Détection et suivi d'événements dans des documents de presse historiques

Guillaume Bernard

LA ROCHELLE UNIVERSITÉ

ÉCOLE DOCTORALE EUCLIDE

LABORATOIRE L3i

THÈSE

présentée pour obtenir le grade de

Docteur de La Rochelle Université

par

Guillaume BERNARD

soutenue le **10 novembre 2022**

Discipline : **Informatique et Applications**

Détection et suivi d'événements dans des documents de presse historique

---

Rapporteurs	Cédrick Fairon	Professeur ordinaire (eq. : Professeur des universités)	Université catholique de Louvain Wallonie, Belgique
	Els Lefever	*Hoofddocent* (eq. : Maîtresse de conférences HDR)	*Universiteit Gent Vlaanderen, België*
Examinateurs	Karell Bertet	Maîtresse de conférences HDR	La Rochelle Université
	Benoit Favre	Professeur des universités	Aix-Marseille Université
	Glenn Roe	Professeur des universités	Sorbonne Université
Encadrement	Antoine Doucet	Professeur des universités	La Rochelle Université
	Cyril Faucher	Maître de conférences	La Rochelle Université
	Cyrille Suire	Enseignant-chercheur	La Rochelle Université

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

Liberté
Égalité
Fraternité

Thèse réalisée au Laboratoire L3i
La Rochelle Université - Institut LUDI
Avenue Michel Crépeau
17 042 La Rochelle Cedex 01

Tél : +33 5 46 45 82 62
Site Web : `http://l3i.univ-larochelle.fr`

Sous la direction de Antoine Doucet Professeur des universités

Financement Ministère de l'enseignement supérieur, de la recherche et de l'innovation

Soutenance `www.guillaume-bernard.fr/soutenance-de-these-2022`

Résumé

Reconstruire le fil des événements dans la presse, retracer le parcours de fausses informations ou identifier le point d'origine d'une publication, véridique ou non, sont des enjeux de ce début de siècle. La numérisation croissante des contenus numériques favorise des analyses massives de la presse. La mise à disposition depuis le début des années 2010 de fonds documentaires de presse par les bibliothèques du monde entier ouvre de nouvelles opportunités. Il devient possible d'adopter un autre regard sur des événements du passé, en découdre le fil et en connaître à la fois les chronologies et les diffusions géographiques.

L'analyse des documents historiques relève de processus singuliers. En apparences similaires à des articles issus de la presse numérique, ils diffèrent pourtant en tout point : sur la forme (type de document, format des données) et sur le fond (style journalistique, thèmes abordés, personnages mentionnés). Ce travail de thèse est adossé au projet *NewsEye*[1] dédié à l'analyse de la presse historique pour les humanités numériques. Au sein de ce projet a émergé le besoin d'analyser les événements et la manière dont on en parle : connaissons-nous vraiment tout des événements du passé. Pouvons-nous identifier des relais et réseaux propices à l'introduction de fausses informations ? Ce sont des questions auxquelles les travaux engagés dans cette thèse permettront de répondre.

Cette recherche ne se limite pas à exploiter des mécanismes déjà connus et éprouvés de détection et de suivi d'événements rapportés par la presse historique. Au-delà de ces contraintes évidentes, l'accent est mis sur deux éléments essentiels. Nous évaluons l'impact des documents historiques sur des algorithmes et processus adoptés pour la presse nativement numérique. Nous fournissons des procédés sobres en ressources informatiques. Ce travail produit une première ébauche de système adapté au contexte écologique dans lequel nous vivons.

Cette thèse s'associe aux travaux de recherche liés à la numérisation des fonds de presse historique. Elle a pour objectif de proposer différentes architectures répondant aux enjeux de suivi de mentions d'événements dans la presse. Nous proposons un cadre d'étude intégral comprenant d'abord la synthétisation de documents. Nous créons artificiellement des articles pour qu'ils présentent des spécificités communes aux documents historiques numérisés. Nous décrivons les processus d'analyse de ces derniers et une évaluation qui tient compte de différentes contraintes et des effets de ces dégradations. Nous cherchons à formuler par ce travail de premières recommandations pour orienter la recherche dédiée à l'analyse d'événements dans la presse ancienne. Nous proposons deux méthodes différentes, chacune adaptée pour répondre au même besoin. En tant qu'utilisateur ou utilisatrice d'archives numériques de presse, comme spécialiste ou simple particulier, comment puis-je découvrir tous les articles qui mentionnent ce même événement historique. Ces méthodes, ajustées à des contextes différents, sont plutôt classiques, par génération d'un modèle d'apprentissage ou plus orthodoxes, fonctionnant à la façon d'un moteur de recherche dédié aux événements du passé.

Toutes les données que nous avons générées, tous les algorithmes implémentés, les différents modèles et résultats sont librement partagés conjointement à cette thèse afin de s'intégrer dans les mouvements de science ouverte. Nous publions les codes sources sur l'archive publique *Software Heritage* et les autres données et documents sur la plate-forme *Zenodo*.

1. `https://newseye.eu` (archive sur `https://web.archive.org`)

Detection and Tracking of Events in Historical Press Documents

Abstract

Reconstructing the thread of events in the press, tracing the path of false information or identifying the point of origin of a piece of information, whether true or not, are the challenges of this new century. The increasing digitisation of digital contents is favourable to massive analyses of the press. The availability since the beginning of the 2010s of large press collections by libraries around the world opens another perspective. To be able to take another look at past events, to unravel the thread, to know both the chronologies and the geographical distribution becomes possible.

