 - d_2^2$, où d_2 est la distance de Hellinger (voir le livre de Kuo [1975] pour plus de détails), ce qui nous permet de relier Z et d par l'inégalité suivante :

$$1 - Z \leq d \leq \sqrt{1 - Z}.$$

Cette inégalité induit une forme d'équivalence entre Z et d : on peut notamment observer que si Z vaut 1, d vaut 0 (et inversément).

En réalité, d et d_2 (la distance de Hellinger reliée à Z) sont des dissimilarités qui n'impliquent que la distribution des personnages (c'est-à-dire leur poids d'apparitions) et pas leurs interactions. Dans la deuxième approche du transport optimal ainsi que dans le coefficient RV, on ajoute les informations liées à la structure du réseau (et donc aux interactions) en utilisant la distance carrée itérée (formule 7.6).

8 Étude de cas : comparaison d'adaptations

Comme l'expose Chan dans sa revue des *adaptation studies* (2012), l'étude des films adaptés de romans est un champ qui connaît une popularité grandissante depuis les années 1990. À l'intersection des études de traduction et d'interculturalité (pour les aspects de transposition d'une œuvre d'un contexte à un autre), cette discipline encore très jeune et profondément hybride peine à trouver ses méthodes et son périmètre, ainsi que le montre Elliott (2017).

La question de la fidélité occupe et divise la communauté scientifique depuis les débuts de la discipline. Johnson (2017) en parle comme d'un concept qui a commencé par réunir les premiers chercheurs et chercheuses du domaine avant de devenir plus récemment une notion à rejeter, en considérant notamment que le fait de placer la fidélité parmi les critères d'appréciation d'une œuvre force à la voir sous le prisme du médium de départ et crée des attentes biaisées, comme l'expose par exemple Cardwell (2002). D'autres chercheurs comme Kranz (2007) acceptent plus volontiers les outils de comparaison et la notion de fidélité, pour autant qu'ils ne s'accompagnent pas d'un discours évaluatif.

Boyd thématise également cette question dans son article (2017) en prenant pour parallèle l'adaptation en biologie. Il considère qu'une bonne adaptation doit mêler fidélité et fertilité : une part de fidélité pour s'appuyer sur ce qui a fonctionné dans l'itération précédente, mais également une part de fertilité pour emmener la nouvelle itération plus loin et l'ajuster aux contraintes du nouvel environnement.

Dans le périmètre de cet ouvrage, l'idée n'est pas de faire une étude qualitative et évaluative des adaptations par rapport au livre d'origine, mais de proposer une approche basée sur des méthodes quantitatives afin de collecter des données et des mesures servant d'indicateurs de distance entre les deux œuvres. Comme dans le reste de ce travail, l'interprétation de ces résultats est en grande partie laissée aux experts et expertes du domaine, avec l'espoir que ces premières explorations éveillent leur intérêt sur les outils proposés.

8.1 Récolte des données

Il est bien plus difficile de trouver des scripts de film intégraux que des romans au format numérique. Ainsi, pour constituer le jeu de données utilisé dans l'étude des adaptations, notre choix s'est d'abord porté sur la base de données IMSDB [26] pour chercher de manière automatique les scripts de film qui contiennent le texte « *on the novel by* », indiquant que le script est adapté d'un roman. Notons que IMSDB comporte uniquement des sources en anglais, ce qui conditionne la langue du jeu de données. Parmi les fichiers ainsi identifiés, dix candidats ont été sélectionnés, tous basés sur un roman facilement accessible au format numérique (sur Gutenberg ou par achat de livre numérique) et présentant une diversité à différents niveaux :

— genre : contemporain, fantastique ou polar,

[26] https://imsdb.com/, consulté le 23.10.2023.

— narration du roman : première ou troisième personne,
— nombre de personnages du roman : moins de 50, entre 50 et 100, au-delà de 100,
— écart présumé du nombre de personnages entre le livre et le film : nombre plus important de personnages dans une œuvre ou dans l'autre,
— public cible : enfant ou adulte,
— année de parution du roman : avant 1900, entre 1900 et 2000, après 2000.

Comme on peut le constater dans le tableau 8.1, toutes les œuvres sélectionnées présentent une combinaison différente des critères de diversité et chaque modalité de ces critères a une représentation minimale de 20 % sur l'ensemble du jeu de données. Notons également la présence de deux œuvres littéraires de l'écrivain James Ellroy, *The Black Dahlia* et *L.A. Confidential*, qui diffèrent surtout par leur narration. L'emploi de la première personne a d'ailleurs été observé pour mesurer l'importance du narrateur dans l'histoire : dans *The Black Dahlia* justement, le roman est raconté par Bucky Bleichert, qui enquête sur le meurtre et tient une place cruciale dans le livre comme dans le film ; dans *The Other Boleyn Girl*, la narratrice est Mary Boleyn, également un des personnages principaux de l'histoire dans ses deux formats ; enfin, *The Help* propose une narration à trois voix, et la première personne désigne tantôt Aibileen Clark, tantôt Minny Jackson, et tantôt Skeeter Phelan, le trio central du livre comme du film.

Il est évident que pour tirer des conclusions solides de phénomènes à large échelle, un jeu de données bien plus important serait nécessaire. Comme souvent, c'est la limite du temps et de l'étendue du projet qui s'oppose à la constitution d'un tel jeu de données, tant l'étape de mise en correspondance des personnages est chronophage (voir section 8.2).

Un autre souhait plus difficile à exaucer est la présence dans le jeu de données d'un triplet, permettant des comparaisons deux

TABLE 8.1 Répartition des dix œuvres selon les différents critères de diversité. Les couleurs servent ici à mettre en évidence les modalités des critères de diversité présentés plus haut.

Œuvre	Genre	Narr.	# perso	f/l	Public	Parution
Romeo and Juliet	contemporain	3	24	0,95	adulte	1597
Sense and Sensibility	contemporain	3	54	0,35	adulte	1811
Anna Karenina	contemporain	3	194	0,23	adulte	1877
Narnia	fantastique	3	33	1	enfant	1950
The Black Dahlia	polar	1	204	0,18	adulte	1987
The Silence of the Lambs	polar	3	93	0,31	adulte	1988
L.A. Confidential	polar	3	153	0,17	adulte	1990
The Other Boleyn Girl	contemporain	1	82	0,22	adulte	2001
Coraline	fantastique	3	24	1,33	enfant	2002
The Help	contemporain	1	174	0,16	adulte	2009

à deux d'une œuvre déclinée sur trois supports. De telles œuvres sont évidemment courantes, mais l'acquisition de trois sources textuelles exploitables présente un défi non négligeable. Un seul triplet est donc constitué dans ce jeu de données : *Romeo and Juliet*, dont j'ai réuni la pièce de théâtre de William Shakespeare (1567) [27], l'adaptation en film de Baz Luhrmann (1996) et l'opéra de Charles Gounod (1867) (qui est la seule source en français : toutefois, la technique de récupération des personnages par expressions régulières n'est pas influencée par la langue du texte).

Les réseaux de personnages des romans et des films sont extraits avec Charnetto (voir l'annexe B pour le détail des opérations). Le livret d'opéra et la pièce de théâtre ont, quant à eux, fait l'objet d'ajustements manuels pour identifier la structure du texte et récupérer automatiquement le découpage des scènes ainsi que les mentions des différents personnages.

[27] Nous appellerons cette pièce de théâtre « livre » à travers l'ensemble du travail pour uniformiser avec les autres romans.

8.2 Correspondance des personnages

Avant de pouvoir comparer les livres et les films, il faut établir une correspondance entre les personnages à travers les œuvres. Cette tâche a été réalisée manuellement et pose une question de design : que faire lorsqu'un personnage est remplacé par un autre, au nom différent mais au rôle similaire, dans l'adaptation ? Dans le cadre de cette étude, la règle a été fixée ainsi : si un personnage est renommé pour des raisons de contexte ou d'époque (par exemple, dans le cas de l'adaptation *Romeo + Juliet* de Baz Luhrmann qui se déroule à Los Angeles dans les années 1990, certains noms de personnages sont modernisés, à l'image de Paris qui devient Dave), mais que son rôle est clairement identifié comme identique à travers les versions, on accepte une correspondance avec le personnage de l'œuvre originale. Dans les autres cas (création d'un personnage dans le film qui prend un rôle approchant d'un personnage du livre mais avec une autre identité, notamment), on considère que lesdits personnages n'ont pas de correspondant.

Les listes ainsi créées sont à trouver dans le répertoire character_network au format Excel et sont composées de deux colonnes (ou trois dans le cas de *Romeo and Juliet*, voir tableau 8.2), celle du livre et celle du film. Pour les personnages qui n'existent que dans une des versions, la case est laissée vide. Il est à noter que pour des romans extrêmement denses comme *L.A. Confidential*, beaucoup de noms considérés comme des noms de personnes sont en réalité des références à des personnalités publiques, acteurs et actrices ou musiciens et musiciennes, dont certaines entrées ont peut-être échappé à la vérification des correspondances. Toutefois, ces « faux personnages » ont une présence marginale ou inexistante dans les réseaux de personnages associés, si bien que la fiabilité des résultats obtenus sur les comparaisons n'est pas péjorée.

8.3 Résultats et analyses

L'analyse des résultats se divise en trois parties : on commence par examiner en détail le tableau général des résultats à la sous-section 8.3.1, selon les différentes mesures choisies, pour émettre des hypothèses sur les interprétations possibles. Ensuite, à la section 8.4, on se concentre sur l'exemple de *Romeo and Juliet*, seul

TABLE 8.2 Tableau de correspondance des personnages dans les trois versions de *Romeo and Juliet*.

film	**livre**	**opéra**
ROMEO	Romeo	ROMEO
JULIET	Juliet	JULIETTE
MERCUTIO	Mercutio	MERCUTIO
BENVOLIO	Benvolio	BENVOLIO
FATHER LAURENCE	Friar Laurence	FRÈRE LAURENT
NURSE	Nurse	GERTRUDE
CAPULET	Capulet	CAPULET
GLORIA	Lady Capulet	
TYBALT	Tybalt	TYBALT
CAPTAIN PRINCE	Escalus	LE DUC
DAVE	Paris	PARIS
SAMPSON	Sampson	
GREGORY	Gregory	GREGORIO
BALTHASAR	Balthasar	STEPHANO
MONTAGUE	Montague	
ABRA	Abram	
SUSAN		
CAROLINE	Lady Montague	
APOTHECARY	Apothecary	
COP		
	Peter	
	Servant	
	Friar John	FRÈRE JEAN

triplet du jeu de données, avant de faire un aparté sur le transport de personnages à la section 8.5.

8.3.1 Tableau général des résultats

Dans un souci de lisibilité, j'ai réparti les résultats obtenus en plusieurs tableaux. Dans chacun d'entre eux, la légende de la colonne média indique la paire d'œuvres concernées : « lo » pour livre-opéra, « fo » pour film-opéra, et « lf » pour livre-film. Les trois œuvres surlignées en jaune sont celles qui possèdent une narration à la première personne.

Dans cette section, je vais d'abord examiner en détail chacune des mesures et je proposerai ensuite une mise en perspective de ces différentes approches pour identifier leurs forces, leurs faiblesses et proposer enfin à la sous-section 8.3.2 des pistes pour les futures études d'adaptations.

Distance d'édition de graphes

Le tableau 8.3 regroupe à la fois les informations sur le nombre global de personnages (déjà observables dans le tableau 8.1 mais reproduites ici pour une meilleure vue d'ensemble) et sur ce qui concerne la distance d'édition de graphe (GED) présentée à la sous-section 7.2.1. Les colonnes sont définies comme suit :

- **#pers** : nombre total de personnages contenu dans chaque titre (en fusionnant les deux listes de la paire concernée),
- **f/l** : rapport entre le nombre de personnages du film et celui du livre (ou opéra sur film/livre dans le cas de *Romeo and Juliet*),
- **GED** : distance d'édition de graphe entre les deux réseaux de personnages,
- **GED/#pers** : GED divisée par le nombre total de personnages.

Le calcul de la GED a été fait avec l'algorithme [28] présenté par Abu-Aisheh *et al.* (2015). De ce premier tableau, on peut déjà remarquer la grande amplitude dans la taille des œuvres étudiées. De manière quasi systématique, plus les œuvres ont de personnages, plus le ratio film/livre est bas, ce qui indique un tri de plus en plus important effectué lors de l'adaptation à l'écran. Les œuvres les plus intimistes ont un ratio plus proche de 1, avec même dans le cas de *Coraline* un nombre plus important de personnages dans le film que dans le livre.

TABLE 8.3 Distance d'édition de graphes. Les trois tons de vert servent à mettre en évidence l'amplitude des valeurs, en les répartissant manuellement en trois catégories, le plus foncé représentant la distance la plus grande entre livre et film. Les trois tons de jaune, quant à eux, marquent la variation du rapport film/livre.

Titre	Média	#pers	f/l	GED	GED/#pers
Romeo and Juliet	lo	21	0,62	74	3,52
Romeo and Juliet	fo	21	0,65	50	2,38
Romeo and Juliet	lf	23	0,95	60	2,61
Coraline	lf	24	1,33	47	1,96
Narnia	lf	33	1,00	70	2,12
Sense and Sensibility	lf	54	0,35	363	6,72
The Other Boleyn Girl	lf	82	0,22	621	7,57
The Silence of the Lambs	lf	93	0,31	412	4,43
L.A. Confidential	lf	153	0,17	2838	18,55
The Help	lf	174	0,16	1102	6,33
Anna Karenina	lf	194	0,23	1338	6,90
The Black Dahlia	lf	204	0,18	1756	8,61

La distance d'édition de graphe, sans grande surprise, semble très influencée par la taille des graphes initiaux, si bien que les trois groupes identifiés sur le nombre total de personnages se retrouvent parfaitement dans la répartition des valeurs de GED (les

[28] Les détails d'implémentation sont à trouver sur le répertoire character_network.

œuvres à moins de 50 personnages obtiennent une GED dans les dizaines, celles à moins de 100 personnages produisent une GED dans les centaines, et au-delà de 150, on passe dans les milliers). Si l'on observe l'intérieur des groupes à la recherche d'interprétations plus ciblées, on peut voir par exemple que *Romeo and Juliet* (livre-opéra), avec ses 21 personnages, a un score plus élevé que *Narnia* malgré ses 33 personnages, et que *Coraline* semble avoir le plus de similarité entre ses deux réseaux. On peut également voir que même si le ratio de *Narnia* est à 1, il est le deuxième titre le plus haut en termes de GED. Dans le groupe intermédiaire, *The Other Boleyn Girl* sort un peu du lot avec son score de 621 et dans le groupe le plus fourni en personnages, *L.A. Confidential* bat des records de GED avec un score qui dépasse d'un facteur de 1,6 la deuxième valeur la plus haute (*The Black Dahlia*), avec 50 personnages de moins et un ratio film/livre relativement équivalent.

Lorsque l'on met en perspective la GED et le nombre initial de personnages, on peut tout de même voir que les deux valeurs ne sont pas entièrement proportionnelles. En réalité, ce calcul fait émerger très fortement *L.A. Confidential*, ainsi que, dans une certaine mesure, *The Black Dahlia* et *The Other Boleyn Girl*. À l'inverse, *The Silence of the Lambs* semble subitement plus similaire dans son adaptation que ne le laissait penser sa position initiale, son score se rapprochant de *Romeo and Juliet* (livre-opéra), bien que l'écart entre leurs deux listes complètes de personnages soit très important.

Transport optimal

Le tableau 8.4 concerne cette fois le transport optimal entre les deux réseaux de personnages de chaque titre. Comme présenté à la sous-section 7.2.2, plusieurs approches ont été mises en parallèle : trois coûts différents (coût discret, coût lié à la distance euclidienne carrée et le même coût « rectifié », c'est-à-dire en corrigeant la diagonale pour qu'elle soit nulle) et deux types de poids

TABLE 8.4 Transport optimal. Les trois tons de vert servent à mettre en évidence l'amplitude des valeurs, en les répartissant manuellement en trois catégories, le plus foncé représentant le coût de transport optimal le plus important.

Titre	Média	TOd	TOd#	TO$_o$	TO$_o$#	TO	TO#
Romeo and Juliet	lo	0,35	0,29	0,09	0,07	0,17	0,14
Romeo and Juliet	fo	0,29	0,25	0,06	0,05	0,11	0,07
Romeo and Juliet	lf	0,20	0,12	0,05	0,03	0,11	0,09
Coraline	lf	0,29	0,21	0,09	0,04	0,21	0,13
Narnia	lf	0,21	0,19	0,04	0,04	0,11	0,11
Sense and Sensibility	lf	0,32	0,24	0,07	0,07	0,16	0,22
The Other Boleyn Girl	lf	0,54	0,53	0,30	0,28	0,54	0,51
The Silence of the Lambs	lf	0,36	0,27	0,06	0,05	0,13	0,14
L.A. Confidential	lf	0,59	0,54	0,16	0,18	0,24	0,29
The Help	lf	0,33	0,34	0,12	0,11	0,33	0,29
Anna Karenina	lf	0,49	0,44	0,08	0,09	0,15	0,16
The Black Dahlia	lf	0,46	0,49	0,18	0,20	0,30	0,35

pour chacun de ces coûts (poids d'interactions ou poids d'apparitions, tels que décrit à la sous-section 7.2.2). Plus précisément, les colonnes font référence aux démarches suivantes :

— **TOd** : transport optimal calculé avec le coût discret d (voir équation 7.4) et les poids d'interactions,
— **TOd#** : transport optimal calculé avec le coût discret et les poids d'apparitions (#),
— **TO$_o$** : transport optimal calculé avec la distance euclidienne carrée itérée (voir équation 7.6) rectifiée, avec les poids d'interactions,
— **TO#$_o$** : identique à TO$_o$, avec les poids d'apparitions,
— **TO** : identique à TO$_o$, mais non rectifié,
— **TO#** : identique à TO#$_o$, mais non rectifié.

On peut déjà remarquer que le coût de transport optimal est généralement moindre en prenant pour les personnages le poids d'apparitions plutôt que le poids d'interactions. Il semblerait donc que leur distribution au sein des deux œuvres soit une information plus claire ou plus tranchée, permettant plus facilement à un algorithme de passer d'un réseau à l'autre.

La première approche, celle du coût discret, semble une fois de plus un peu trop triviale pour permettre des observations fines. Dans les valeurs qui échappent au tri initial sur le nombre de personnages par œuvre, on peut tout de même citer *Romeo and Juliet* (livre-opéra) et *The Help*, qui obtiennent des coûts de transport optimal très similaires alors que l'un a 21 personnages de départ et l'autre, 174. *L.A. Confidential* garde sa position d'adaptation la plus éloignée, suivie de près par *The Other Boleyn Girl*.

Lorsqu'on passe à l'autre approche basée sur la distance euclidienne itérée, toutefois, les classements évoluent. Regardons d'abord les colonnes centrales, qui concernent le transport optimal rectifié. Avec cette correction, tous les coûts globaux diminuent, et certaines œuvres commencent à sortir du lot. L'exemple le plus flagrant est *The Other Boleyn Girl*, qui dépasse nettement *L.A. Confidential*, mais aussi *The Black Dahlia* et *The Help*, dont le transport optimal reste coûteux malgré les ajustements. Or ces trois œuvres sont celles dont le roman est raconté à la première personne et dont les personnages principaux ont certainement été sous-capturés par les modèles de NER : il semblerait donc que si les personnages les plus importants sont stables d'une œuvre à l'autre, les scores de transport optimal basés sur la distance euclidienne itérée soient plutôt bas (à l'image de *Anna Karenina* qui obtient des coûts de transformation très raisonnables alors que ce titre se plaçait systématiquement dans les plus éloignés jusqu'à présent), bien que, pour les œuvres qui voient des fluctuations importantes dans la présence des personnages centraux, le coût de transformation reste conséquent. Notons également qu'en rectifiant la diagonale, les deux versions de poids des nœuds produisent le même ordre de grandeur pour les résultats et qu'il n'y a que pour *The Black Dahlia*, *Anna Karenina* et *L.A. Confidential* que le poids d'apparitions produit un coût total plus haut.

Dans les deux dernières colonnes du tableau 8.4, qui ne sont pas rectifiées, les coûts de transformation remontent un peu de manière attendue, mettant toujours en évidence claire les trois

œuvres dont le roman est écrit à la première personne. La seule surprise supplémentaire vient peut-être de *Coraline*, qui présente un coût relativement haut dans la variante non rectifiée et avec poids d'interactions. Pour comprendre ce phénomène, en nous penchant sur la distribution des poids des personnages pour le livre et le film, on peut voir que si, dans la plupart des adaptations, les trois ou quatre personnages au poids le plus haut sont également dans le top 5 du livre d'origine, la situation est différente dans *Coraline*. Il y a très peu de personnages en tout : Coraline est celle qui se détache le plus, et tous les autres ont moins de 10 % de présence totale, avec des classements qui varient beaucoup du livre au film. Il y a également la question de Wybie, personnage très présent dans le film (troisième en poids d'apparitions, cinquième en poids d'interactions) et totalement absent du livre. Ces différences importantes et ramenées à l'échelle d'une petite liste totale de personnages pourraient expliquer un score qui se détache du lot lorsqu'on ne facilite pas la correspondance des personnages entre livre et film en rectifiant la diagonale.

TABLE 8.5 Coefficient RV. Les trois tons de vert servent à mettre en évidence l'amplitude des valeurs, en les répartissant manuellement en trois catégories, le plus foncé représentant les dissimilarités les plus importantes. Par construction, un coefficient RV élevé indique une grande *similarité* : son amplitude est donc marquée par un dégradé de bleu pour éviter toute confusion.

titre	média	RV	-ln(RV)	-ln(RVh)	RV#	-ln(RV#)	-ln(RVh#)
Romeo and Juliet	lo	0,64	0,45	0,35	0,71	0,35	0,30
Romeo and Juliet	of	0,54	0,62	0,29	0,60	0,50	0,18
Romeo and Juliet	lf	0,73	0,32	0,36	0,62	0,48	0,30
Coraline	lf	0,51	0,67	0,40	0,68	0,38	0,39
Narnia	lf	0,39	0,94	0,38	0,64	0,45	0,39
Sense and Sensibility	lf	0,35	1,05	0,58	0,65	0,42	0,76
The Other Boleyn Girl	lf	0,23	1,45	1,81	0,16	1,84	2,04
The Silence of the Lambs	lf	0,35	1,06	0,41	0,40	0,93	0,28
L.A. Confidential	lf	0,23	1,46	0,63	0,19	1,65	0,89
The Help	lf	0,28	1,28	0,88	0,15	1,90	1,00
Anna Karenina	lf	0,23	1,46	0,38	0,18	1,72	0,39
The Black Dahlia	lf	0,35	1,06	0,50	0,29	1,23	0,63

Coefficient RV

La dernière mesure de comparaison figure dans le tableau 8.5. Comme dans la partie sur le transport optimal, on distingue ici les deux approches pour le poids des nœuds (poids d'interactions et poids d'apparitions). Les colonnes sont organisées de la manière suivante :

- **RV** : le coefficient RV (expliqué à la sous-section 7.2.3) associé à la paire de réseaux de personnages (en bleu pour rendre plus évident le fait qu'une valeur élevée signifie cette fois-ci une similarité plus grande),
- **-ln(RV)** : le logarithme naturel du coefficient RV, dont on modifie le signe pour donner des valeurs positives (comme dans l'équation 7.11),
- **-ln(RVh)** : la dissimilarité de compromis telle qu'exprimée dans l'équation 7.11,
- **$RV^{\#}$** : le coefficient RV calculé à l'aide des poids d'apparitions,
- **-ln($RV^{\#}$)** : le logarithme naturel de ce coefficient RV,
- **-ln($RVh^{\#}$)** : la dissimilarité de compromis basée le poids d'apparitions.

Si l'on retrouve là aussi une tendance dans les scores qui semble dépendre de la taille initiale des réseaux (les valeurs en vert du haut sont globalement plus basses que les valeurs du bas), certains écarts commencent à se creuser. Prenons par exemple deux cas intéressants, ceux de *Anna Karenina* et de *The Silence of the Lambs*. Deux œuvres plutôt grandes en ce qui concerne le nombre de personnages, mais dont les valeurs de similarité sont globalement plus basses que leurs voisins : le second a un score très bas en distance d'édition de graphes ramenée au nombre de personnages dans le tableau 8.3, tous deux ont une ligne très claire dans le tableau du transport optimal 8.4 lorsqu'on prend la distance euclidienne carrée. Ce sont les seules œuvres à présenter un écart très important entre le coefficient RV et sa dissimilarité de

compromis RVh, pour les deux versions de pondération des nœuds. Si l'on interprète la dissimilarité de compromis comme une dissimilarité intrinsèque entre les deux réseaux, faisant abstraction des dispersions relatives des deux œuvres et se basant uniquement sur les personnages communs, cette information laisse peut-être penser que malgré des écarts importants entre le livre et le film, le « noyau dur » des personnages communs aux deux œuvres est plutôt stable tant dans *Anna Karenina* que dans *The Silence of the Lambs*.

À l'inverse, les valeurs calculées pour *The Other Boleyn Girl* confirment l'impression des tableaux précédents : les dissimilarités de compromis sont même supérieures au score cumulé, ce qui témoigne d'un écart majeur parmi les personnages communs au livre et au film, par rapport à l'écart mesuré sur l'ensemble des réseaux. Un regard sur la décomposition du log de RV nous indique par ailleurs des scores très élevés pour les cinq termes de l'équation 7.11, ce qui classe définitivement *The Other Boleyn Girl* parmi les adaptations les plus éloignées du roman de départ (compte tenu de la difficulté à récolter correctement les occurrences du personnage de Mary Boleyn, bien entendu).

Quant à *Coraline*, qui se distinguait parmi les « petits » romans dans les mesures de transport optimal, sa position est moins évidente dans ce tableau-ci. On voit tout de même une différence assez nette entre les deux versions de poids des nœuds : si sa similarité est la deuxième plus haute avec les poids d'apparitions, le phénomène est différent pour les poids d'interactions, et il se détache nettement des trois paires de *Romeo and Juliet*. Mais c'est surtout *Narnia* qui marque une grande dissimilarité dans son adaptation (là encore, seulement avec les poids d'interactions), surtout pour les personnages qui ne sont pas communs aux deux œuvres si l'on en croit les valeurs du RV et du RVh.

En dehors de ces quelques cas, la distinction entre les deux options de poids pour les nœuds ne semble pas mettre en lumière beaucoup de situations, et l'on remarque plus ou moins la même

tendance dans les deux parties du tableau. La version RV# semble surtout creuser les écarts entre les différentes œuvres : les œuvres plutôt similaires le sont encore davantage avec les poids d'apparitions, les œuvres peu similaires le sont encore moins, avec pour seule exception *Romeo and Juliet* (livre-film) qui obtient un meilleur score de similarité lorsqu'on prend en compte les poids d'interactions.

Comparaison des mesures

Reste la question de la comparaison des mesures calculées sur chaque réseau. Le tableau 8.6 recense les différences entre les deux valeurs pour chaque mesure et chaque paire d'œuvres [29] ; il est ensuite colorisé sur la base de la valeur absolue (pour identifier aussi bien les grands écarts positifs que négatifs). Les mesures en question sont les suivantes :

— **diam** : diamètre du réseau (voir définition 6, page 30),
— **excm** : excentricité moyenne du réseau (voir définition 14, page 34),
— **distm** : distance moyenne (moyenne des plus courts chemins) dans le réseau (voir définition 7, page 31),
— **densité** : densité du réseau (voir définition 8, page 31),
— **transit** : transitivité du réseau (voir définition 9, page 31),
— **mclique** : taille maximale des cliques dans le réseau (voir définition 9, page 31).

Les valeurs reportées dans le tableau 8.6 ne concernent que des mesures globales sur le réseau ; les mesures ciblées sur les nœuds et les arêtes ont aussi été calculées pour chaque titre et sont à retrouver dans les tableaux complets sur le répertoire

[29] La direction de la soustraction est indiquée par la colonne « média », en prenant toujours l'œuvre de droite moins l'œuvre de gauche. Par exemple, -0,28 correspond à la densité du film à laquelle on soustrait la densité du livre pour *Coraline*.

TABLE 8.6 Différence des mesures pour chaque paire de réseaux. Le dégradé de gris sert à mettre en évidence l'amplitude des valeurs, en les répartissant manuellement en trois catégories, le plus foncé représentant la dissimilarité la plus grande entre deux œuvres.

Titre	Média	Diam	Excm	Distm	Densité	Transit	mclique
Romeo and Juliet	lo	0	0,00	-0,08	0,12	0,04	-3
Romeo and Juliet	fo	0	-0,03	-0,38	0,31	0,26	3
Romeo and Juliet	lf	0	0,03	0,30	-0,19	-0,23	-6
Coraline	lf	1	0,97	0,45	-0,28	-0,29	-2
Narnia	lf	0	-0,08	-0,04	0,02	-0,02	1
Sense and Sensibility	lf	0	0,05	-0,31	0,32	0,22	-3
The Other Boleyn Girl	lf	-1	-0,62	-0,39	0,35	0,20	-7
The Silence of the Lambs	lf	0	-0,06	-0,12	0,04	-0,05	-5
L.A. Confidential	lf	1	0,68	0,33	-0,06	-0,15	-19
The Help	lf	-1	-0,36	-0,29	0,18	0,28	-3
Anna Karenina	lf	0	-0,28	-0,01	0,07	0,13	-5
The Black Dahlia	lf	-2	-0,93	-0,11	0,03	-0,12	-15

character_network. Elles sont parfois utilisées ici pour étayer certaines hypothèses ou observer des phénomènes locaux.

Les œuvres étudiées peuvent être distinguées en trois groupes sur la base des résultats de ce tableau. On a d'abord les adaptations qui semblent fidèles à l'original, ou qui présentent en tout cas peu de différences dans les mesures appliquées aux deux réseaux. C'est le cas par exemple de *Narnia*, qui obtient très peu d'écart sur toutes les mesures, ainsi que de *Romeo and Juliet* (livre-opéra), *The Silence of the Lambs*, et peut-être *Anna Karenina*, qui présente un écart principalement sur l'excentricité moyenne des personnages, ce qui peut laisser penser que la répartition d'importance des personnages n'est pas tout à fait identique entre les deux œuvres. Cette impression est confirmée par l'observation des nœuds les plus importants qui montre notamment une présence bien plus importante de Kitty dans le livre que dans le film.

Viennent ensuite les œuvres qui changent de taille d'un média à l'autre, sans pour autant bousculer totalement la structure globale des interactions entre les personnages. Ainsi, toutes les

notions de distance (diamètre, excentricité, distance moyenne) augmentent ou diminuent de manière importante en changeant de média, de même que la taille maximale des cliques qui évolue avec le nombre de personnages à disposition. Mais la densité et la transitivité ne changent que très peu, indiquant que la proportion de connexions et de cliques reste stable. On trouve dans cette catégorie *The Black Dahlia*, dont les distances évoluent de manière très nette (avec la clique maximale qui passe de 20 à 5 personnages dans le film), mais dont la densité reste quasiment identique. *L.A. Confidential* suit le même comportement, avec une clique maximale qui tombe aussi de 24 à 5 personnages. Les deux œuvres étant des histoires policières (toutes deux écrites par James Ellroy), on peut s'imaginer que les protagonistes principaux (enquêteurs, policiers et principaux suspects) sont restés stables dans l'adaptation, mais que beaucoup de trames secondaires ont été retirées à l'écran.

Reste la catégorie des œuvres qui présentent un grand écart de densité et de transitivité entre les versions, pas forcément assorti de fluctuations importantes dans la taille du réseau. *Coraline* et *The Other Boleyn Girl* en sont les représentants les plus parlants, de même que *Romeo and Juliet* (opéra-film). *The Help* et *Romeo and Juliet* (livre-film) peuvent également être rattachés à cette catégorie. Si le groupe précédent témoignait d'un changement d'échelle (sans pour autant modifier totalement la structure des interactions), ce groupe-ci présente des différences plus marquées dans les liens entre les personnages : dans le cas de *The Help* et *The Other Boleyn Girl*, il faudrait procéder à une analyse plus ciblée des personnages des romans pour diminuer ce problème de narration à la première personne, afin de voir ce qu'il reste de cet écart des mesures avec des réseaux de personnages plus précis. Dans le cas de *Coraline* toutefois, les changements narratifs sont nombreux entre le livre et le film, plusieurs personnages importants (dont Wybie) ayant été ajoutés dans le film et l'importance de chaque personnage fluctuant beaucoup entre les deux œuvres.

Pour *Romeo and Juliet*, les deux paires d'œuvres représentées ici concernent le film, qui a été transposé à une époque contemporaine, occasionnant également des changements importants dans l'existence et l'importance de plusieurs personnages.

Observations générales

Après avoir examiné de près les différents tableaux et les différents scores de similarité, certaines forces et faiblesses de chaque mesure semblent se dégager. Une première remarque est que la taille des réseaux a un impact non négligeable sur la plupart des mesures. Si la distance d'édition de graphes est la moins robuste à cette variation, le transport optimal et le coefficient RV fournissent également des valeurs plus importantes pour une taille plus grande de réseau. À l'instar du rééchantillonnage appliqué à l'exemple textuel de *Consider the Consequences* à la section 5.2, il existe des méthodes de rééchantillonnage de réseaux pour corriger cet effet de taille (voir par exemple l'article de Shan et Levina [2022]) ; ces méthodes n'ont pas été testées dans le cadre de ce projet, mais pourraient s'avérer prometteuses pour de futures recherches.