The analysis of historical documents is a singular process. Although similar in appearance to articles from the digital press, they differ in all respects : in form (document, data format) and in content (journalistic style, themes addressed, characters mentioned). This thesis is associated with the *NewsEye*[2] project dedicated to the analysis of the historical press for the digital humanities. Within this project, the need to analyse events and the way they are talked about emerged : do we know everything about past events ? Can we identify information relays conducive to the introduction of false information ? The work undertaken in this thesis will help answer.

This work is not limited to identifying and exploiting already known and proven mechanisms for detecting and tracking events mentioned and reported in the historical press. Beyond these obvious constraints, the focus is on two essential elements. We evaluate the impact of historical documents on algorithms, and processes necessarily oriented towards the natively digital press. We also provide low-computing resources processes, as a first draft towards an adaptation to the ecological context in which we live.

In this acceleration of the research work associated with the digitisation of historical press collections, this thesis aims at proposing different architectures answering the challenges of tracking mentions of events in the press. We propose a complete study framework including the synthesis of documents with specificities related to digitised historical press articles. We describe the analysis process of these documents and an evaluation that considers different constraints and the effects of these degradations. Through this work, we seek to formulate initial recommendations to guide research dedicated to the analysis of events in the historical press. We propose three methods, each adapted to answer the same need : as a user of digital press archives. As a researcher or as a simple private individual, how can I discover all the articles that mention this same historical event ? These three methods, adjusted to different contexts, are rather classical, by the generation of a learning model, or more orthodox, functioning as a search engine dedicated to past events.

With this thesis, we share all the data we have generated and every algorithm, the different models and results, to be integrated in the movements dedicated to open science. The source codes are published on the public archive *Software Heritage* and the other data and resources on the platform *Zenodo*.

2. `https://newseye.eu` (archive sur `https://web.archive.org`)

Remerciements

Après ces trois années de thèse, une portion de ma vie s'achève. J'ai appris, je me suis instruit, j'ai conçu, fabriqué, tout au long de ces années à l'école, au lycée puis à l'université. J'ai été entouré par une foule de personnes remarquables qu'il convient de citer. Sans la participation de chacun d'entre eux, peut-être n'aurais-je pas pu mener à bien ce travail et cette longue formation académique.

Je tiens tout d'abord à remercier mes deux encadrants, Antoine Doucet et Cyril Faucher qui m'ont accompagné durant ces trois années au laboratoire de recherche. Sans leur implication, mon travail aurait été impossible. Je tiens à remercier également chaleureusement Cyrille Suire qui m'a toujours guidé, sur le plan technique et scientifique. Au-delà de la recherche, vous avez contribué à forger mon intérêt pour l'informatique et la science. Vous m'avez accompagné depuis l'obtention de mon DUT et jusqu'à aujourd'hui. J'ai pris un grand plaisir à travailler sur cette thématique avec vous et j'espère sincèrement que cette collaboration pourra se poursuivre dans le futur. Merci enfin de m'avoir confié dès mes premières semaines, une partie de vos charges d'enseignement. J'ai pu m'épanouir dans une autre discipline, exigeante et complexe, mais passionnante. Je remercie également Paolo Rosso pour son accueil et ses précieux conseils lors de mon séjour à Valencía en Espagne.

Je tiens à remercier également les membres de mon jury qui ont pris le temps d'évaluer mes travaux de recherche. Merci à Cédrick Fairon et Els Lefever pour leurs rapports et à Karell Bertet, Benoit Favre ainsi que Glen Roe pour avoir examiné cette thèse. Je profite des ultimes modifications de ce manuscrit pour les remercier pour leurs questions, leurs encouragements et pour leurs mots, précieux, qui m'ont porté à la fin de ma soutenance et ensuite. Ces moments resteront à jamais gravés dans ma mémoire. Vous y avez également votre place.

Je remercie Jean-Loup Guillaume et à nouveau Benoit Favre pour leurs précieux conseils, ainsi que pour avoir accepté d'encadrer mon jury de comité de suivi individualisé (CSI). Vos conseils ont toujours une place dans mon esprit : j'essaie désormais chaque jour d'être un peu plus pragmatique, et d'avoir davantage confiance en moi.

Ce sont mes stages de Licence puis de Master qui m'ont donné le goût pour la science et la recherche. À ce titre, je remercie Manon Laëron, alors doctorante au LIENSS à La Rochelle et Samuel Nowakowski de l'Université de Lorraine à Nancy pour leur accompagnement et leurs précieux conseils. Mon travail de recherche n'aurait pas été le même sans la présence de mes chers camarades de thèse. Beatriz d'abord, qui m'a transmis une partie de la fibre espagnole. C'est grâce à toi que j'ai pu aller à València (Espagne). David G-G et Silvia C. que j'ai rencontrés là-bas, je vous remercie pour l'aide que vous m'avez apportée. Damien, nous avons mis du temps pour nous apprécier, mais c'est sans aucun regret. Merci pour ces moments de rire partagés. Je tiens à remercier aussi quelques élèves de DUT. Je pense en particulier à Léo qui restera gravé comme mon premier souvenir peu glorieux d'enseignement, et le groupe composé de Maxime, Oscar et Virgile qui en est le dernier. À tous les autres, merci pour l'intérêt que vous avez montré et pour l'enrichissement que vous m'avez apporté.