En l'état, il paraît donc difficile de découper les scores en intervalles absolus et de décréter que l'adaptation serait significativement proche ou éloignée de l'œuvre originale au-delà d'un certain seuil. Toutefois, lorsque l'on regroupe les titres selon l'ordre de grandeur de leur liste de personnages, on peut dégager des observations intéressantes sur la base d'une comparaison de scores afin de cibler les mesures qui s'écartent de la tendance du groupe. Notons tout de même que le transport optimal et le coefficient RVh peuvent produire des valeurs très basses malgré de grands réseaux de départ, comme c'est le cas pour *The Silence of the Lambs* et *Anna Karenina*.

La distance d'édition de graphes ne fournit que peu d'informations pertinentes, même lorsqu'elle est ramenée au nombre

total de personnages. Longue à calculer pour des réseaux importants et très influencée par la taille de chaque réseau, elle peine à capturer des différences pourtant bien visibles dans les autres approches, ce qui en fait un candidat peu intéressant pour la comparaison d'œuvres adaptées.

Le transport optimal offre des pistes plus intéressantes, d'autant plus lorsque l'on compare ses différentes variantes (rectifié ou non, et avec les deux options de poids des nœuds). Il permet en outre des observations sur la transformation de certains personnages en d'autres rôles, comme présenté à la section 8.5.

Le coefficient RV permet également différents niveaux d'observation par son découpage en cinq termes. Si le score de base, bien que compris entre 0 et 1, ne fournit pas un « pourcentage de similarité » extrêmement marqué (car bien influencé par la taille des réseaux), le passage en logarithme et la décomposition offrent une grille de lecture riche et précise. La différence entre RV et RVh permet de voir si la dissimilarité provient majoritairement des personnages communs aux deux œuvres ou de ceux qui sont exclusifs à l'une ou à l'autre, et les autres termes de la décomposition mettent également l'accent sur la dispersion de chaque œuvre, notamment.

Globalement, il est rassurant de voir que les trois romans écrits à la première personne obtiennent des dissimilarités notables tant dans le transport optimal que dans le coefficient RV : on s'attendait à une disparité exagérée dans l'importance du personnage principal entre livre et film, et ces scores témoignent bien de cet écart, inspirant ainsi confiance dans la pertinence générale de leurs valeurs.

Enfin, la démarche la plus robuste face aux variations de taille des listes de personnages est la comparaison des mesures sur chaque réseau, puisqu'il s'agit ici de comparer directement des scores propres à chacune des deux œuvres, plutôt que de calculer un score de similarité sensible à leur échelle commune. Cette dernière méthode, qui semble permettre d'identifier trois groupes

distincts parmi notre jeu de données à dix titres, met en évidence les différences d'amplitude, mais également de structure entre les deux œuvres étudiées, permettant là aussi des observations pertinentes.

8.3.2 Stratégie pour les futures études d'adaptations

Quelle stratégie adopter pour étudier de nouveaux couples d'œuvres parmi ces différentes options ? Sur la base de ces observations, notre recommandation dépend de la taille du jeu de données à disposition. Si l'on veut étudier plusieurs œuvres (une dizaine de paires, comme ici, ou davantage), une analyse regroupant transport optimal, coefficient RV et comparaison de mesures permettra des regards complémentaires et un nombre important de pistes d'analyse à confirmer ensuite en *close reading*. Si toutefois on souhaite se pencher sur un seul couple d'œuvres, la comparaison de mesures est l'option la plus susceptible de mettre en évidence des disparités concrètes, là où un score unique de transport optimal ou de coefficient RV sera peu parlant, à moins de tirer profit des résultats déjà calculés ici pour comparer les nouvelles valeurs avec les tableaux existants, ce qui devrait permettre une comparaison par ordre de grandeur et une intuition de l'écart entre valeurs obtenues et valeurs attendues pour une taille de liste de personnages donnée.

8.4 Exemple pratique : le triplet *Romeo and Juliet*

Un titre se détache dans le jeu de données constitué pour ce travail : *Romeo and Juliet*, qui a été analysé sous trois formes (la pièce de théâtre originale de Shakespeare, le livret d'opéra de Gounod et l'adaptation en film de Luhrmann). Que valent les observations de la sous-section 8.3.1 lorsqu'elles sont mises à l'épreuve d'un trio

TABLE 8.7 Le triplet *Romeo and Juliet*.

Média	livre-opéra	opéra-film	livre-film
#pers	21	21	23
f/l	0,62	0,65	0,95
GED	74	50	60
GED/#pers	3,52	2,38	2,61
TOd	0,35	0,29	0,20
TOd#	0,29	0,25	0,12
TO_o	0,09	0,06	0,05
TO_o#	0,07	0,05	0,03
TO	0,17	0,11	0,11
TO#	0,14	0,07	0,09
RV	0,64	0,54	0,73
-ln(RV)	0,45	0,62	0,32
-ln(RVh)	0,35	0,29	0,36
RV#	0,71	0,60	0,62
-ln(RV#)	0,35	0,50	0,48
-ln(RVh#)	0,30	0,18	0,30
Diam	0,00	0,00	0,00
Excm	0,00	-0,03	0,03
Distm	-0,08	-0,38	0,30
Densité	0,12	0,31	-0,19
Transit	0,04	0,26	-0,23
mclique	-3,00	3,00	-6,00

d'œuvres ? Les indices de similarité sont-ils cohérents les uns par rapport aux autres ?

Pour avoir une meilleure vue d'ensemble, j'ai reporté les résultats des tableaux précédents dans le tableau 8.7, en choisissant la teinte la plus foncée de vert pour la valeur maximale par ligne. Ce tableau se révèle très intéressant, car il permet de voir des distinctions nettes entre la méthode du transport optimal et celle du coefficient RV.

Le couple livre-opéra est celui qui a le plus grand écart de nombre de personnages (21 dans le livre contre 13 dans l'opéra). Toutes les variantes du transport optimal en font le couple le plus distant, alors même que c'est celui qui présente le moins de différences (et de loin) en matière de comparaisons de mesures ; il semblerait qu'il reste plus coûteux de transformer un réseau très grand en un réseau plus petit (ou vice versa) que d'ajuster des réseaux de taille similaire. À l'inverse, les mesures du film sont très éloignées des deux autres, avec un réseau bien moins dense et interconnecté.

Si l'on regarde le coefficient RV, les deux versions des poids de nœuds mettent en évidence cette fois le couple opéra-film, ce qui rejoint les conclusions des comparaisons de mesure (même si la distinction est moins marquée pour le coefficient $RV^{\#}$ calculé à l'aide des poids d'apparitions). Et enfin, les coefficients RVh construits sur les poids de compromis tendent quant à eux à identifier le couple livre-film comme le moins similaire : sachant que le personnage de Juliet est beaucoup moins présent dans le livre que dans le film ou l'opéra, et que le RVh s'intéresse principalement aux personnages en commun, il est possible que cette différence, couplée aux écarts de mesure plus importants entre livre et film qu'entre livre et opéra, soit à l'origine de ce score.

Bien que les valeurs soient parfois très proches et ne permettent donc pas des interprétations trop tranchées, il semblerait donc que l'on puisse retrouver dans ce triplet le même genre de phénomènes que dans l'étude plus large, avec des scores qui dépendent plus ou moins des personnages communs, des personnages centraux ou de la taille des réseaux selon les situations. Enfin, si une étude restreinte à ces trois œuvres ne permet pas de dire si elles sont « globalement similaires » pour ce qui touche aux réseaux de personnages, l'étude comparée menée à la sous-section 8.3.1 donne quelques éléments de réponse, lorsqu'ils sont mis en perspective avec *Coraline* ou *Narnia*.

8.5 Aparté : transport optimal de personnages

Comme présenté à la sous-section 7.2.2, le transport optimal est initialement prévu pour calculer des déplacements de matière d'un endroit à un autre. Si le coût de ce déplacement induit une mesure de distance entre deux configurations, ce qui présente un intérêt en soi, il est en outre possible de regarder de plus près quelle solution optimise le transport : dans le contexte de réseaux de personnages, il s'agit de voir sur quel(s) personnage(s) du film est envoyé tel personnage du roman d'origine, dans une démarche de minimisation des efforts requis pour cette transformation. En d'autres termes, si l'on considère les personnages du livre comme de la « matière première » dont la quantité de départ correspond à leur poids (d'apparitions ou d'interactions), et que l'on fixe comme objectif le réseau de personnages du film, dont les poids (d'apparitions ou d'interactions) des nœuds sont également spécifiés, l'algorithme va « distribuer » les personnages du livre sur les personnages du film. Sans cette information de poids, la démarche pourrait sembler naturelle et tout à fait directe ; or l'interprétation devient plus exotique dès lors que l'on réalise que chaque personnage est associé à une quantité, et que l'on va donc travailler avec des « pourcentages de personnages », pour le dire simplement.

En effet, imaginons que Romeo ait un poids de 0,2 dans la pièce de Shakespeare (ce qui revient à dire qu'il représente 20 % des apparitions totales des personnages de la pièce), et 0,1 dans le film, et que ces valeurs indiquent le poids de ce nœud dans les deux réseaux correspondants. Si l'algorithme considère que la solution la moins coûteuse est de transporter le Romeo de la pièce de théâtre sur son homologue cinématographique, remplissant intégralement le poids de 0,1 du nœud du film, il restera tout de même la moitié de la matière première « Romeo de la pièce » à distribuer sur les autres personnages du film, et il est tout à

fait possible que plusieurs personnages reçoivent une petite portion de Romeo dans leur « composition » finale. Dans la même logique, les personnages qui n'existent que dans une des deux œuvres créeront forcément des situations similaires, soit comme sources, soit comme récipients d'un ou plusieurs autres personnages, afin que tout le réseau de personnages de départ soit réparti dans le réseau de personnages d'arrivée.

On peut donc se demander si l'examen détaillé de ces répartitions a un sens, et s'il est possible d'interpréter les résultats sur un plan littéraire. Pour aborder cette question plus concrètement, j'ai choisi d'analyser le transport optimal de la pièce *Romeo and Juliet* de Shakespeare en son adaptation *Romeo + Juliet* de Luhrmann. Les tableaux suivants représentent la transformation des personnages de la pièce de théâtre (les lignes) en personnages du film (les colonnes), avec des valeurs allant de 0 à 1 pour indiquer les pourcentages (de sorte que la somme de chaque ligne soit égale à 1). Les trois tableaux sont basés sur le transport optimal avec distance euclidienne carrée itérée :

— le tableau 8.8 avec les poids d'apparitions (et la diagonale rectifiée ou non, ce qui produit exactement les mêmes valeurs),
— le tableau 8.9 avec les poids d'interactions,
— le tableau 8.10, avec les poids d'interactions et la diagonale rectifiée.

La première observation directe est liée à ce premier point : en se basant sur les poids d'apparitions (et donc sur le nombre total de répliques de chaque personnage, indépendamment de ses interlocuteurs), le fait d'encourager la correspondance entre les personnages identiques en rectifiant la diagonale ne fournit pas d'information supplémentaire au modèle, qui converge exactement sur la même répartition. Il semblerait donc que, dans cet exemple précis, la solution qui minimise le coût de transport soit stable même sans cette correction de diagonale. Le résultat se

TABLE 8.8 Transformation des personnages de Shakespeare en ceux de Luhrmann, avec poids d'apparitions (rectifier la diagonale ou non donne les mêmes résultats).

$OT_{ij}^{\#} / OT^{\#}$	ROMEO	JULIET	MERCUTIO	BENVOLIO	FATHER LAURENCE	NURSE	CAPULET	GLORIA	TYBALT	CAPTAIN PRINCE	DAVE	SAMPSON	GREGORY	BALTHASAR	MONTAGUE	ABRA	SUSAN	CAROLINE	APOTHECARY	COP
Romeo	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Juliet	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mercutio	0	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Benvolio	0	0	0,036	0,964	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Friar Laurence	0	0	0	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Nurse	0,334	0,041	0,101	0	0	0,497	0	0	0,156	0	0	0	0	0	0	0	0,007	0	0	0,020
Capulet	0	0	0	0	0	0	0,844	0	0	0,0112	0	0	0	0	0	0	0	0	0	0
Lady Capulet	0	0	0	0	0	0	0	0,989	0	0	0	0	0	0	0	0	0	0	0	0
Tybalt	0	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0	0	0	0	0	0
Escalus	0	0	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0	0	0	0	0
Paris	0	0	0	0	0	0	0	0	0	0	0,895	0	0	0	0	0	0,105	0	0	0
Sampson	0	0	0	0	0	0	0	0	0	0	0	0,823	0	0	0	0,177	0	0	0	0
Gregory	0	0	0	0	0	0	0	0	0	0,015	0	0	0,914	0	0	0,071	0	0	0	0
Balthasar	0	0	0	0	0	0	0	0	0	0	0	0	0	0,914	0	0	0	0	0,010	0,076
Montague	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,000	0	0	0	0	0
Abram	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,000	0	0	0	0
Lady Montague	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,000	0	0	0
Apothecary	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,000	0	0
Peter	0	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Servant	0	0	0,113	0	0	0	0	0	0,193	0,437	0	0	0	0	0,081	0	0	0,176	0	0
Friar John	0	0	0	0,744	0	0	0	0	0	0	0	0	0	0	0	0	0,256	0	0	0

réplique d'ailleurs sur les autres paires d'œuvres : à part de petites exceptions comme le personnage d'Aslan dans *Narnia* qui se répartit sur différents personnages si l'on n'encourage pas sa correspondance avec son homologue du film, presque tous les transports optimaux basés sur les poids d'apparitions découlent vers des transformations identiques [30].

En regardant les valeurs du tableau 8.8, on remarque en outre que la répartition est plutôt claire, organisée en majeure partie autour de la diagonale, avec assez peu de personnages de la pièce qui se retrouvent distribués sur plusieurs personnages du film. Parmi les personnages les plus fréquents, l'exception la plus notable est celle de la Nurse (à la sixième ligne). Mais la colonne correspondante montre bien qu'il s'agit essentiellement d'une différence d'importance de ce personnage au sein des deux œuvres : la NURSE du film apparaît deux fois moins souvent que celle de la pièce, les presque 50 % restants de la « quantité » initiale de Nurse sont donc forcés de combler des trous ailleurs (par exemple, chez ROMEO, qui a un nombre d'apparitions plus important dans le film et n'est donc pas totalement couvert par le Romeo de la pièce de théâtre), sans qu'il faille pour autant imaginer que l'algorithme voit un rapprochement majeur entre la Nurse et ROMEO sur un plan narratif.

Les personnages qui n'existent que dans l'une des deux œuvres, sans grande surprise, se retrouvent identifiés par des mélanges assez insolites : ainsi, SUSAN serait composée à la fois de Friar John, Paris et Nurse, alors que COP serait une fusion de Nurse et Balthasar. Dans l'autre sens, les personnages qui n'existent que dans la pièce (comme Servant et Friar John) ne trouvent pas de correspondant direct dans le film. Ils sont alors répartis sur plusieurs personnages, comme pour compléter les présences de ces derniers. Par exemple, même si Mercutio (de la pièce) est trans-

[30] Ces résultats sont à trouver dans le tableau des adaptations sur le répertoire character_network pour davantage d'informations.

formé intégralement en MERCUTIO (du film), il ne « suffit » pas à occuper toutes ses apparitions, et les petits personnages comme Servant et Peter permettent de combler ce manque. À ce stade, il semble clair que les petits pourcentages témoignent plutôt d'un écart d'importance d'une œuvre à l'autre que d'une réelle correspondance entre les personnages qu'ils représentent. Il faudrait donc surtout retenir de ce tableau que le transport général de la pièce de théâtre au film se fait de manière plutôt évidente lorsque l'on s'appuie sur des poids d'apparitions.

En revanche, les tableaux 8.9 et 8.10 permettent des observations plus fines. On peut immédiatement remarquer une diagonale un peu moins nette, avec d'autres personnages qui émergent : avec les poids d'interactions, Nurse a une importance bien plus similaire à celle de NURSE, mais c'est par contre Capulet qui est deux fois plus présent dans la pièce de théâtre que dans le film (et qui doit donc se distribuer, là aussi, en fonction des trous à combler). On pourrait là aussi regarder en détail chaque phénomène local, mais ce qui semble le plus porteur de sens, en matière d'interprétation littéraire, c'est l'analyse des *différences* entre les deux tableaux, c'est-à-dire les changements opérés selon que l'on rectifie ou non la diagonale.

En réalité, ces différences sont peu nombreuses ; il n'y a que trois personnages qui présentent des modifications dans leur répartition : Escalus, Balthasar et Montague. Les trois rôles existent également dans le film (CAPTAIN PRINCE, BALTHASAR et MONTAGUE). Dans le cas d'Escalus et de Montague, on est à nouveau dans une situation où leur correspondant du film est moins important : ils le remplissent en entier et distribuent ensuite la quantité restante aux endroits les plus appropriés, qui peuvent effectivement varier selon les informations que l'on donne à l'algorithme, sans pour autant avoir un effet notable sur l'interprétation du tableau. Mais le cas de Balthasar est plus intéressant : lorsque l'on prend la version rectifiée (et que l'on indique donc la correspondance entre Balthasar et BALTHASAR à l'algorithme),

TABLE 8.9 Transformation des personnages de Shakespeare en ceux de Luhrmann, avec poids d'interactions.

OT	ROMEO	JULIET	MERCUTIO	BENVOLIO	FATHER LAURENCE	NURSE	CAPULET	GLORIA	TYBALT	CAPTAIN PRINCE	DAVE	SAMPSON	GREGORY	BALTHASAR	MONTAGUE	ABRA	SUSAN	CAROLINE	APOTHECARY	COP
Romeo	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Juliet	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mercutio	0	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Benvolio	0	0	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Friar Laurence	0	0	0	0	0,946	0	0	0	0	0	0	0	0	0,054	0	0	0	0	0	0
Nurse	0	0,156	0	0	0	0,844	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Capulet	0,401	0	0,101	0	0	0	0,445	0,010	0	0	0,016	0	0	0,027	0	0	0	0	0	0
Lady Capulet	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0
Tybalt	0	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0	0	0,085	0	0	0,154
Escalus	0	0	0	0	0	0	0	0	0,334	0,427	0	0	0	0	0	0	0	0	0	0
Paris	0	0	0	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0	0	0	0
Sampson	0	0	0	0	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0	0	0
Gregory	0	0	0	0	0	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0	0
Balthasar	0	0	0	0	0	0	0	0	0	0	0	0	0	0,633	0	0	0	0	0,367	0
Montague	0	0	0	0	0	0	0	0	0	0	0	0,004	0	0,066	0,683	0,144	0	0	0	0,103
Abram	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,000	0	0	0	0
Lady Montague	0	0	0	0	0	0	0	0	0	0	0	0,360	0,230	0	0	0	0	0,410	0	0
Apothecary	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,000	0
Peter	0	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Servant	0	0,330	0	0,493	0	0	0	0	0	0	0	0	0	0,177	0	0	0	0	0	0
Friar John	0	0	0	0	0	0	0	0	0	0	0	0	0	0,000	0	0	0	0	0	0

TABLE 8.10 Transformation des personnages de Shakespeare en ceux de Luhrmann, avec poids d'interactions et diagonale rectifiée.

OT_0	ROMEO	JULIET	MERCUTIO	BENVOLIO	FATHER LAURENCE	NURSE	CAPULET	GLORIA	TYBALT	CAPTAIN PRINCE	DAVE	SAMPSON	GREGORY	BALTHASAR	MONTAGUE	ABRA	SUSAN	CAROLINE	APOTHECARY	COP
Romeo	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Juliet	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mercutio	0	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Benvolio	0	0	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Friar Laurence	0	0	0	0	0,946	0	0	0	0	0	0	0	0	0,054	0	0	0	0	0	0
Nurse	0	0,156	0	0	0	0,844	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Capulet	0,401	0	0,101	0	0	0	0,445	0,010	0	0	0,016	0	0	0,027	0	0	0	0	0	0
Lady Capulet	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0
Tybalt	0	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0	0	0	0	0	0
Escalus	0	0	0	0	0	0	0	0	0,334	0,427	0	0	0	0	0	0	0,085	0	0,046	0,108
Paris	0	0	0	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0	0	0	0
Sampson	0	0	0	0	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0	0	0
Gregory	0	0	0	0	0	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0	0
Balthasar	0	0	0	0	0	0	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0
Montague	0	0	0	0	0	0	0	0	0	0	0	0,004	0	0,020	0,683	0,144	0	0	0	0,148
Abram	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,000	0	0	0	0
Lady Montague	0	0	0	0	0	0	0	0	0	0	0	0,340	0,230	0	0	0	0	0,410	0	0
Apothecary	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,000	0
Peter	0	0	1,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Servant	0	0,330	0	0,493	0	0	0	0	0	0	0	0	0	0,177	0	0	0	0	0	0
Friar John	0	0	0	0	0	0	0	0	0	0	0	0	0	1,000	0	0	0	0	0	0

il se transporte intégralement sur son homonyme du film. Mais si l'on enlève cette contrainte, il se répartit entre BALTHASAR et APOTHECARY, ce qui donne à penser que l'association Balthasar/BALTHASAR est moins évidente pour le modèle et qu'un coup de pouce est nécessaire pour encourager ce transport de personnage.

De manière générale, les informations que l'on peut tirer de l'observation détaillée du transport de personnages restent plutôt anecdotiques. Pour en extraire des points d'intérêt, il faut veiller au poids respectif de chaque paire de personnages afin de ne pas se laisser troubler inutilement par des distributions qui découlent surtout de la présence plus importante du personnage d'origine (forcé de distribuer le reste de son poids sur d'autres personnages). Influencée parfois par les personnages qui n'existent que dans l'une des œuvres, souvent par ceux dont le poids est très différent d'une œuvre à l'autre, la distribution du transport optimal peut en tout cas permettre de se rassurer sur la qualité de l'algorithme (lorsque la diagonale est, comme ici, nettement marquée), et parfois mettre l'accent sur un personnage en particulier dont la correspondance n'est pas si évidente, comme c'est le cas pour Balthasar dans cet exemple.

Quatrième partie

Conclusion et ouverture

9 Continuer l'exploration

9.1 Rappel de la problématique

Ce travail s'ancrait à l'origine dans plusieurs démarches, toutes liées aux approches de *distant reading* et en particulier aux réseaux de personnages. Partant d'une ambition globale d'adapter cet outil mathématique à des terrains qui lui étaient encore inaccessibles, la recherche s'est articulée en plusieurs axes.

Un premier axe préliminaire et indispensable était l'exploration des possibilités d'automatisation dans la construction d'un réseau de personnages basé sur des sources textuelles. Afin de rendre possible la génération de réseaux nécessaires à l'exemplification et à l'expérimentation des deux autres axes, il fallait d'abord s'intéresser aux méthodes existantes pour faciliter et accélérer chaque étape d'un processus qui n'aurait pu se faire à la main dans le cadre temporel de ce travail.

Après ces étapes préparatoires, un moteur important de cette recherche était d'étendre l'usage des réseaux de personnages à un catalogue plus large de types d'œuvres. Jusqu'alors abondamment employé pour des narrations « linéaires » (films, romans, pièces de théâtre, etc.), le réseau (comme la plupart des outils de *distant reading*) n'était pas capable de capturer la dimension interactive de récits tels que les livres dont vous êtes le héros ou les

jeux vidéo, car sa structure ne permettait pas de tenir compte des différentes options possibles (et donc du caractère incertain ou variable de certaines interactions en fonction des choix à disposition). Pour construire des réseaux de personnages qui rendent justice à cet aspect interactif, il fallait donc imaginer un modèle théorique intégrant cette notion de choix, mais aussi trouver des sources textuelles adaptées pour mettre ce modèle en pratique.

Enfin, le dernier axe reposait sur l'idée d'utiliser le réseau de personnages comme moyen de comparaison entre deux œuvres de supports médiatiques différents. Il peut être difficile de construire un langage commun pour mettre en perspective un livre et son adaptation en film, ou une pièce de théâtre et sa déclinaison en opéra. Par sa structure mathématique définie de manière stable sur tous les types d'œuvres compatibles, le réseau de personnages offrait donc des pistes alléchantes pour s'émanciper des contraintes inhérentes à l'hétérogénéité des médias étudiés et proposer une mise en perspective des œuvres à travers la comparaison de leurs réseaux de personnages respectifs. Or, la comparaison de réseaux est un problème mathématique vaste et encore en exploration. Il fallait donc identifier les approches appropriées pour ce cas d'étude précis et trouver une manière d'interpréter les résultats obtenus qui puisse avoir du sens pour les données choisies.

9.2 Résultats principaux

En réponse à cette problématique, un premier résultat de cette recherche a été le développement de Charnetto (voir l'annexe B), un module codé en Python et disponible en libre accès pour la génération automatisée de réseaux de personnages en partant de différents types de sources textuelles. Fruit de plusieurs mois de travail, ce module s'est avéré précieux pour les milliers de réseaux générés dans les étapes suivantes. Il peut s'utiliser de plusieurs manières en fonction du type de données à traiter : sur des

romans, il appelle au choix entre Flair ou spaCy (deux algorithmes de reconnaissance des entités nommées) pour détecter automatiquement les occurrences de personnages ; sur des scripts de films suivant le format présent sur IMSDB, il repère les personnages grâce à un code sur mesure ; pour tout autre type de données, il reconnaît une syntaxe fournie dans la documentation de Charnetto et récupère les personnages qui y sont mis en évidence (par une annotation manuelle ou tout autre processus de reconnaissance externe). Dans tous les cas, il génère ensuite une liste de personnages qui peut être modifiée à la main (pour rectifier les erreurs de regroupement des alias possibles d'un même personnage) et s'appuie sur cette liste corrigée pour générer un réseau de personnages basé sur les cooccurrences (selon des paramètres et des seuils qui peuvent être ajustés en fonction des besoins et de la nature des analyses).

La recherche s'est ensuite portée sur l'intégration de réseaux de personnages dans le cas de narrations interactives, avec l'élaboration d'un modèle théorique intitulé « flux narratif » (voir la section 5.1) servant d'étape préliminaire à l'usage d'outils comme Charnetto pour construire les réseaux eux-mêmes. Ce flux narratif est une schématisation de la structure narrative de l'œuvre étudiée, tenant compte de ses embranchements et intégrant toutes les données nécessaires aux analyses prévues, qui s'appuie sur la théorie mathématique des réseaux de flot. Le flux peut ensuite être parcouru du début à la fin, formant une « traversée » qui représente une manière d'expérimenter l'œuvre (en ayant opéré des choix à chaque fois que la narration le réclamait). Chaque traversée donne ainsi lieu à une version « linéaire » de l'œuvre et permet de produire un réseau de personnages propre à cette expérience. Il est ainsi possible de générer de nombreux réseaux de personnages pour une même œuvre et la dernière partie du chapitre 5 propose une façon de regrouper ces réseaux en un seul « réseau moyen », témoignant de la variation (absolue et

relative) des poids des nœuds et des arêtes à travers tous les réseaux d'origine.

Pour illustrer ce modèle théorique, le chapitre 6 présente un exemple pratique basé sur quatre jeux vidéo de la franchise *Life is Strange*. Un premier résultat de ce chapitre est la mise à disposition de données nettoyées et structurées, basées sur les scripts rédigés par des communautés de joueurs et de joueuses et disponibles en libre accès pour de futures recherches. Ces données rassemblent les occurrences des différents personnages ainsi que la structure narrative des jeux, avec des marqueurs témoignant des différentes options de dialogue et leurs conséquences pour la narration. Ces données ont ensuite été transformées en un flux narratif pour chacun des quatre jeux (le code permettant de transformer ces données en un flux narratif lié à ce processus est également disponible publiquement pour permettre une réplication des résultats). Enfin, les explorations de ces différents flux et des centaines de milliers de réseaux de personnages générés automatiquement à partir de chacun d'entre eux donnent lieu à des pistes de réflexion quant à l'importance des choix laissés aux joueurs et joueuses sur leur expérience de jeu et sur la structure générale des scénarios de cette franchise.

Dans le dernier axe de recherche, les apports de ce travail sont également à distinguer entre la partie théorique et la mise en pratique. Sur le plan théorique, les mesures mathématiques sélectionnées pour la comparaison de réseaux de personnages ont dû être adaptées à ce contexte précis. En particulier, le coefficient RV, jusqu'alors prévu pour des nœuds de poids identique entre les deux réseaux, a fait l'objet d'ajustements pour devenir calculable dans le cas de réseaux de personnages qui admettent presque systématiquement un poids différent pour un même nœud d'une œuvre à l'autre. Plus largement, le chapitre 7 propose différentes mesures choisies et ajustées pour former une sorte de boîte à outils de comparaison d'œuvres cross-médiatiques.

Enfin, la mise en application de ces mesures à différentes adaptations a donné lieu à plusieurs résultats dans le chapitre 8. Comme dans le cas de *Life is Strange*, une base de données a été rendue disponible, regroupant dix paires d'œuvres livre/film ainsi qu'un triplet théâtre/opéra/film, et incluant les listes de personnages de chaque œuvre ainsi que les occurrences de ces personnages, de sorte à faciliter la création de réseaux de personnages pour des recherches ultérieures. De plus, l'étude réalisée sur ces différentes adaptations a révélé l'utilité des mesures précitées, dans la complémentarité des éléments qu'elles mettent en évidence et l'interprétation fine qu'elles permettent d'obtenir sur la similarité entre chaque adaptation et son œuvre d'origine. Il est notamment possible d'identifier trois groupes d'adaptations en partant des mesures propres à chaque réseau : les adaptations très similaires au livre d'origine, celles qui présentent un changement conséquent de taille en gardant tout de même l'essentiel des structures principales, et celles qui marquent un changement important de structure par rapport à l'œuvre initiale. Les différents tableaux regroupant les valeurs de ces mesures offrent en outre une base de comparaison précieuse pour de futures recherches sur des œuvres additionnelles, en permettant la mise en perspective des mesures obtenues avec celles de ce travail (afin de pouvoir rattacher ces nouvelles œuvres au comportement de l'une ou l'autre des dix adaptations étudiées). Le chapitre 8 offre également une évaluation de la qualité de ces mesures par rapport à ce qui est attendu pour certaines œuvres très clairement polarisées en matière de similarité, ainsi qu'une étude plus approfondie du triplet vérifiant les hypothèses d'interprétation des premiers résultats.

9.3 Limites et périmètre du travail

Le problème le plus important durant l'élaboration de ce travail a été l'accès à des données pertinentes et de bonne qualité. Si l'on

trouve aujourd'hui de bonnes sources pour des romans libres de droits ou des scripts de film, la quête de sources textuelles témoignant des interactions entre des personnages de jeux vidéo et gardant une trace des options laissées aux joueurs et joueuses est bien plus ardue. De telles données sont rarement partagées avec le public, et si certaines communautés de passionnées et passionnés produisent des scripts pour des jeux à grand succès, il n'existe pas de syntaxe standardisée qui rendrait tous ces scripts faciles à parcourir automatiquement. La création de code sur mesure pour interpréter correctement chaque script prend donc un temps considérable. De plus, le souhait de travailler avec des œuvres cross-médiatiques limite également le nombre d'œuvres éligibles, car l'intersection des bases de données de livres et de films n'offre que peu de candidats. C'est la raison pour laquelle il n'a pas été possible, dans le périmètre de ce travail, d'intégrer des jeux vidéo en comparaison avec d'autres médias dans le chapitre 8, bien que l'envie de regrouper les précédents résultats avec une étude cross-médiatique intégrant des réseaux de personnages tirés de narrations interactives était présente.

Il a également fallu faire des choix lors de l'élaboration du code de Charnetto, pour obtenir en quelques mois un outil suffisant pour aborder les questions de recherches principales du projet. Le souhait initial de produire automatiquement des réseaux conversationnels (basés donc sur les dialogues) a été revu à la baisse devant l'ampleur de la question sur un plan informatique, puisque la détection automatique de dialogues et d'interlocuteurs dans le cas de romans s'appuie aujourd'hui sur des stratégies complexes (voir par exemple l'article de Ek et Wirén [2019] qui ouvrent la voie de la distinction automatique des passages narratifs et discursifs dans la fiction, et le chapitre de Weimer *et al.* [2024] sur l'attribution des dialogues). Il a donc fallu se baser sur des réseaux de cooccurrences, avec les limites que ce choix présente en matière de pertinence des résultats (les cooccurren-

ces n'impliquant pas toujours de vraies interactions dans la narration). De la même manière, il a été décidé de détecter uniquement les entités nommées, sans étendre les capacités de Charnetto aux pronoms ou aux anaphores : le gain de qualité des résultats aurait probablement été considérable, mais là aussi, les défis techniques se sont avérés trop ambitieux pour la nature du projet, qui ne prétendait pas résoudre des questions ouvertes en sciences computationnelles pures (l'attribution correcte des pronoms aux personnages auxquels ils se réfèrent n'est, là non plus, pas une question triviale). L'effet de cette concession est particulièrement visible dans les exemples pratiques du chapitre 8, dans les réseaux de personnages des romans dont la narration est à la première personne et pour lesquels on constate une trop faible importance du personnage principal par rapport à l'adaptation en film. Toutefois, les limites imposées par le périmètre du projet dans le cas de Charnetto n'ont pas empêché d'obtenir des résultats intéressants et interprétables pour servir d'illustration aux approches théoriques des deuxième et troisième parties, ce qui était le principal objectif de ce module.