Sur un plan plus personnel, je remercie bien évidemment Benjamin B. et Louis B., deux amis fidèles auprès de qui je veux renouveler mon immense affection. Vous étiez toujours présents, malgré les tempêtes. Ma motivation et mon travail auraient été bien différents sans votre présence à mes côtés. Merci encore, infiniment. C'est grâce à Marine C. que je suis « tombé » dans l'informatique et c'est elle qui m'a forcé à réordonner mes vœux post-baccalauréat : je suis donc allé à l'IUT de La Rochelle. Merci à toi, tu as visiblement bien fait. Je tiens à remercier également Aurélien R. : je ne te comprenais pas toujours lorsque nous nous sommes connus. Tu m'as pourtant apporté des connaissances en informatique que je n'aurais obtenues de personne d'autre. Je remercie enfin mes incroyables amis d'enfance et le « noyau dur », Nikita F., Célia B., Clothilde I. et Valentin B.. Vous avez tous joué un rôle déterminant dans ma vie, vous avez contribué, chacun et chacune, à me faire devenir qui je suis. Sachez que dans les temps moroses au travail, votre petite voix d'encouragement résonnait toujours dans mon esprit. C'est à vous que je pense quand je regarde en arrière et nous revois au lycée. Je suis fier de vous et vous remercie de m'avoir tant appris. Je termine enfin par des

amis que cette thèse m'a fait connaître, Nolwenn M., Olivia B., Raphaël P. et Thomas B.. Je vous souhaite le meilleur dans vos études et dans votre vie. Merci, Thomas, pour ton travail. Nous savons que ton aide a été précieuse pour cette thèse. Antoine B. enfin, merci pour ton soutien sans failles, ton énergie, tes encouragements et ta patience, tu as toi aussi fortement contribué à la réussite de cette thèse. C'est enfin fini, j'espère que tu es soulagé.

À mes parents, j'adresse les plus sincères remerciements. Merci de m'avoir toujours soutenu et encouragé dans la voie que je voulais suivre, dans mes bizarreries associatives et pour tout ce que je vous ai fait subir. Vous m'avez permis de m'épanouir et j'espère que vous êtes fiers du petit bonhomme que vous avez élevé. Merci à vous deux.

Je remercie enfin les membres du laboratoire L3i de l'université ainsi que les membres du département informatique de l'IUT de La Rochelle. Vous m'avez tous fait confiance, enseigné et m'avez guidé. Je vous en serai à jamais reconnaissant. J'adresse un remerciement tout spécial à Mélanie, qui m'a aidé à commencer ma thèse puis à Alain, Christophe, Jean-Loup, Karell, Mourad, Petra, Ronan et Yacine. Merci à vous.

À ma grand-mère Monique.

Table des matières

Table des figures

Liste des tableaux

Note sur la rédaction

Dans cet ouvrage, j'ai utilisé le pronom dit « *nous de modestie* » pour rappeler que, bien qu'étant l'unique auteur de ce document et de ce qui y est décrit, celui-ci n'aurait jamais existé sans les nombreux travaux précédents et qui ont été réalisés par la communauté scientifique dans son ensemble. C'est une manière de rendre hommage à tous les auteurs qui m'ont précédé ainsi qu'à leurs recherches.

Vous y trouverez peut-être d'étranges accords du *nous* au singulier, conséquence directe de cet usage.

Chapitre 1

Introduction

LE JAPON RAVAGÉ
par un tremblement de terre

OSAKA, 1er septembre. — Aujourd'hui, à midi, on a ressenti un termblement de terre, qui a duré plus de six minutes, accompagné d'un mouvement de bas en haut. Toutes les horloges ont été arrêtées.

Une seconde secousse sismique a été ressentie à 2 h. 25 cet après-midi. D'après l'Observatoire d'Osaka, le centre du tremblement a été probablement la péninsule Inzu.

Un sismographe a enregistré des secousses qui ont duré presque une heure et demie.

Les dégâts à Tokio, Yokohama, Yokosuka sont, paraît-il, importants.

La ligne de chemins de fer de To-Raido a été sérieusement endommagée en divers endroits. Non seulement les communications téléphoniques, mais aussi les communications télégraphiques sont totalement interrompues entre Osaka et Tokio.

(La secousse avait été enregistrée par différents observatoires européens, notamment par celui d'Uccle (Belgique), et par celui du parc Saint-Maur qui, d'après l'amplitude des oscillations, l'avait située à une dizaine de mille kilomètres, et l'estimait d'une intensité au moins égale à celle qui avait été ressentie au Chili l'an dernier.)

L'incendie à Yokohama

LONDRES, 1er septembre. — Un radiotélégramme de San-Francisco annonce que la ville de Yokohama est pour ainsi dire entièrement détruite à la suite d'un tremblement de terre. Il y a de nombreuses victimes.

NAGASAKI, 1er septembre. — Un radiotélégramme annonce qu'un incendie a éclaté à Yokohama et que les habitants se réfugient à bord du *London-Maru* et du *Paris-Naval*, qui ont jeté l'ancre dans le port.

On est sans nouvelles de Tokio

OSAKA, 1er septembre. — L'interruption complète des communications télégraphiques, téléphoniques et ferroviaires prouve combien a été violente la secousse de tremblement de terre à Tokio, dont jusqu'ici on n'a encore reçu aucune nouvelle.

Toutes les voies ferrées aboutissant à Tokio sont désorganisées dans un périmètre d'une centaine de milles autour de la ville.

SAN-FRANCISCO, 1er septembre. — On annonce que toutes les communications radiotélégraphiques d'Amérique sont interrompues avec le Japon.

La dernière dépêche reçue du Japon, et qui est parvenue à 9 heures ce matin, rapportait que toutes les lignes télégraphiques terrestres du Japon étaient apparemment coupées.

Le Figaro, 69me Année, 3me Série - N° 245 - Dimanche 2 septembre 1923.
Source : gallica.bnf.fr/BnF

Sommaire

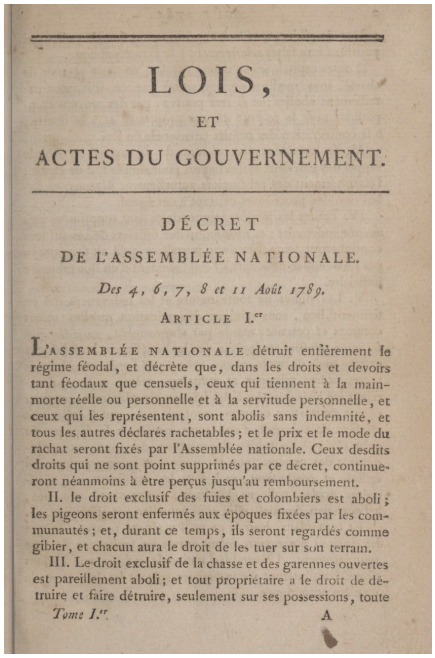
1.1 Contexte

La numérisation des fonds patrimoniaux est une activité en expansion depuis l'apparition des sciences du numérique à la fin du XX^{e} siècle [Fra21]. Des manuscrits de plusieurs centaines d'années, des gravures, images négatives, registres patrimoniaux de l'Église ou de l'administration, presse, tous ces objets sont numérisables. La numérisation, utilisée par différents publics, scientifiques ou amateurs, permet une préservation à long terme des supports et limite leur manipulation, donc leur dégradation. Elles sont destinées à toutes et à tous et sont des éléments de notre culture commune. La numérisation et l'archivage numérique des documents est une problématique traitée dans le monde entier. En ce sens, le ministère de la Culture en France encourage et aide les organisations, bibliothèques et archives publiques à numériser leurs contenus [Min22]. Des archives numérisées sont aujourd'hui disponibles pour toute la France et accessibles sur la plateforme `francearchives.fr`. Soutenu par l'État, cet organisme accompagne et édite des recommandations à destination des acteurs voulant numériser leurs fonds historiques. En parallèle, la communauté scientifique s'organise pour collecter, traiter, indexer, publier et exploiter des numérisations d'objets historiques [Die20]. Le développement des humanités numériques a permis de renouveler certaines des méthodes employées et de remettre en question l'impact de l'accès numérique à une documentation étendue pour les sciences humaines et sociales. Cet impact porte principalement sur les méthodes et non sur les résultats par l'accès à de grands volumes de documentation, par la réduction des frais de recherche ou par l'automatisation de certains processus scientifiques.

Le volume des données archivées est monumental et extrêmement varié. Les actes de l'état civil et les tables décennales apparues après la Révolution française de 1789 accompagnent les registres paroissiaux de baptêmes, mariages et décès. Les premiers cadastres, les cartes historiques et militaires occupent des rayonnages entiers d'archives départementales, aux côtés de fonds iconographiques, de cartes postales ou de photographies. L'apparition de la télévision a fait émerger le besoin d'archiver également les programmes audiovisuels, menant à la création de l'INA en France, l'Institut National de l'Audiovisuel. Cet organisme collecte et archive des programmes télévisuels sur sa plate-forme dédiée, l'Inathèque [1]. L'émergence du Web, média transnational utilisé entre autres pour diffuser de l'information, a fait naître une nouvelle problématique d'archivage

1. `https://inatheque.ina.fr/` (archive sur `https://web.archive.org`)

du Web [2] au niveau mondial, modifiant les voies classiques d'accès à la documentation.



LOIS,

ET

ACTES DU GOUVERNEMENT.

DÉCRET

DE L'ASSEMBLÉE NATIONALE.

Des 4, 6, 7, 8 et 11 Août 1789.

ARTICLE I.er

L'ASSEMBLÉE NATIONALE détruit entièrement le régime féodal, et décrète que, dans les droits et devoirs tant féodaux que censuels, ceux qui tiennent à la main-morte réelle ou personnelle et à la servitude personnelle, et ceux qui les représentent, sont abolis sans indemnité, et tous les autres déclarés rachetables; et le prix et le mode du rachat seront fixés par l'Assemblée nationale. Ceux desdits droits qui ne sont point supprimés par ce décret, continueront néanmoins à être perçus jusqu'au remboursement.

II. le droit exclusif des fuies et colombiers est aboli; les pigeons seront enfermés aux époques fixées par les communautés; et, durant ce temps, ils seront regardés comme gibier, et chacun aura le droit de les tuer sur son terrain.

III. Le droit exclusif de la chasse et des garennes ouvertes est pareillement aboli; et tout propriétaire a le droit de détruire et faire détruire, seulement sur ses possessions, toute

Tome I.er A

(a) Extrait d'un décret de l'Assemblée nationale déclarant l'abolition des privilèges le 4 août 1789.
Source : `gallica.bnf.fr`/BnF

LE GAULOIS

RÉDACTION

ABONNEMENTS

ÉPITRES BORDELAISES

(b) Première page du journal Le Gaulois, publié le 19 mai 1867.
Source : `gallica.bnf.fr`/BnF

FIGURE 1.1 – Deux documents provenant du fond ancien numérisé de la BnF.