Plus généralement, les illustrations pratiques restent également limitées par le temps qui leur était réservé dans le calendrier du projet et par leur nature même d'exemples. Les chapitres 6 et 8 auraient pu faire l'objet de thèses complètes, ce qui aurait permis une étude bien plus approfondie des œuvres concernées et donc des résultats plus complets. Il aurait également été appréciable de travailler en collaboration avec des experts et expertes en *game studies* et en *adaptation studies* notamment, pour offrir une interprétation plus riche et contextualisée des valeurs obtenues. En l'état, il faut voir ces chapitres comme une invitation à de plus amples recherches, avec l'espoir que la curiosité de personnes liées à ces domaines soit éveillée par ces premières explorations.

9.4 Perspectives

Comme la section précédente le suggère déjà, de nombreuses pistes sont envisageables pour prolonger le travail déjà produit. Sur le plan technique, un développement plus ambitieux pourrait permettre à Charnetto de proposer une palette de fonctionnalités bien plus large, même si d'autres modules centrés sur la génération automatique de réseaux de personnages ont vu le jour récemment, comme Renard (2024) qui promet de belles performances (ainsi que la possibilité de construire des réseaux dynamiques).

Les perspectives qui concernent les jeux vidéo et plus globalement les narrations interactives sont bien plus prometteuses, dans la mesure où les territoires de ces recherches restent encore peu explorés. Le modèle du flux narratif demande à être éprouvé, sur des données aussi variées que possible, pour en dessiner les limites et les améliorations à envisager. Il serait également précieux d'établir un standard d'encodage pour les fictions interactives, afin de pouvoir y appliquer un flux narratif, ainsi que de travailler à assouplir la structure du flux narratif pour le rendre plus accessible de façon standardisée, et le décliner plus aisément à d'autres types d'analyses que les réseaux de personnages. En ce qui concerne le *distant reading* appliqué à des narrations interactives, si le flux narratif s'avère robuste et exploitable, de nombreuses pistes semblent intéressantes pour observer la manière dont ces œuvres s'articulent et proposent des histoires différentes selon les choix : variation du vocabulaire rencontré, importance des personnages, évolution dans la nature de leurs relations, fréquence d'apparition de certains mots, etc. Tous ces axes devraient permettre d'aborder la narration interactive sous un angle plus littéraire, offrant ainsi un regard complémentaire aux études existantes sur l'importance des choix sur le plan psychologique ou sociologique.

Pour le cas particulier des réseaux de personnages dans les narrations interactives, là aussi, de nombreuses pistes d'approfondissement sont envisageables. Sur un plan théorique, la synthèse des informations contenues dans la multitude de réseaux découlant des traversées peut se faire de plusieurs manières : ce travail présente l'option du réseau moyen, mais on pourrait aider beaucoup d'autres stratégies, par exemple du *clustering* de réseaux pour identifier des groupes de traversées qui produisent un réseau globalement identique, et ainsi compter le nombre de scénarios sensiblement différents pour une même œuvre. Un tel outil pourrait aider les *narrative designers* dans leur travail. L'exemple particulier de *Life is Strange* invite lui aussi à des analyses plus fines, prenant en compte la dépendance de certains choix sur les options ultérieures, s'appuyant peut-être sur les répliques pour observer la nature des interactions (en plus de leur existence) et plaçant ces résultats dans une perspective plus large sur l'accueil général de ces jeux par le public, la liberté perçue par les joueurs et les joueuses ou d'autres angles mêlant *game studies* et psychologie, sociologie ou histoire des médias, entre autres axes possibles.

Enfin, la question des comparaisons de réseaux de personnages dans le cas de l'étude d'adaptations bénéficierait également d'un jeu de données plus grand, permettant de vérifier ou d'infirmer les hypothèses d'interprétation présentées dans ce travail pour les différentes mesures étudiées. Le dialogue avec des spécialistes des *adaptation studies*, ou du moins des médias représentés dans ces données, enrichirait également le discours général et permettrait en outre de voir si de telles méthodes trouvent un écho dans les travaux existants et captent l'intérêt des domaines de sciences humaines qui travaillent actuellement sur de tels objets. Là encore, les ouvertures semblent trop nombreuses pour en faire une liste précise, mais la réplication des résultats sur une base de données de centaines ou de milliers de paires livres/films permettrait de dégager des tendances générales, d'identifier des

paires singulières (qui se distingueraient de la masse), ou de visualiser ces œuvres sur des graphes avec différents axes de similarité, d'une part pour mieux comprendre les finesses de ces mesures, d'autre part pour en tirer des observations sur les œuvres étudiées. En outre, un élargissement aux œuvres trans-médiatiques (c'est-à-dire aux univers narratifs déployés sur plusieurs médias, avec des scénarios différents à chaque fois) pourrait également produire des résultats intéressants sur l'importance de personnages récurrents entre les œuvres et les supports médiatiques, ou sur l'évolution générale de la narration à travers les médias utilisés. Les scénaristes qui adaptent un roman au cinéma pourraient également se servir du réseau de personnages du livre pour observer les changements dans la dynamique de groupe en retirant un personnage, ou en fusionnant deux rôles, et ainsi veiller à ce que les interactions cruciales soient préservées.

Plus largement, notre espoir est d'avoir proposé avec ce travail une direction pour un dialogue bilatéral entre sciences et lettres, une forme d'interdisciplinarité qui respecte à la fois le formalisme des mathématiques et les objets de sciences humaines en essayant au mieux de ne dénaturer ni l'un ni les autres. La perspective la plus enthousiasmante, en ce sens, serait de voir des chercheurs et des chercheuses des différentes traditions approcher ce travail avec leur propre bagage, trouver de l'intérêt dans les résultats présentés et prolonger les investigations dans la direction qui leur correspond, pour étendre et consolider sa nature de passerelle entre ces disciplines.

Cinquième partie

Annexes

A Flair ou spaCy ?

Si Flair semble être le choix idéal pour détecter automatiquement les personnages (comme présenté à la section 4.2), il a toutefois un défaut : son exécution est coûteuse en ressources et traiter un roman entier demande un temps considérable sur une machine faiblement équipée en CPU et en GPU. J'ai donc décidé de comparer ses performances avec un autre algorithme très populaire et bien plus léger : celui de spaCy (Honnibal et Montani, 2017), qui ne figure pas dans l'article de Stanislawek *et al.* (2019).

Pour les besoins de cette comparaison, j'ai annoté les 100 premiers paragraphes de huit romans, sélectionnés selon leurs particularités narratives et dans l'idée de proposer un échantillon varié au niveau des genres et des défis posés aux modèles de NER. La liste complète des romans est dans le tableau A.1.

Trois de ces romans sont en français : *Comment le dire à la nuit* (Vincent Tassy, 2018), écrit à la troisième personne sous forme de roman choral avec des prénoms anciens (comme Egmont ou Athalie), *Que passe l'hiver* (David Bry, 2017), se déroulant dans un univers imaginaire dans lequel la plupart des noms sont inventés et présentant beaucoup de personnages, et *L'apprenti assassin* (Robin Hobb, 1998), écrit à la première personne et dans lequel

Titre	Langue	Narration	Genre	Exemples de personnages
Comment le dire à la nuit	FR	3e pers.	fantasy gothique	Egmont, Athalie, Adriel
Que passe l'hiver	FR	3e pers.	fantasy nordique	Stig, Umbre, Ewald
L'apprenti assassin	FR	1re pers.	fantasy médiévale	Prudence, Subtil, Royal
Assassin's apprentice	EN	1re pers.	fantasy médiévale	Caution, Shrewd, Regal
The Terror	EN	3e pers.	fantastique historique	John Franklin, Francis Crozier, Graham Gore
A Swift Pure Cry	EN	1re pers.	contemporain	Shell, Declan, Father Rose
The Nightingale Girls	EN	3e pers.	historique	Helen, Dora, Millie
The Green Mile	EN	1re pers.	fantastique	Paul Edgecombe, Del, Billy the Kid

TABLE A.1 Description des données pour la comparaison entre Flair et spaCy.

presque tous les prénoms sont tirés de mots du lexique commun (comme Royal ou Prudence).

La version anglaise de ce dernier, *Assassin's Apprentice* (Robin Hobb, 1995), a aussi été annotée pour voir la différence de performance entre les deux langues. Pour les autres romans anglophones, on trouve dans le jeu de données *The Terror* (Dan Simmons, 2007), basé sur des faits historiques, qui contient une grande quantité de personnages (dont beaucoup s'appellent John, James ou Thomas), *A Swift Pure Cry* (Siobhan Dowd, 2006), qui se passe en Irlande avec une poignée de personnages différents, *The Nightingale Girls* (Donna Douglas, 2012), construit comme un roman choral historique et enfin *The Green Mile* (Stephen King, 1996), qui contient beaucoup de surnoms différents pour les mêmes personnages.

Les annotations manuelles se concentrent sur quatre catégories d'entités nommées : personnages, lieux, organisations et autres. Ce découpage correspond à celui de Flair, et un tableau de correspondance a été mis en place avec les catégories données par spaCy pour uniformiser les résultats. Pour l'anglais, les modèles utilisés sont `ner-english-large` (Schweter et Akbik, 2020) pour Flair et `en_core_web_trf` [31] pour spaCy ; pour le français, Flair

31 https://github.com/explosion/spacy-models/releases/tag/en_core_web_trf-3.7.3, consulté le 11.06.2024.

utilise `ner-french` (Akbik *et al.*, 2018) et spaCy utilise `fr_core_news_lg`[32].

Pour chaque modèle, on construit une matrice de confusion indiquant la répartition des prédictions entre toutes les catégories d'entités nommées par rapport à ce qui était attendu (en référence aux annotations manuelles). Plusieurs mesures d'évaluation ont été utilisées pour observer les différences entre les deux modèles pour chaque langue. Les mesures les plus classiques sont la précision et le rappel : la précision est la proportion d'entités nommées correctement catégorisées sur l'ensemble des entités nommées *détectées* (ce qui atteste donc de la qualité des résultats), et le rappel est la proportion d'entités nommées correctement catégorisées sur l'ensemble des entités nommées *à détecter* (ce qui atteste de la quantité de résultats obtenus). Afin de distinguer les erreurs de détection et les erreurs de catégorisation des entités, j'ai calculé deux précisions et rappels différents, de la manière suivante :

— La matrice de confusion est d'abord transformée en une matrice 3x3 simplifiée (voir figure A.1) qui réunit les catégories LOC, ORG et MISC sous l'étiquette NPER (pour non-personnage).
— Ensuite, pour calculer la précision et le rappel de *détection* (préfixés *d_*), on réunit PER et NPER pour former une matrice 2x2 décrivant la capacité des modèles à détecter correctement les entités, sans distinction des catégories.
— Enfin, la précision et le rappel de *catégorisation* (préfixés *c_*) se basent sur la matrice 2x2 obtenue en enlevant NONE, pour observer les performances des modèles uniquement sur l'attribution des catégories pour les entités détectées.

En outre, les matrices de confusion des deux modèles nous permettent également de calculer le kappa de Cohen, défini

[32] https://github.com/explosion/spacy-models/releases/tag/fr_core_news_lg-3.7.0, consulté le 11.06.2024.

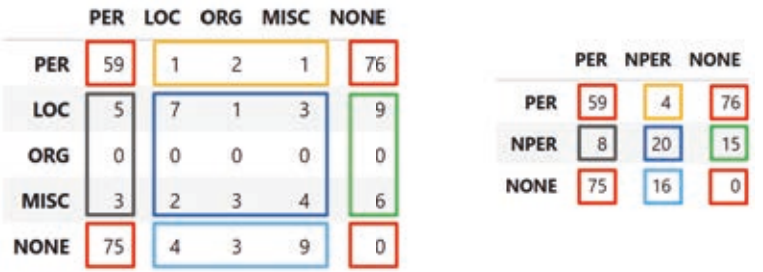


FIGURE A.1 À gauche, matrice de confusion pour les 100 premiers paragraphes de *Assassin's Apprentice*. À droite, matrice simplifiée en agrégeant les catégories qui ne sont pas des personnages sous l'étiquette NPER.

comme

$$\kappa := \frac{Pr(a) - Pr(e)}{1 - Pr(e)},$$

où $Pr(a)$ est la « proportion d'accord », c'est-à-dire le nombre de prédictions correctes (diagonale de la matrice de confusion) par rapport à l'ensemble des observations (somme de la matrice), et $Pr(e)$ est la probabilité d'un accord au hasard, donc la probabilité que les prédictions tombent juste « accidentellement » (pour chaque ligne, cette valeur correspond à la probabilité que les prédictions et les annotations correspondent, on multiplie donc la somme de la ligne et la somme de la colonne, on divise le résultat par le nombre total d'observations au carré, puis on additionne toutes ces valeurs pour obtenir un score global d'accord au hasard). Le kappa de Cohen indique ainsi l'accord entre le modèle et la réalité, avec un score qui va de 0 (aucun accord entre les deux au-delà du hasard) à 1 (accord parfait).

Or, dans notre cas, la détection correcte de lieux ou d'organisations est moins importante que celle des personnages : en cas de scores très proches entre les deux modèles, il peut donc être intéressant de les départager à l'aide d'une mesure calibrée spécifiquement sur la capacité à détecter et catégoriser correctement les personnages. Pour ce faire, on choisit de « filtrer » le kappa de

Prédiction

Vérité	PER	LOC	ORG	MISC	NONE
PER	126	1	0	3	3
LOC	3	25	0	8	3
ORG	1	0	1	0	0
MISC	8	4	1	9	4
NONE	1	1	0	2	0

Prédiction

Vérité	PER	LOC	ORG	MISC	NONE
PER	126	1	0	3	3
LOC	3	36	0	0	0
ORG	1	0	1	0	0
MISC	8	0	0	18	0
NONE	1	0	0	0	3

FIGURE A.2 À gauche, matrice de confusion pour les 100 premiers paragraphes de *The Green Mile*, avec Flair. À droite, matrice filtrée sur la détection de personnages.

Cohen sur la détection de personnages, en simulant des scores de détection parfaits pour les autres catégories, comme illustré dans la matrice de droite de la figure A.2 (les valeurs contenues dans le cadre sont rassemblées sur la diagonale). Notons qu'il aurait également été possible de réduire les matrices de départ en des matrices 2x2 pour ne garder que la détection ou non des personnages, mais il me semblait utile de garder la répartition entre les différentes catégories, pour préserver la granularité des erreurs de catégorisation et pouvoir observer d'éventuels phénomènes récurrents (confusion systématique entre personnages et lieux dans certains romans, par exemple).

Les résultats de toutes ces mesures sont reportés dans le tableau A.2. Comme on peut le voir, les performances sont globalement meilleures pour Flair, autant en français qu'en anglais. Les modèles réagissent mieux en anglais dans l'ensemble, et à part quelques valeurs très basses comme *L'apprenti assassin* qui pose beaucoup de difficultés aux deux modèles (certainement à cause des noms de personnages tirés de noms communs), les scores sont globalement plutôt bons. Une étude plus détaillée des erreurs précises pour chaque modèle dans les extraits des huit romans montre en outre que Flair a tendance à rater quelques occurrences de personnages, là où spaCy passe totalement à côté de certains noms, ce qui fait également pencher la balance du

Titre	Modèle	kappa	kappa (PER)	d_précision	d_rappel	c_précision	c_rappel
Comment le dire à la nuit	flair	0,28	**0,55**	**1**	0,75	0,32	**0,64**
	spacy	**0,39**	0,49	0,8	**0,89**	**0,59**	0,58
Que passe l'hiver	flair	**0,3**	**0,76**	**0,94**	**0,69**	0,40	**0,67**
	spacy	0,24	0,48	**0,94**	0,68	**0,55**	0,64
L'apprenti assassin	flair	**0,2**	**0,51**	**0,93**	**0,70**	0,32	**0,61**
	spacy	0,01	0,26	0,71	0,56	**0,60**	0,21
Assassin's Apprentice	flair	**0,63**	**0,71**	**0,88**	**0,94**	0,5	**0,5**
	spacy	0,41	0,57	0,85	0,79	**0,55**	0,38
The Terror	flair	**0,56**	0,88	0,9	**0,95**	**0,57**	**0,62**
	spacy	0,4	**0,89**	**1,0**	0,7	0,55	**0,62**
A Swift Pure Cry	flair	**0,55**	**0,72**	**0,9**	**0,78**	**0,43**	**0,24**
	spacy	0,44	0,54	0,86	0,75	0,41	0,17
The Nightingale Girls	flair	**0,77**	**0,91**	**1,0**	**0,9**	**0,6**	**0,65**
	spacy	0,54	0,83	0,96	0,96	**0,6**	0,53
The Green Mile	flair	**0,59**	**0,81**	**0,95**	**0,98**	**0,91**	**0,97**
	spacy	0,39	0,73	0,81	0,83	0,89	0,93

TABLE A.2 Comparaison des scores entre Flair et spaCy pour les 8 romans annotés.

côté de Flair : il est toujours possible de faire une récupération automatique des occurrences lorsque le nom est présent dans la liste des personnages et qu'il manque certaines de ses apparitions dans le texte, mais il est bien plus difficile d'identifier un personnage qui aurait totalement échappé au modèle. C'est par exemple le cas dans *The Green Mile* de Stephen King, avec le personnage de Brutus qui est parfois surnommé « Brutal » : Flair capture la plupart des occurrences de ce surnom, alors que spaCy ne l'identifie pas comme un nom propre.