En France, la Bibliothèque Nationale de France (BnF) met à disposition du public des numérisations de livres, de journaux et recueils, à travers une plate-forme publique, *Gallica* [3]. *RetroNews* [4] est quant à lui dédié spécifiquement à la diffusion et à l'analyse des titres de presse. Les deux figures présentées en 1.1a et 1.1b sont extraites de ces plate-formes. La BnF conserve des manuscrits et des reproductions de documents extrêmement variés, couvrant plusieurs siècles de publication. La figure 1.1b représente une page du journal Le Gaulois publié dans la seconde moitié du XIXe siècle.

La presse offre un aperçu unique des événements du passé. La compréhension des événements historiques par l'analyse de la presse est une occasion d'interroger nos connaissances historiques. Retracer le parcours des informations diffusées dans la presse peut entraîner des conséquences sur la manière dont les faits historiques sont interprétés de nos jours. Des malversations, des fraudes à l'information ou l'introduction de faux détails et d'informations trompeuses sont possibles dans tous les maillons de la chaîne de

2. `https://web.archive.org`
3. `https://gallica.bnf.fr/` (archive sur `https://web.archive.org`)
4. `https://www.retronews.fr/` (archive sur `https://web.archive.org`)

diffusion [Pin20]. Retracer ce parcours est non seulement un moyen efficace de reconstruire la chronologie d'un événement, mais également un moyen de vérifier la chaîne de l'information [Oiv+19]. Ces travaux, lorsque réalisés manuellement, sont nécessairement limités. Le temps et la quantité de journaux contraignent le nombre de documents et d'événements analysables [Bol19]. L'introduction des sciences du numérique, en automatisant le processus d'indexation, peut simplifier ou accélérer les analyses. L'exploration des événements mentionnés au sein de milliers de documents devient alors possible.

Pour faire face à l'énorme volume de données numérisées archivées, les traitements numériques et informatiques sont indispensables. L'analyse des événements est un point d'entrée possible pour mener une recherche historique et ces événements sont presque toujours rapportés par la presse. Du plus petit fait divers à un événement majeur comme le tremblement de terre qui, en 1923, a fait trembler le Japon, tous sont rapportés par la presse. C'est ce type spécifique de document, la presse historique numérisée, que nous allons étudier tout au long de cet ouvrage à travers un élément particulier, les événements qui y sont décrits.

Hors de la recherche historique, l'idée d'analyser des événements rapportés dans la presse en les ordonnant et en les classant n'est pas limitée à la presse ancienne et trouve des applications dans nos sociétés contemporaines. Nous le verrons par la suite, les projets consacrés à la détection et au suivi des événements mentionnés dans la presse sont nombreux. Certains, comme *Europe Media Monitor* (*EMM*) [SPV09] ont donné lieu à de réelles applications informatiques comme la détection d'événements émergents ou de signaux faibles. Le programme *NewsBrief*[5] est issu des travaux de recherche. Il permet à tout un chacun de suivre l'évolution des événements dans la presse numérique. Sur son interface, montrée en figure 1.2, on retrouve un thème, ici défini par le premier article publié en lien avec l'événement montré. Viennent ensuite tous les articles, triés par ordre de parution : tous sont liés au même événement. Le graphe présente l'intensité des publications dans le temps : il identifie les temps forts de chaque événement.

1.1.1 Les documents historiques et les articles de presse

Le siècle qui débute en 1850 et se termine en 1950 est considéré comme l'âge d'or de la presse écrite en France et dans le monde. La transmission sans fil et la télévision n'existent pas. Les organes de presse sont en situation de monopole pour relayer l'information [CHK12]. Cette période est marquée par la libéralisation politique qui aboutit, en France, à une loi dite « sur la liberté de la presse », promulguée le 29 juillet 1881 [81]. L'augmentation de la taille du lectorat a entraîné une profonde mutation de la presse. Elle s'est transformée en entreprise commerciale. C'est à cette époque qu'apparaît effectivement la profession de journaliste dont le rôle est de rapporter les faits observés et de les narrer pour le public. La qualité des écrits et contenus n'est pas toujours au rendez-vous. Les rédactions misent parfois sur le sensationnel pour publier du contenu [Pin20]. Bien avant déjà, Honoré de Balzac s'inspire de ces transformations pour son roman critique Les Illusions Perdues, publié entre 1834 et 1843. Des recherches histo-

5. `https://emm.newsbrief.eu/` (archive sur `https://web.archive.org`)