En conclusion, avec un matériel suffisamment performant, Flair est un choix plus intéressant pour capturer des entités nommées de meilleure qualité. Toutefois, l'alternative beaucoup plus accessible de spaCy offre des résultats satisfaisants si l'on utilise une machine plus modeste. Dans le cadre de ce travail, les réseaux de personnages basés sur des romans seront générés avec Flair, mais le module présenté au chapitre suivant, Charnetto, a été pensé pour fonctionner avec les deux modèles, dans l'idée d'autoriser les deux scénarios en fonction des contraintes et des besoins de chaque projet.

B Charnetto

Dans l'optique de faciliter la génération de réseaux et d'avoir la main sur les différentes parties du processus, la première étape pratique de cette recherche a été de construire un module en Python (Van Rossum et Drake, 2009) prenant en entrée un texte de fiction et produisant en sortie un réseau de personnages enrichi de plusieurs types de métadonnées. Dans cette section, nous allons examiner en détail chaque partie de ce module pour être en mesure d'exploiter au mieux les réseaux ainsi extraits, en comprenant la façon dont ils ont été conçus.

Le module en lui-même, appelé « Charnetto », n'a pas pour ambition de révolutionner les méthodes computationnelles actuelles. Il s'appuie sur des algorithmes existants et son intérêt est surtout de rendre accessible la génération de réseaux de personnages à un public ayant quelques bases de programmation, en limitant autant que possible le code à produire. Il est publié sous licence open source et accessible sur la plateforme Pypi [33].

D'autres outils d'automatisation ont été élaborés ces dernières années et permettent également de traiter automatiquement des

33 https://pypi.org/, consulté le 20.09.2024.

textes pour en extraire des réseaux de personnages ou des analyses plus générales de traitement du langage naturel. En Python, on peut notamment mentionner BookNLP (Bamman, 2021) qui propose une palette d'outils de traitement automatique du langage basée sur spaCy, avec notamment de l'attribution de dialogues (mais pas de génération de réseaux de personnages ni d'intégration de Flair), ou Renard (Amalvy *et al.*, 2024) qui n'était pas encore disponible au moment d'élaborer le code de Charnetto et fait un grand travail d'automatisation de chaque étape de la génération d'un réseau de personnages (offrant en outre la possibilité de générer des réseaux dynamiques). À notre connaissance, Charnetto est toutefois le seul outil en Python qui propose une extraction complète de réseaux de personnages en partant de scripts de films et en intégrant une syntaxe simplifiée pour des annotations manuelles, comme détaillé à la section B.1.

Les sections suivantes sont pensées pour donner une idée générale du fonctionnement de Charnetto. Pour des consignes précises d'utilisation, un fichier README, une documentation et un exemple détaillé en Python sont disponibles sur le répertoire charnetto.

B.1 Données

Charnetto fonctionne sur différents types de données textuelles :

- pour des romans, il peut exécuter Flair ou spaCy sur le texte afin d'identifier les occurrences des personnages automatiquement, ou interpréter une syntaxe d'annotations manuelles présentée plus bas,
- pour les scripts de films, il contient un code d'extraction des personnages pour le format trouvé notamment sur IMSDB [34],

34 https://imsdb.com/, consulté le 23.10.2023.

— pour tous les autres types de textes (comme des pièces de théâtre ou des scripts de jeux vidéo dont la structure peut varier énormément d'un texte à l'autre), il permet la génération du réseau de personnages sur la base d'un tableau recensant les occurrences des différents personnages (l'extraction de ces informations étant alors faite à part).

Si l'on veut utiliser Charnetto à partir d'un roman, l'idéal est d'utiliser un format comme celui du projet Gutenberg [35], à savoir un fichier `.txt` contenant uniquement le texte du roman.

Charnetto contient également un code permettant d'identifier les personnages des films, selon les conventions majoritairement utilisées dans IMSDB, qui est la base de données libre d'accès la plus complète actuellement en matière de scripts de films : noms de personnages en majuscule, avec leur réplique à la ligne, et indications de changement de scène également en majuscule (avec certains mots-clés repérables comme INT/EXT pour intérieur/extérieur, NIGHT/DAY pour jour/nuit, etc.). Une étude exploratoire a été menée sur les données de IMSDB pour assurer la stabilité du code et de la syntaxe identifiée [36], en testant le code d'extraction sur plusieurs centaines de scripts de films afin d'affiner les mots-clés indiquant des changements de scène, les mises en pages rencontrées et les manières de contrôler que les résultats sont conformes aux attentes.

Si l'on choisit d'annoter un roman à la main, Charnetto peut également se charger d'extraire les annotations pour générer le réseau associé. Pour cela, je propose une syntaxe simple reposant

35 https://www.gutenberg.org/, consulté le 23.11.2023.

36 Cette étude a eu lieu dans le cadre du cours « Programmation Orientée Objet (Python) » enseigné par Davide Picca, en collaboration avec Melinda Femminis, Florian Rieder et Andres Stadelmann, que je remercie grandement pour leur travail. Le répertoire du projet est disponible sur ce lien : https://github.com/dpicca/imsdb.

sur les URLs dans MarkDown, tel qu'illustré ci-dessous dans un extrait de *The Lion, the Witch and the Wardrobe* (C. S. Lewis, 1950) :

> "**[Narnia](LOC)** What's that?" said **[Lucy](PER)**.
> "This is the land of **[Narnia](LOC)**," said the **[Faun](PER)**, "where we are now; all that lies between the lamppost and the great castle of **[Cair Paravel](LOC)** on the eastern sea. And you—you have come from the wild woods of the west?"
> "I—I got in through the wardrobe in the spare room," said **[Lucy](PER)**.

À noter que Charnetto n'impose aucune contrainte sur la langue du texte à traiter. Si l'on désire passer par un algorithme de NER, on peut indiquer le modèle de langue choisi (pour autant qu'il soit mis à disposition par le modèle de NER sélectionné), et si l'on choisit de procéder soi-même à l'extraction des personnages, on peut fournir un tableau de données à Charnetto, comme celui décrit dans la section B.2, et procéder à la suite des étapes de manière automatique.

B.2 Extraction des entités et tableau de données

Comme présenté dans la section 4.2, Flair est l'algorithme le plus performant pour détecter les entités nommées dans les textes de fiction. Toutefois, il repose sur un framework coûteux en ressources (PyTorch), raison pour laquelle je n'ai pas voulu imposer son emploi dans le cadre de Charnetto. Le module a donc été conçu pour fonctionner avec Flair et avec spaCy, bien plus léger et rapide d'exécution, sur simple mention de l'algorithme souhaité au moment de lancer le processus.

Concrètement, la première étape d'extraction de Charnetto dépend du type de données mis à disposition :

— dans le cas d'un roman non annoté, il applique l'algorithme de NER choisi, muni du modèle de langage sélectionné,

— dans le cas d'un script de film, il utilise des expressions régulières en s'appuyant sur les conventions de IMSDB,
— dans le cas d'un texte annoté, il repère automatiquement les annotations manuelles en MarkDown,

et à l'issue de ces trois scénarios, il récolte pour chaque occurrence de personnage détectée les informations suivantes :

— `name` : l'intitulé de l'occurrence,
— `start_pos` : la position de départ de l'occurrence dans le fichier,
— `end_pos` : la position de fin de l'occurrence dans le fichier,
— `tag` : la catégorie attribuée par l'algorithme de NER à l'entité nommée (personnage, lieu, etc.), le cas échéant,
— `score` : le score de confiance si fourni par l'algorithme de NER (c'est le cas pour Flair),
— `block` : le numéro du bloc de texte qui contient l'occurrence (paragraphe, scène, etc.).

Enfin, ces informations sont réunies dans un tableau, sous forme d'objet Python de type `pandas.DataFrame` (comme l'exemple

TABLE B.1 Exemple de tableau pour le début du roman *The Nightingale Girls*.

name	start_pos	end_pos	score	tag	block
Chapter One	0	11	0,78	MISC	1
Doyle	28	33	0,53	PER	2
Bethnal Green	123	136	1,00	LOC	3
Dora Doyle	158	168	0,99	PER	3
Matron	230	236	0,79	PER	5
Nightingale Teaching Hospital	244	273	0,99	LOC	5
Dora	486	490	0,71	PER	6
Dora	493	497	1,00	PER	6
Matron	607	613	1,00	PER	8
Victoria Park	762	775	0,99	LOC	8
Nightingale	785	796	0,98	LOC	8
Dora Doyle	1030	1040	1,00	PER	11
East End	1057	1065	0,98	MISC	12
Thames	1172	1178	1,00	LOC	12
Gold's Garments	1236	1251	0,54	ORG	14
...	...	...	...	...	...

du tableau B.1 pour le roman *The Nightingale Girls* (Donna Douglas, 2012)) et facilement exportables sous différents formats.

Notons que dans le cas des films, la colonne `tag` sera toujours associée à la catégorie des personnages, car le code ne cherche pas à extraire d'autres types d'entités nommées. Dans la même logique, on fixe le score de confiance à 1 pour chaque ligne du tableau lorsque cette information n'est pas renseignée automatiquement.

B.3 Alias et liste de personnages

Une fois le tableau construit, il s'agit de constituer une liste de personnages, comme décrit à la section 4.1. La première étape consiste à parcourir les différentes occurrences d'une même entité au sein du tableau, afin d'uniformiser ses catégories : il peut arriver qu'un algorithme de NER manque de régularité dans l'attribution des catégories et considère que telle occurrence est un lieu, alors que la même entité a été catégorisée comme un personnage lors des dix occurrences précédentes. Pour uniformiser les résultats et éviter de passer à côté de certaines mentions d'un personnage, on effectue donc un vote à majorité, en recensant les différentes catégories attribuées à une même entité et en réaffectant à chacune de ses occurrences la catégorie la plus fréquemment prédite, dans une nouvelle colonne `utag` (pour « tag uniformisé ») pour éviter d'écraser l'information originale.

On filtre ensuite le tableau pour ne garder que les entités qui ont été catégorisées comme des personnages, ce qui donne une série d'alias, qu'il va falloir grouper en des identités distinctes. La validation de ces regroupements nécessite de toute façon une étape manuelle : il serait désirable de construire un code qui ne réclame aucune manipulation humaine de bout en bout, mais le problème de la résolution de coréférences est malheureusement loin d'être trivial et les meilleurs modèles actuels, coûteux en ressources, n'ont pas un taux de réussite qui permette, dans notre

cas, de se passer de vérification. Une amorce de regroupement est néanmoins incluse dans Charnetto pour faciliter cette manipulation. Ainsi, une première « liste de listes » est élaborée : chaque personnage y est représenté par une liste de ses alias (le premier étant le principal) et Charnetto repère les noms qui sont intégralement contenus dans d'autres noms pour les mettre directement dans la même liste d'alias. Cela permet par exemple de rassembler automatiquement le prénom, le nom et la combinaison « prénom nom » d'un même personnage, si les trois appellations sont détectées dans le texte. Notons qu'il peut tout à fait arriver qu'un même alias soit présent dans différentes listes, par exemple si deux personnages partagent le même prénom ou le même nom de famille.

```
# Les 4 héros
Lucy
  Lu
  LUCY
Edmund
  Ed
  EDMUND
Susan
  Susan the Gentle
  Su
Peter
  Peter the Magnificent
  PETER'S
  Peter Wolf's-Bane

Aslan #Le lion
  ASLAN
  O Aslan
  Lion Aslan
```

FIGURE B.1 Extrait de liste de personnages pour *The Lion, the Witch and the Wardrobe*.

Il est possible à ce stade de fixer un seuil minimal sur le nombre d'apparitions de chaque alias dans le texte pour qu'il soit pris en considération dans la liste. Ce seuil permet de filtrer les résultats qui seraient trop rares pour être pertinents (comme la mention d'un personnage extérieur à l'histoire et qui n'est évoqué qu'une fois dans le cadre d'une conversation, sans jamais avoir

d'effet sur la progression de l'histoire) ou qui ne correspondent pas aux occurrences que l'on cherche à capturer (notamment des noms de personnalités publiques, identifiées correctement comme des personnes, mais ne faisant pas référence à un personnage de l'histoire, par exemple le nom d'un chanteur qu'un personnage entendrait à la radio).

Cette première liste est ensuite déposée dans un fichier texte, avec cette amorce de hiérarchie entre l'alias principal et les autres alias présents dans le texte. C'est ce fichier qu'il faut ensuite parcourir manuellement pour compléter le travail de regroupement et qui a été pensé avec une syntaxe simple et facilement éditable. Les alias principaux sont classés par ordre décroissant de fréquence pour faciliter le travail d'annotation et rencontrer les personnages les plus présents en premier.

Comme on peut le voir dans la figure B.1, chaque alias est écrit sur une ligne, les alias principaux (c'est-à-dire ceux qui représenteront les personnages pour la suite des analyses) sont placés tout à gauche et les éventuels autres alias d'une même identité sont sous l'alias principal, avec une indentation qui permet de repérer la hiérarchie visuellement. L'alias principal est par défaut le premier alias rencontré dans le texte, mais on peut modifier la hiérarchie si l'on souhaite afficher un nom plus pertinent dans le réseau final. On peut également choisir de laisser des lignes vides ou de commenter la liste (avec le symbole #) pour clarifier la liste des personnages au besoin. Il est utile de noter que la suite du processus est sensible à la casse (le contraire risquerait d'introduire beaucoup d'erreurs pour des textes dans lesquels certains personnages ont des noms qui peuvent également être utilisés comme des noms communs, à l'image de *L'assassin royal*), raison pour laquelle on trouve dans la liste des occurrences en majuscules (comme « LUCY » à la figure B.1).

Il arrive qu'en corrigeant la liste manuellement, on s'aperçoive qu'un personnage a été totalement oublié par l'algorithme de NER. Dans *Coraline* (Neil Gaiman, 2002) par exemple, les

« autres » parents de Coraline dans le monde parallèle (« *other mother* » et « *other father* ») ne sont pas identifiés par Flair, qui ne retient que « *mother* » et « *father* », englobant à tort les deux mères sous la même identité. Pour récupérer ces personnages manquants, on peut donner à Charnetto une ou plusieurs expressions régulières décrivant précisément la syntaxe du ou des alias à ajouter à la liste, générant par la même occasion un nouveau tableau d'occurrences actualisé.

B.4 Cooccurrences

Nous avons à présent une liste des personnages, avec leurs différents alias, ainsi que le tableau de toutes les occurrences de ces alias dans le texte initial, organisés en « blocs » (paragraphes, scènes, etc. selon les types de texte et les choix d'analyse). Il reste à choisir la taille de la fenêtre d'observation qui permettra de déterminer si deux entités sont cooccurrentes : si le découpage est fait par scène, on va préférer fixer la taille à 1 pour considérer en cooccurrence les personnages présents dans la même scène. Dans un roman que l'on observerait à l'échelle du paragraphe, en revanche, il est possible que l'on préfère mettre en cooccurrence des personnages présents dans le même groupe de quatre ou cinq paragraphes, selon la taille du texte, la longueur des paragraphes et d'autres critères propres à chaque œuvre.

Pour établir la liste des cooccurrences, Charnetto divise le tableau selon la fenêtre d'observation choisie [37] et récupère la liste de tous les alias présents dans cette fenêtre. Il faut ensuite les

[37] Il a été choisi ici de ne pas faire de fenêtre mobile pour que le nombre d'interactions entre deux personnages ne soit pas artificiellement gonflé par des observations successives du même paragraphe. Ainsi, en l'état, on fixe une taille de fenêtre (en nombre de paragraphes) et on découpe l'ensemble du texte en blocs de cette taille, sans recoupement. Une amélioration du code permettant de choisir entre fenêtre fixe et fenêtre mobile est toutefois envisagée.

comparer à la liste de personnages pour déterminer à quel personnage ils font référence et les remplacer par l'alias principal correspondant. Il est possible de rencontrer des ambiguïtés lors de cette attribution : si deux personnages s'appellent John, l'alias « John » sera présent dans les deux listes d'alias et chaque occurrence de « John » peut donc pointer sur l'un ou l'autre des personnages selon les situations. Là encore, il n'existe pas de solution entièrement fiable pour traiter ces cas : j'ai donc pris le parti de remonter les lignes du tableau jusqu'à l'occurrence la plus proche d'un des candidats possibles de cette ambiguïté et d'attribuer cette identité à l'occurrence concernée. Cette stratégie repose sur l'hypothèse que dans un roman, pour ne pas perdre le lectorat, on s'autorise à utiliser un alias pouvant porter à confusion uniquement lorsque plus tôt dans le texte, on a spécifié plus clairement à quel personnage on faisait référence. La méthode peut bien sûr introduire des erreurs dans la définition des cooccurrences, mais le résultat final étant un agrégat de toute cette récolte de données, les tendances générales devraient rester globalement stables.

Lorsque chaque personnage de la fenêtre d'observation a été identifié, une simple combinaison de toutes les paires possibles de personnages permet de créer les cooccurrences de cette fenêtre, qui sont ajoutées à une liste globale des cooccurrences avant de passer à la fenêtre suivante. On considère ici que les cooccurrences ne sont pas orientées. On se retrouve finalement avec une liste de toutes les cooccurrences entre deux personnages trouvées dans l'intégralité du tableau, comme dans l'exemple suivant (extrait des cooccurrences du roman *The Terror* de Dan Simmons [2007]) :

```
[('Captain Crozier', 'Francis Rawdon Moira Crozier'),
 ('Captain Crozier', 'James Fitzjames'),
 ('Captain Crozier', 'James Ross'),
```

```
('Captain Crozier', 'John Franklin'),
('Captain Crozier', 'Sophia Cracroft'),
('Francis Rawdon Moira Crozier', 'James Fitzjames'),
('Francis Rawdon Moira Crozier', 'James Ross'),
('Francis Rawdon Moira Crozier', 'John Franklin'),
('Francis Rawdon Moira Crozier', 'Sophia Cracroft'),
('James Fitzjames', 'James Ross'),
('James Fitzjames', 'John Franklin'),
('James Fitzjames', 'Sophia Cracroft'),
('James Ross', 'John Franklin'),
('James Ross', 'Sophia Cracroft'),
('John Franklin', 'Sophia Cracroft'),
('Captain Crozier', 'Cornelius Hickey'),
...]
```

Le poids associé aux arêtes du réseau de personnages représente le nombre d'apparitions d'une même paire (non orientée) de personnages au sein de cette liste. Il est également possible à ce stade de fixer un seuil minimal sur ce poids, pour filtrer les interactions trop rares et identifier plus aisément les phénomènes globaux du récit. La figure B.2 illustre la variation de ce seuil pour le roman *Anna Karenina* (Tolstoï, 1878) : au vu du nombre de personnages, un seuil de 1 (signifiant l'affichage de toutes les cooccurrences) produit un réseau très fourni, alors qu'un filtre à 10 ou 20 cooccurrences minimales permet une vision d'ensemble plus claire et épurée, mettant mieux en avant les interactions les plus fréquentes.

B.5 Génération du réseau

La dernière étape est la plus directe, puisque tous les éléments sont prêts. On s'appuie sur le package Networkx (Hagberg *et al.*, 2008) qui permet de générer des réseaux : la liste de personnages sert de base pour construire les nœuds (de sorte que même des

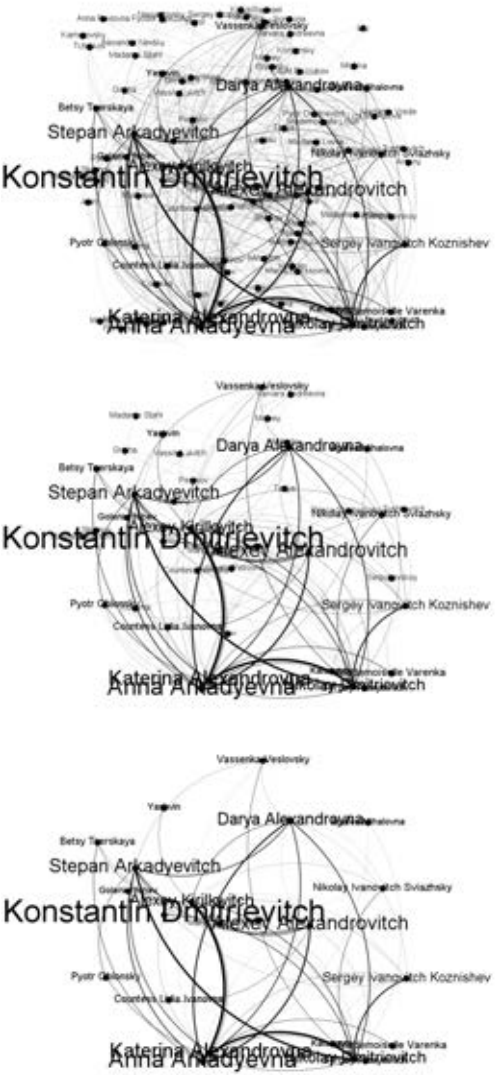


FIGURE B.2 Variation du seuil minimal de poids des arêtes pour le roman *Anna Karenina* : de gauche à droite, un poids minimal de 1, 10 et 20.

personnages qui ne seraient en interaction avec personne seront représentés dans le réseau) et la liste de cooccurrences permet de définir les arêtes. On enrichit ce réseau avec un poids des nœuds défini par le nombre d'occurrences de chaque personnage dans le tableau des occurrences et un poids des arêtes donné par le nombre de fois où les interactions sont présentes dans la liste des cooccurrences.

Le choix du meilleur outil de visualisation pour le réseau est laissé libre, en fonction des besoins d'interactivité et des limitations techniques. La solution la plus simple est de l'afficher directement avec Python (grâce à des librairies comme Matplotlib [Hunter, 2007]), mais on peut également l'exporter pour utiliser des outils externes comme Gephi (Bastian *et al.*, 2009) ou construire sa propre visualisation dynamique en javascript avec d3.js (Bostock, 2012), par exemple.

B.6 Contrôle de qualité des résultats

Si l'étape d'extraction des entités nommées a déjà fait l'objet d'un contrôle de qualité, en comparant les résultats de l'outil avec des données annotées dans l'annexe A, on peut se demander dans quelle mesure les écarts observés dans la récupération de ces entités se retrouvent dans le réseau final, et à quel point le fait d'agréger les informations permet de compenser les erreurs individuelles.

Un tel contrôle nécessite une annotation manuelle qui servira de « *gold standard* » : j'ai donc annoté manuellement l'intégralité de *The Lion, the Witch and the Wardrobe*, roman également utilisé à la sous-section 8.3.1 dans le cadre de sa comparaison avec le film *The Chronicles of Narnia: The Lion, the Witch and the Wardrobe*. Le fruit de cette annotation (centrée exclusivement sur la récupération des personnages) est à retrouver sur le répertoire character_network, sous forme de tableau Excel. En utilisant le

modèle `ner-english-large` (Schweter et Akbik, 2020) de Flair pour extraire automatiquement les entités nommées, on obtient une précision de 0,91 (qui correspond à 9 % de mots faussement interprétés comme des noms de personnages) et un rappel de 0,93, c'est-à-dire que 7 % des occurrences de personnages identifiées dans la version manuelle n'ont pas été détectées par l'outil.

La figure B.3 présente les deux réseaux de personnages issus de cette démarche : en haut, celui généré grâce à Charnetto en



FIGURE B.3 Réseaux de personnages du roman *The Lion, the Witch and the Wardrobe* : en haut, avec Flair, en bas, en annotation manuelle.

utilisant Flair pour l'extraction des entités nommées, en fixant les différents seuils à 1 pour le nombre d'apparitions minimales des nœuds et des arêtes, et à 3 pour la taille de la fenêtre d'observation ; en bas, celui généré grâce à Charnetto sur la base des annotations manuelles, avec les mêmes seuils pour assurer une continuité dans la création des arêtes.

Reste à savoir comment s'y prendre pour comparer ces réseaux. Un simple regard ne donne qu'une vague idée des similarités et des différences entre les deux résultats, et peut même donner un sentiment biaisé de leur proximité : sur la base des rendus de la figure B.3, on peut en effet présumer qu'ils sont sensiblement différents, alors que les informations quantitatives rapportées ci-dessous indiquent le contraire.

La question de la comparaison de plusieurs réseaux est traitée en détail au chapitre 7, dans le cadre de l'étude sur les adaptations. Ici, plutôt que d'utiliser des mesures plus ou moins fines pour estimer la distance entre les deux graphes, on choisit de se concentrer sur la comparaison des mesures calculées sur chaque réseau. La logique est la suivante : un même réseau peut avoir un aspect très différent selon les choix opérés dans sa visualisation (la disposition des nœuds, l'intensité des forces d'attraction et de répulsion, etc.), mais les valeurs des mesures calculées sur ce réseau resteront stables à travers toutes ses représentations. Ainsi, si l'objectif de la génération d'un réseau de personnages est de pouvoir en mesurer des propriétés comme la densité, la transitivité ou les nœuds les plus centraux, il faut en priorité que notre modèle de génération du réseau assure une stabilité dans la production (et ainsi l'interprétation) de ces mesures, davantage que sur un aspect visuel ou sur la structure exacte du réseau de référence.

Différentes mesures ont donc été calculées sur chaque réseau et reportées dans le tableau B.2. Pour les mesures globales, on retrouve le nombre de nœuds, le diamètre (voir définition 6, page 30), l'excentricité moyenne (voir définition 14, page 34), la

TABLE B.2 Comparaison des mesures entre Flair et les annotations manuelles pour *The Lion, The Witch and the Wardrobe*.

Mesures	**Flair**	***Gold standard***
Nb de nœuds	24	24
Diamètre	3	3
Excentricité moy.	2,21	2,29
Densité	0,45	0,41
Transitivité	0,69	0,65
Distance moy.	1,56	1,64
Clique maximale	11	9
Degré nœuds	Lucy 20 Peter 20 Edmund 19 Susan 17 Tumnus 17	Peter 19 Lucy 18 Edmund 17 Susan 17 Tumnus 14
Force nœuds	Lucy 327 Peter 271 Susan 220 Edmund 200 Aslan 196	Lucy 314 Peter 252 Susan 224 Edmund 207 Aslan 189
Poids arêtes	Lucy - Susan 59 Lucy - Peter 57 Susan - Peter 49 Lucy - Aslan 43 Lucy - Edmund 41	Lucy - Susan 58 Lucy - Peter 55 Susan - Peter 48 Edmund - White Witch 46 Lucy - Aslan 44
Centralité prox.	Lucy 0,88 Peter 0,88 Edmund 0,85 Susan 0,79 Tumnus 0,79	Peter 0,85 Lucy 0,82 Edmund 0,79 Susan 0,79 Tumnus 0,72
Centralité interm.	Edmund 0,14 Peter 0,12 Lucy 0,12 Tumnus 0,08 Susan 0,04	Peter 0,15 Lucy 0,14 Edmund 0,11 Tumnus 0,09 Susan 0,08

densité (voir définition 8, page 31), la transitivité (voir définition 9, page 31), la distance moyenne (voir définition 7, page 31) et la taille de la plus grande clique. Viennent ensuite les mesures sur les nœuds et les arêtes, avec les cinq valeurs les plus élevées pour le degré (voir définition 10, page 32), la force (voir définition 11, page 32), la centralité de proximité (voir définition 12, page 33) et la centralité intermédiaire (voir définition 13, page 33) des nœuds, ainsi que le poids des arêtes.

En observant le tableau, on peut voir que les deux réseaux sont relativement similaires si l'on se concentre sur la comparaison de leurs mesures globales. Le degré et la force des nœuds sont également stables sur les cinq valeurs les plus hautes, de même que la centralité de proximité. On observe par contre quelques changements dans le poids des arêtes, avec l'émergence dans la version manuelle de l'arête entre Edmund et la Sorcière Blanche (White Witch), dont on peut soupçonner que ses différentes appellations (jamais par un nom propre) ont été plus délicates à capturer à l'aide d'un algorithme de NER. La centralité intermédiaire subit quelques permutations dans le classement, même si les valeurs elles-mêmes sont globalement équivalentes.

En résumé, on peut apprécier la qualité du résultat automatisé en sachant qu'il manquait 7 % des entités nommées au départ (ce qui représente un peu moins d'une centaine d'occurrences) et que 9 % des mentions de personnages étaient en réalité incorrectes. En affinant la détection de certains personnages comme la Sorcière Blanche à l'aide d'expressions régulières comme proposé dans Charnetto, on peut s'attendre à des résultats encore plus précis dans les mesures sur les nœuds et les arêtes, rendant l'usage de solutions automatisées d'autant plus intéressant.

Sources

Romans

Austen, J. (1811). *Sense and Sensibility*. Thomas Egerton.
Bry, D. (2017). *Que passe l'hiver*. Éditions de l'Homme Sans Nom.
Douglas, D. (2012). *The Nightingale Girls*. Arrow.
Dowd, S. (2006). *A Swift Pure Cry*. David Fickling Books.
Ellroy, J. (1987). *The Black Dahlia*. The Mysterious Press.
Ellroy, J. (1990). *L.A. Confidential*. The Mysterious Press.
Gaiman, N. (2002). *Coraline*. Bloomsbury Publishing.
Gregory, P. (2001). *The Other Boleyn Girl*. Scribner.
Harris, T. (1988). *The Silence of the Lambs*. St. Martin's Press.
Hobb, R. (1995). *Assassin's Apprentice*. Spectra.
Hobb, R. (1998). *L'apprenti assassin*. Éditions Pygmalion.
King, S. (1996). *The Green Mile*. Signet Books.
Lewis, C. S. (1950). *The Lion, the Witch and the Wardrobe*. Geoffrey Bles.
Simmons, D. (2007). *The Terror*. Little Brown & Company.
Stockett, K. (2009). *The Help*. Penguin Books.
Tassy, V. (2018). *Comment le dire à la nuit*. Éditions le Chat Noir.
Tolstoï, L. (1878). *Anna Karenina*. The Russian Messenger.
Webster, D., & Hopkins, M. A. (1930). *Consider the Consequences*. The Century Company.

Films

Chadwick, J. (2008). *The Other Boleyn Girl*.
De Palma, B. (2006). *The Black Dahlia*.

Demme, J. (1991). *The Silence of the Lambs*.
Hanson, C. (1997). *L.A. Confidential*.
Lee, A. (1995). *Sense and Sensibility*.
Luhrmann, B. (1996). *William Shakespeare's Romeo + Juliet*.
Selick, H. (2009). *Coraline*.
Taylor, T. (2011). *The Help*.
Wright, J. (2012). *Anna Karenina*.

Jeux vidéo

Life is Strange. (2015). Dontnod Entertainment.
Life is Strange 2. (2019). Dontnod Entertainment, Feral Interactive.
Life is Strange : Before the Storm. (2017). Deck Nine Games.
Life is Strange : True Colors. (2021). Deck Nine Games.
Phoenix Wright : Ace Attorney. (2006). Capcom.

Pièces de théâtre

Shakespeare, W. (1567). *Romeo and Juliet*.

Opéras

Gounod, C. (1867). *Roméo et Juliette*.

Bibliographie

Aarseth, E. (1997). *Cybertext : Perspectives on Ergodic Literature*. Johns Hopkins University Press.

Abdi, H. (2010). Congruence : Congruence Coefficient, RV Coefficient, and Mantel Coefficient. *Encyclopedia of research design*, *3*, 222-229.

Abu-Aisheh, Z., Raveaux, R., Ramel, J.-Y., & Martineau, P. (2015). An Exact Graph Edit Distance Algorithm for Solving Pattern Recognition Problems. *4th International Conference on Pattern Recognition Applications and Methods 2015*.

Agarwal, A., Corvalan, A., Jensen, J., & Rambow, O. (2012). Social Network Analysis of Alice in Wonderland. *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, 88-96.

Ahuja, R., Magnanti, T., & Orlin, J. (1993). *Network Flows : Theory, Algorithms, and Applications*. Prentice Hall.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR : An Easy-to-Use Framework for State-of-the-Art NLP. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54-59.

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. *COLING 2018, 27th International Conference on Computational Linguistics*, 1638-1649.

Akoglu, L., Tong, H., & Koutra, D. (2014). Graph-based Anomaly Detection and Description : A Survey. *Data Mining and Knowledge Discovery*, *29*.