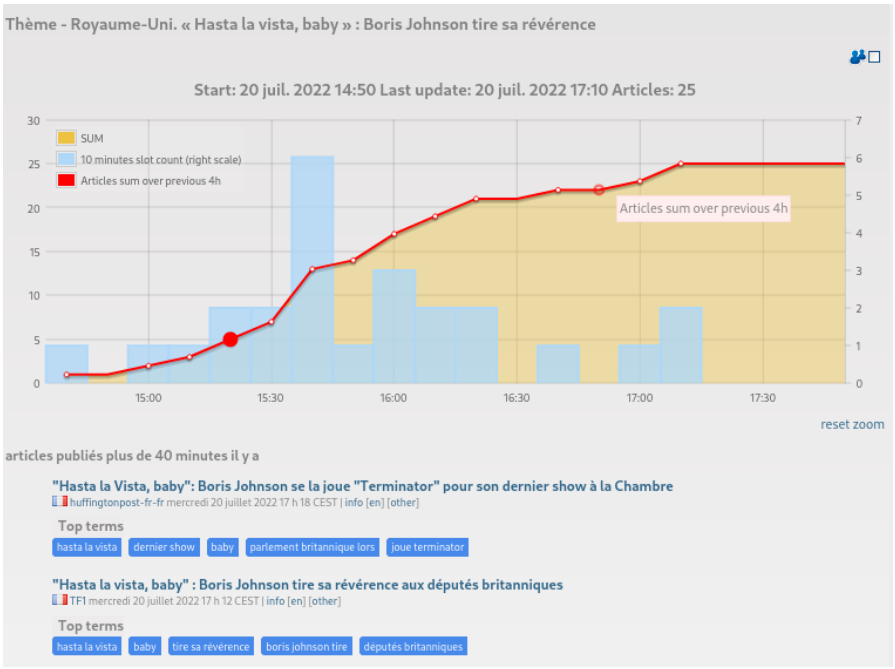


FIGURE 1.2 – Capture d'écran du programme *NewsBrief*, développé en lien avec le projet *Europe Media Monitor* (*EMM*) [SPV09]

riques plus récentes mentionnent à ce propos la place faite aux fausses informations ou aux histoires virales relayées par la presse de cette époque [Pin20]. Le déclin s'amorce dans la première moitié du XX^e^ siècle où les actualités sont progressivement partagées sans fil par télégraphe. L'apparition de la radio vient confirmer cet état de fait dans les années 1920.

Les articles de presse sont rédigés dans un style spécifique que l'on peut qualifier de « journalistique ». Chaque auteur se doit d'expliquer son sujet en répondant à des questions de base. Selon l'Institut Supérieur de Formation au Journalisme, la règle de « 5W » est une méthode empirique de rédaction qui consiste à répondre à cinq questions dans le contenu d'un article [ISF21]. Les « W » valent pour les termes anglais *what* (quoi), *why* (pourquoi), *who* (qui), *when* (quand) et *where* (où). Cette stratégie rédactionnelle est une pratique commune qui a peu évolué depuis le milieu du XIX^e^ siècle et la professionnalisation du métier de journaliste.

Les articles de presse font partie intégrante des fonds documentaires conservés par les bibliothèques. Ces articles sont une mine d'or pour les scientifiques des sciences humaines et sociales (SHS) : ils reflètent les opinions à propos d'événements passés ou éclairent des actions politiques vécues par les individus à ces époques. Tout le monde a accès à des reproductions d'articles de presse et à des transcriptions réalisées par ordinateur. La difficulté d'analyse de ces articles historiques notamment par les chercheurs et chercheuses en sciences humaines et sociales a mené à la création de nombreux projets de recherche que nous verrons par la suite. La numérisation, la transcription et l'ap-

plication de méthodes numériques d'extraction d'information simplifient l'analyse et la compréhension de ces documents. Cependant, la qualité des acquisitions ou le volume de données posent, entre autres, de nouveaux enjeux aux sciences du numérique.

1.1.2 Les événements en presse écrite

La presse informe le public à propos des événements qui ont lieu dans le monde. Elle mentionne des événements d'ampleur aussi bien locale qu'internationale. À titre d'exemple, l'attentat qui a entraîné la mort de l'Archiduc François Ferdinand d'Autriche et de sa femme à Sarajevo s'est répandu à grande échelle et dans le monde à travers les réseaux de presse. Un extrait de Le Figaro en figure 1.3 relate cet événement. La presse regorge cependant d'événements de plus faible audience : faits divers généraux, actualités régionales, comptes rendus de débats politiques, etc.

DOUBLE ATTENTAT

L'Archiduc héritier d'Autriche et sa femme assassinés en Bosnie

Une bombe et deux coups de revolver

Pendant des semaines, tout le peuple austro-hongrois et l'Europe entière ont attendu avec anxiété, matin et soir, le bulletin de santé de l'empereur François-Joseph. Les médecins, eux-mêmes, pendant longtemps ne cachaient pas leur inquiétude. D'un jour à l'autre on attendait l'issue fatale de la maladie dont était atteint ce souverain de quatre-vingt-quatre ans.

François-Joseph, complètement rétabli, est parti samedi pour Ischl, où chaque année il va faire une cure et se livre à son sport favori, la chasse ; son neveu qui était appelé à lui succéder, dont l'avènement au trône semblait imminent, est mort.

Il a été tué hier à Serajevo, capitale de la Bosnie, qu'il a tant contribué à faire autrichienne. On le disait de tendances slavophiles et c'est sous la balle d'un Slave qu'il est tombé.

provincial a été grièvement blessé par les éclats de la bombe.

Evidence du complot

Le complot est, en effet, évident, et il n'est pas douteux que ce complot a été préparé par la jeunesse nationaliste de Serajevo. L'âge des deux meurtriers le prouve, et quant à la réalité du complot elle paraît nettement établie d'abord par le fait que Prinzip a tiré presque immédiatement après que Cabrinovitch eut jeté sa bombe, et puis par cet autre fait qu'une autre bombe a été trouvée à peu de distance de l'endroit où a eu lieu le second attentat. Il ne paraît pas douteux qu'elle a été jetée par un troisième complice qui guettait le passage de l'archiduc et qui s'est enfui lorsqu'il a su que le crime préparé était devenu inutile.

FIGURE 1.3 – Numérisation de la première page du journal Le Figaro, publié le 29 juin 1914.
Source : `gallica.bnf.fr`/BnF

L'analyse de la presse historique ou des événements qu'elle mentionne est à l'origine

de découvertes sur la propagation de l'information elle-même [Bol19]. La presse rapporte les événements sous diverses formes, que l'on peut classer en trois catégories [Lej13]. La première est la dépêche télégraphique : à cette époque, l'information qui parvient aux agences de presse est succincte et publiée rapidement. Puis viennent les articles longs dans un second temps, avec davantage d'informations. Enfin, des articles d'opinion peuvent être publiés, des éditoriaux et discussions à propos des événements qui ont eu lieu.

Dans cet ouvrage, nous nous intéressons particulièrement aux événements historiques en tentant de les définir dans notre contexte informatique. Les contours de cette notion sont flous, nous tenterons d'en proposer une définition adaptée à la problématique du suivi d'événements. La presse, support principal sur lequel ces événements étaient mentionnés, est l'objet de notre étude. Nous nous intéressons particulièrement à la manière dont nous pouvons identifier ces événements et les tracer, les suivre pour comprendre les modalités de diffusion entre les différents journaux et peut-être les différents pays. Par ce travail de recherche, nous souhaitons faire évoluer les connaissances actuelles dans ce domaine. Ces travaux débouchent naturellement sur quelques contributions scientifiques.

1.2 Contribution de la thèse

Les contributions scientifiques de cette thèse s'inscrivent à plusieurs niveaux. L'objectif de ce travail est de fournir une approche scientifique à la question du suivi des événements mentionnés dans la presse, mais également une réalisation technique.

Les questions de recherche soulevées dans cette thèse et les contributions sont les suivantes :

— **Publication d'algorithmes pour le suivi de mentions d'événements**. Ce travail de recherche s'inscrit au sein d'une communauté qui œuvre à la détection et au suivi d'événements dans la presse. Peu de travaux s'intéressent au cas particulier du suivi dans la presse ancienne en faisant face aux difficultés inhérentes à ces supports : bruit, difficulté à extraire les articles, etc. Nous publions l'ensemble des algorithmes réemployés et issus de l'état de l'art ainsi que ceux conçus de toute pièce. Si des entraînements de modèles sont nécessaires, nous proposons également d'en diffuser non seulement les protocoles d'apprentissage, mais aussi tous les logiciels d'entraînement.

— **Étude de la qualité des algorithmes dans le contexte particulier des documents anciens**. L'analyse de documents historiques impose l'existence de corpus d'articles de presse ancienne numérisés et prétraités. Comme nous allons le voir, les prétraitements consistent à extraire les articles des pages numérisées de journaux puis le texte de ces articles. En présence de documents historiques, ces deux tâches introduisent des erreurs. Le cas spécifique du suivi des événements sur ce type de document n'a pas fait jusqu'alors l'objet d'études dédiées. En effet, les articles de presse historiques annotés pour notre objectif de recherche n'existent pas. Pour contrer cette limitation, nous proposons de synthétiser des erreurs typiques présentes dans des documents historiques et d'appliquer les algo-

rithmes retenus sur chacun d'eux. Cette première analyse permettra d'entrevoir les pistes pour améliorer les algorithmes en attendant la publication de données annotées et exploitables dans ce contexte.

— **Définition et automatisation de la représentation d'événements historiques à partir de sources ouvertes**. La définition des événements dans la littérature scientifique consacrée au suivi des informations de presse est sujette à débat. Leur représentation, qui découle logiquement de ces définitions l'est tout autant. Nous proposons dans cette thèse, après avoir pris soin de définir la notion d'événement adaptée à la problématique de recherche, un moyen de les représenter numériquement. Nous nous basons sur une ontologie existante et utilisons les ressources et connaissances disponibles sur l'encyclopédie Wikipédia ou des graphes de connaissance dédiés à l'analyse d'événements.

— **Analyse des jeux de données adaptés au suivi d'événements**. Les différents corpus utilisés pour le suivi d'événements depuis une vingtaine d'années ont tous des spécificités qui peuvent restreindre leur utilisation : limitations d'usage ou indisponibilité des corpus utilisés dans des publications passées. Nous proposons dans cette thèse de ne traiter que des corpus de données utilisés par la communauté, publics et disponibles. Nous explorons leurs contenus pour identifier des biais qui pourraient affecter les analyses que nous produisons. Comme nous l'expliquerons, s'intéresser aux données publiques favorise la comparaison des résultats et la réutilisation de nos travaux.

Enfin, cette thèse s'inscrit dans le mouvement de la science ouverte. Les décisions prises sur le choix des données à utiliser le sont toutes avec cet objectif. Les publications issues de ce projet de recherche sont partagées sous des licences permissives et archivées publiquement en ligne. De même, nous publions toutes les données intermédiaires, graphes, modèles, algorithmes, connectés entre eux et décrits par des métadonnées. Nous utilisons les plates-formes *Zenodo* [EO13] pour l'archivage et la publication des données, *Software Heritage* [DZ17] pour les codes sources et leur référencement [6] et l'archive ouverte *HAL* [7] pour le dépôt des documents. Nous voulons également inscrire ce projet dans son époque et traitons des problématiques de sobriété énergétique. Le numérique contribue fortement aux émissions de gaz à effet de serre et est consommateur de ressources minières [FDM13]. Nous contributions, à travers cette étude, à définir une méthode de suivi de mention d'événements n'exigeant pas d'importantes ressources documentaires ou numériques.

1.3 Structure de la thèse

Ce manuscrit de thèse est divisé en quatre chapitres de contributions, précédés d'une introduction et suivis d'une conclusion. Le chapitre 2 présente un état de l'art et introduit

6. Les travaux produits durant cette thèse se sont basés sur le système d'exploitation Red Hat® Entreprise Linux® 8. Les versions des logiciels utilisés sont précisées lorsque cela est nécessaire afin de faciliter la reproduction des expériences présentées.