Alberich, R., Miro-Julia, J., & Rosselló, F. (2002). Marvel Universe Looks Almost Like a Real Social Network. *ArXiv*.

Amalvy, A., Labatut, V., & Dufour, R. (2024). Renard : A Modular Pipeline for Extracting Character Networks from Narrative Texts. *Journal of Open Source Software*, *9*, 65-74.

Arinik, N., Figueiredo, R., & Labatut, V. (2020). Multiple Oartitioning of Multiplex Signed Networks : Application to European Parliament Votes. *Social Networks*, *60*, 83-102.

Bamman, D. (2021). *BookNLP*. https : / / github . com / booknlp / booknlp.

Barabási, A.-L., & Pósfai, M. (2016). *Network Science*. Cambridge University Press.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi : An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the Third International Conference on Weblogs and Social Media*.

Bateman, C. (2006). Keeping the Player on Track. In *Game Writing : Narrative Skills for Videogames*. Charles River Media, Inc.

Bavaud, F. (2023). Exact First Moments of the RV Coefficient by Invariant Orthogonal Integration. *Journal of Multivariate Analysis*, *198*.

Bavaud, F., & Métrailler, C. (2023). A (Dis)similarity Index for Comparing Two Character Networks Based on the Same Story. *Proceedings of the Workshop on Computational Methods in the Humanities 2022*.

Beveridge, A., & Shan, J. (2016). Network of Thrones. *Math Horizons*, *23*, 18.

Birkholz, J. M., & Budke, L. (2021). Distant and Close Reading in Literature : a Case of Networks in Periodical Studies. *Interférences littéraires*, 204-217.

Blom, J. (2023). *The Dynamic Game Character*. Amsterdam University Press.

Bonato, A., D'Angelo, D. R., Elenberg, E. R., Gleich, D. F., & Hou, Y. (2016). Mining and Modeling Character Networks. *ArXiv*.

Bossaert, G., & Meidert, N. (2013). "We Are Only as Strong as We Are United, as Weak as We Are Divided" a Dynamic Analysis of the Peer Support Networks in the Harry Potter Books. *Open Journal of Applied Sciences, 03*, 174-185.

Bostock, M. (2012). *D3.js - Data-Driven Documents*. http://d3js.org/.

Boyd, B. (2017, mai). Making Adaptation Studies Adaptive. In *The Oxford Handbook of Adaptation Studies*. Oxford University Press.

Caïra, O. (2019). Qu'allez-vous faire de Roméo ? *COnTEXTES*.

Caïra, O. (2020). Métamorphoses du chapitre dans la scénarisation interactive. *Itinéraires*.

Cardwell, S. (2002). *Adaptation Revisited : Television and the Classic Novel*. Manchester University Press.

Chan, T. H. L. (2012). A Survey of the 'New' Discipline of Adaptation Studies : Between Translation and Interculturalism. *Perspectives, 20*, 411-418.

Cipresso, P., & Riva, G. (2016). Computational Psychometrics Meets Hollywood : The Complexity in Emotional Storytelling. *Frontiers in Psychology*.

Cuesta-Lazaro, C., Prasad, A., & Wood, T. (2022). What Does the Sea Say to the Shore ? A BERT Based DST Style Approach for Speaker to Dialogue Attribution in Novels. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 5820-5829.

Da, N. Z. (2019). The Digital Humanities Debacle. *The Chronicle of Higher Education, 65*.

Deleris, L., Bonin, F., Daly, E., Deparis, S., Hou, Y., Jochim, C., Lassoued, Y., & Levacher, K. (2018). Know Who Your Friends Are : Understanding Social Connections from Unstructured Text. *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics : Demonstrations, 76-80.

Donnat, C., & Holmes, S. (2018). Tracking Network Dynamics : A Survey Using Graph Distances. *The Annals of Applied Statistics*, *12*, 971-1012.

Ek, A., & Wirén, M. (2019). Distinguishing Narration and Speech in Prose Fiction Dialogues. *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, 124-132.

Elliott, K. (2017, mai). Adaptation Theory and Adaptation Scholarship. In *The Oxford Handbook of Adaptation Studies*. Oxford University Press.

Elson, D., Dames, N., & McKeown, K. (2010). Extracting Social Networks from Literary Fiction. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 138-147.

Elson, D. K. (2012). *Modeling Narrative Discourse* [thèse de doct., Columbia University].

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, *29*, 751-760.

Falk, M. (2016). Making Connections : Network Analysis, the Bildungsroman and the World of The Absentee. *Journal of Language, Literature and Culture*, *63*, 107-122.

Gao, X., Xiao, B., Tao, D., & Li, X. (2010). A Survey of Graph Edit Distance. *Pattern Anal. Appl.*, *13*, 113-129.

Givon, S., & Milosavljevic, M. (2007). Extracting Useful Information from the Full Text of Fiction. *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, 633-638.

Gleiser, P. (2007). How to Become a Superhero. *Journal of Statistical Mechanics Theory and Experiment*, *2007*.

Goldberg, A., Tardos, É., & Tarjan, R. (1989). *Network Flow Algorithms* (T. 216). Princeton University, Department of Computer Science.

Góngora, S., Chiruzzo, L., Méndez, G., & Gervás, P. (2024). PAYADOR : A Minimalist Approach to Grounding Language Models on Structured Data for Interactive Storytelling and

Role-playing Games. *Proceedings of the 15th International Conference on Computational Creativity*.

Guex, G., Courtain, S., & Saerens, M. (2021). Covariance and Correlation Measures on a Graph in a Generalized Bag-of-paths Formalism. *Journal of Complex Networks, 8*.

Hagberg, A., Swart, P., & Chult, D. (2008). Exploring Network Structure, Dynamics, and Function Using NetworkX. *Proceedings of the 7th Python in Science Conference*.

Honnibal, M., & Montani, I. (2017). *spaCy*. https://spacy.io/.

Hunter, J. D. (2007). Matplotlib : A 2D Graphics Environment. *Computing in Science & Engineering, 9*, 90-95.

Iyyer, M., Guha, A., Chaturvedi, S., Boyd-Graber, J., & Daumé III, H. (2016). Feuding Families and Former Friends : Unsupervised Learning for Dynamic Fictional Relationships. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, 1534-1544.

Jannidis, F. (2009). Character. In P. Hühn, J. Pier, W. Schmid & J. Schönert (Éd.), *Handbook of Narratology* (p. 14-29). De Gruyter.

Jeanson, L., Guex, G., & Xanthos, A. (2024). Lexical Diversity Measurement Using Subsample Entropy : Formalism and Evaluation. *JADT 2024, mots comptés, textes déchiffrés, tome 1*, 439-448.

Jenkins, H. (2006). *Convergence Culture : Where Old and New Media Collide*. NYU Press.

Johnson, D. T. (2017, mai). 87Adaptation and Fidelity. In *The Oxford Handbook of Adaptation Studies*. Oxford University Press.

Josse, J., Pagès, J., & Husson, F. (2008). Testing the Significance of the RV Coefficient. *Computational Statistics & Data Analysis, 53*, 82-91.

Kemeny, J. G., & Snell, J. L. (1976 [1960]). *Finite Markov Chains*. Springer New York.

Kranz, D. L. (2007). Trying Harder : Probability, Objectivity, and Rationality in Adaptation Studies. In *The Literature/Film Reader : Issues of adaptation*. Lanham (Md.) : Scarecrow Press.

Kumaran, V., Rowe, J., Mott, B., & Lester, J. (2023). SceneCraft : Automating Interactive Narrative Scene Generation in Digital Games with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *19*, 86-96.

Kuo, H. H. (1975). *Gaussian measures on Banach spaces*. Springer.

Kydros, D., & Anastasiadis, A. (2015). Social Network Analysis in Literature. The Case of The Great Eastern. *Continuities, Discontinuities, Ruptures in the Greek World (1204-2014) : Economy, Society, History, Literature. Proceedings [of the] 5th European Congress of Modern Greek Studies of the European Society of Modern Greek Studies*, *4*.

Kydros, D., Notopoulos, P., & Exarchos, G. (2015). Homer's Iliad – A Social Network Analytic Approach. *International Journal of Humanities and Arts Computing*, *9*, 115-132.

Labatut, V., & Bost, X. (2019). Extraction and Analysis of Fictional Character Networks : A Survey. *ACM Comput. Surv.*, *52*.

Lee, J., & Yeung, C. Y. (2012). Extracting Networks of People and Places from Literary Texts. *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, 209-218.

Lee, O.-J., & Jung, J. J. (2019). Modeling Affective Character Network for Story Analytics. *Future Generation Computer Systems*, *92*, 458-478.

Liu, D., & Albergante, L. (2018). Balance of Thrones : a Network Study on *Game of Thrones*. *ArXiv*.

Marti, M., & Baroni, R. (2014). De l'interactivité du récit au récit interactif. *Cahiers de Narratologie*.

Masías, V. H., Baldwin, P., Laengle, S., Vargas Schüler, A., & Crespo, F. (2016). Exploring the Prominence of Romeo and Juliet 's Characters Using Weighted Centrality Measures. *Digital Scholarship in the Humanities*, *32*.

McKee, R. (1997). *Story : Substance, Structure, Style and the Principles of Screenwriting*. HarperCollins.

Min, S., & Park, J. (2016). Network Science and Narratives : Basic Model and Application to Victor Hugo's Les Misérables. In *Complex Networks VII : Proceedings of the 7th Workshop on Complex Networks CompleNet 2016* (p. 257-265). Springer International Publishing.

Monge, G. (1781). *Mémoire sur la théorie des déblais et des remblais*. Imprimerie royale.

Moretti, F. (2000). Conjectures on World Literature. *New Left Review*.

Nakamura, M., & Go, I. (2011). Tezuka Is Dead : Manga in Transformation and Its Dysfunctional Discourse. *Mechademia, 6*.

O'Keefe, T., Pareti, S., Curran, J. R., Koprinska, I., & Honnibal, M. (2012). A Sequence Labelling Approach to Quote Attribution. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 790-799.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node Centrality in Weighted Networks : Generalizing Degree and Shortest Paths. *Social Networks, 32*, 245-251.

Pante, I. (2020). *Fictions interactives*. https://isaacpante.net/if/.

Park, G.-M., Kim, S.-H., Hwang, H.-R., & Cho, H.-G. (2013). Complex System Analysis of Social Networks Extracted from Literary Fictions. *International Journal of Machine Learning and Computing*, 107-111.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch : An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32* (p. 8024-8035). Curran Associates, Inc.

Propp, V. (1970 [1928]). *Morphologie du conte*. Seuil.

Ransom, J. C. (1941). *The New Criticism*. New Directions.

Riedl, M., & Bulitko, V. (2013). Interactive Narrative : An Intelligent Systems Approach. *AI Magazine*, *34*, 67-77.

Riesen, K. (2016). *Structural Pattern Recognition with Graph Edit Distance : Approximation Algorithms and Applications* (1st). Springer Publishing Company, Incorporated.

Robert, P., & Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods : The RV- Coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *25*, 257-265.

Rochat, Y., & Triclot, M. (2016). Les réseaux de personnages de science-fiction : échantillons de lectures intermédiaires. *ReS Futurae*, *10*.

Schafer, T. (1996). Grim Fandango Puzzle Document.

Schweter, S., & Akbik, A. (2020). FLERT : Document-Level Features for Named Entity Recognition. *ArXiv*.

Shan, Q., & Levina, E. (2022). Network Resampling for Estimating Uncertainty. *ArXiv*.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel Methods for Pattern Analysis* (illustrated edition). Cambridge University Press.

Shoubridge, P., Kraetzl, M., & Ray, D. (1999). Detection of Abnormal Change in Dynamic Networks. *1999 Information, Decision and Control. Data and Information Fusion Symposium, Signal Processing and Communications Symposium and Decision and Control Symposium. Proceedings*, 557-562.

Solomon, J. M. (2018). Optimal Transport on Discrete Domains. *ArXiv*.

Sorlin, S., & Solnon, C. (2005). Similarité de graphes : une mesure générique et un algorithme tabou réactif. *7e rencontres nationales des jeunes chercheurs en intelligence artificielle, RJCIA'2005*, 253-266.

Srivastava, S., Chaturvedi, S., & Mitchell, T. M. (2015). Inferring Interpersonal Relations in Narrative Summaries. *AAAI Conference on Artificial Intelligence*.

Stanislawek, T., Wróblewska, A., Wójcicka, A., Ziembicki, D., & Biecek, P. (2019). Named Entity Recognition — Is There a Glass Ceiling? *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 624-633.

Stiller, J., Nettle, D., & Dunbar, R. (2003). The Small World of Shakespeare's Plays. *Human Nature*, *14*, 397-408.

Suen, C., Kuenzel, L., & Gil, S. C. (2013). Extraction and Analysis of Character Interaction Networks From Plays and Movies. *Digital Humanities Conference*.

Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020). Anaphora and Coreference Resolution : A Review. *Information Fusion*, *59*, 139-162.

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task : Language-Independent Named Entity Recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142-147.

Truby, J. (2007). *The Anatomy of Story*. Faber ; Faber.

Vala, H., Jurgens, D., Piper, A., & Ruths, D. (2015). Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized : On The Difficulty of Detecting Characters in Literary Texts. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 769-774.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Venturini, T., Bounegru, L., Jacomy, M., & Gray, J. (2017). 11. How to Tell Stories with Networks : Exploring the Narrative Affordances of Graphs with the Iliad. In *Studying Culture through Data* (p. 155-170). Amsterdam University Press.

Villani, C. (2008). *Optimal Transport : Old and New*. Springer Berlin Heidelberg.

Weimer, A. M., Varachkina, H., Gödeke, L., & Gittel, B. (2024). The Author's Speech : Empirical Perspectives on Speaker Attribution in Narrative Fiction. In *Ambiguity and Narratology, Interdisciplinary Perspectives and Diachronic Case Studies* (p. 53-78). De Gruyter.

Wilde, L. R. (2019). *Kyara* Revisited : The Pre-Narrative Character-State of Japanese Character Theory. *Frontiers of Narrative Studies*, *5*, 220-247.

Wills, P., & Meyer, F. G. (2019). Metrics for graph comparison : A practitioner's guide. *PLoS ONE*, *15*.

Woloch, A. (2003). *The One vs. the Many : Minor Characters and the Space of the Protagonist in the Novel*. Princeton University Press.

Table des matières

Sommaire 5

Remerciements 7

Première partie

Introduction et bases théoriques 11

1 Introduction 13

2 Bases narratives 19

3 Bases mathématiques 23

3.1 Graphes et réseaux 23

3.2 Réseaux de personnages 26

3.3 Mesures et interprétations 29

4 Bases informatiques 37

4.1 Extraction de réseaux 38

4.2 Reconnaissance d'entités nommées 41

Deuxième partie

Narrations plurielles et interactives 45

5 Approche théorique 47

5.1 Flux narratif 48

5.2 Illustration : taille de vocabulaire 54

5.3 Réseau moyen 57

6 Étude de cas : *Life is Strange* **61**
6.1 Expérience préliminaire : *Phoenix Wright* 61
6.2 Choix de *Life is Strange* 64
6.3 Récolte des données. 68
6.4 Construction du flux 74
6.5 Résultats et analyses 84
6.5.1 Vérification des valeurs empiriques 85
6.5.2 Poids moyen des nœuds et des arêtes 88
6.5.3 Écart-type absolu et relatif 93
6.5.4 Mesures et interprétations 95
6.5.5 Observations par épisode 99
6.6 Ouverture sur la suite 101

Troisième partie

Œuvres plurielles et comparaisons **107**

7 Approche théorique **109**
7.1 Comparaison de mesures 110
7.2 Mesures de comparaison 114
7.2.1 Distance d'édition de graphes 115
7.2.2 Transport optimal 117
7.2.3 Coefficient RV 125

8 Étude de cas : comparaison d'adaptations **131**
8.1 Récolte des données 132
8.2 Correspondance des personnages 135
8.3 Résultats et analyses 136
8.3.1 Tableau général des résultats. 137
8.3.2 Stratégie pour les futures études d'adaptations 150
8.4 Exemple pratique : le triplet *Romeo and Juliet* 150
8.5 Aparté : transport optimal de personnages 153

Quatrième partie

Conclusion et ouverture 161

9 Continuer l'exploration 163
9.1 Rappel de la problématique 163
9.2 Résultats principaux 164
9.3 Limites et périmètre du travail 167
9.4 Perspectives 170

Cinquième partie

Annexes 173

A Flair ou spaCy ? 175

B Charnetto 181
B.1 Données 182
B.2 Extraction des entités et tableau de données 184
B.3 Alias et liste de personnages 186
B.4 Cooccurrences 189
B.5 Génération du réseau 191
B.6 Contrôle de qualité des résultats 193

Sources 199

Bibliographie 201