7. `https://hal.archives-ouvertes.fr/`

les représentations formelles d'événements en traitement automatique des langues et dans les documents de presse. Il couvre également la question du suivi des événements mentionnés dans la presse écrite. Dans ces cas-là, le fil des événements est reconstruit à partir des informations rapportées dans les textes. Nous exposons enfin les propriétés des corpus sélectionnés en mettant en avant les éventuels biais qui affectent nos travaux.

Le chapitre 3 apporte un éclairage sur les spécificités observables dans les documents historiques numérisés. Nous détaillons un cadre expérimental pour simuler celles-ci, étape essentielle lorsque des documents historiques annotés manquent. Nous proposons également, au sein de ce chapitre, une stratégie pour annoter les documents historiques numérisés. Ces annotations sont indispensables pour concevoir et évaluer des algorithmes de suivi d'événements.

Les deux chapitres 4 puis 5 introduisent deux approches distinctes pour suivre des événements dans la presse. Dans le chapitre 4, nous analysons des masses d'articles de presse pour former des groupes d'articles qui décrivent un même événement. Au chapitre 5, nous nous appuyons sur des représentations d'événements obtenues à partir de sources ouvertes. Depuis cette représentation d'événement, nous proposons un moteur de recherche spécialisé dans l'analyse de la presse. Dans les deux cas, nous réalisons une étude pour évaluer l'impact des dégradations des documents historiques. Nous recherchons quelle dégradation affecte davantage le processus et proposons des pistes de solutions pour les contourner.

Enfin nous terminons, au chapitre 6, par un bilan des travaux menés ainsi que des contributions de ce travail. Nous énonçons enfin les perspectives ouvertes par cette thèse.

Chapitre 2

État de l'art

Sommaire

Comme nous l'avons évoqué en introduction de ce document, nous nous intéressons à deux concepts distincts : les événements d'abord et la propagation des mentions de ces événements dans la presse ensuite. Une mention d'un événement est un article, un document ou une phrase qui traite et qui décrit cet événement. Pour comprendre comment ils se propagent et créer des algorithmes spécifiques pour cette tâche, la question même de l'événement est en jeu. Comprendre comment un événement s'exprime formellement est un préalable.

Nous nous intéresserons premièrement à cette représentation d'événement dans la presse et proposerons une analyse à la section 2.1. À partir de toutes les campagnes d'évaluation qui ont eu lieu depuis l'émergence de cette discipline, nous proposerons une définition de ce concept, adaptée à notre cas d'étude. Comme nous allons le montrer, la notion d'événement porte de nombreuses ambiguïtés : en fonction de la discipline ou du but, les définitions en langage naturel et formel varient. Dans un second temps, à la section 2.2 nous explorerons la problématique du suivi d'événements dans la presse. À notre connaissance, les travaux spécialement dédiés à l'analyse de documents historiques sont inexistants. Nous nous intéresserons alors davantage à la presse récente, cherchant les moyens d'appliquer ces techniques à la presse historique numérisée.

Enfin, pour produire une évaluation de nos travaux, des données de presse sont nécessaires. Nous mettrons en avant nos choix de données en section 2.3 et, à travers une analyse exploratoire, nous présenterons les obstacles que ces données posent dans le cadre de notre étude.

2.1 Le concept d'événement en traitement des langues

La notion d'événement en traitement du langage naturel est une source de débats infinis entre spécialistes de divers domaines [ST17]. Cette notion diffère selon les disciplines et les buts. Par nature, le concept d'événement adopte différentes structures, différents types et arguments [XW19]. Cette vision ne capture pourtant pas la variété des événements. Ce champ de recherche a fait l'objet d'études approfondies au cours des décennies passées, arborant différents objectifs. Les événements portent une longue histoire commune en sciences humaines, en philosophie et en sciences du numérique, plus particulièrement en traitement automatisé du langage.

L'étude des événements remonte à Aristote qui en a proposé une typologie basée sur la sémantique des verbes ou les structures temporelles internes aux textes. Cette proposition a influencé la recherche dans la littérature philosophique [Min75 ; Ken03 ; Ryl09]. L'idée que les phrases verbales dans les langues humaines peuvent être considérées comme reflétant un événement a été un point de départ pour les linguistes [Ven67 ; Bac86 ; Dow12]. Leurs articles, très influents, ont marqué le début d'une classification des événements basée sur la cognition, largement acceptée dans la littérature de sémantique lexicale. Les événements peuvent être groupés selon des paramètres temporels : la durée, la terminaison et leur structure interne [Ven57]. Tout verbe exprimé dans une langue naturelle peut être caractérisé dans l'un des trois types d'événements de base : états, processus ou transitions. Dans la nomenclature de Vendler [Ven57 ; Ven67], les verbes

peuvent dénoter des états (événements statiques sans finalité, par exemple « savoir » ou « croire »), des activités (processus dynamiques sans finalité, par exemple « courir »), des réalisations (événements instantanés avec une finalité, tels que « faire » ou « remarquer ») ou des accomplissements (processus qui ont une finalité et sont graduels comme « écrire un livre »).

Les théories philosophiques nées après ont voulu structurer la perception et la cognition que chacun ou chacune se fait du monde qui l'entoure. La théorie des cadres [Min75] formalise des situations appelées cadres (*frame* dans la version originale) organisés en niveaux. Certains représentent la « vérité » (sous forme de prédicats considérés comme réels et vrais par rapport au contexte) et d'autres représentent des créneaux, des intervalles temporels, qui doivent être remplis avec des instances et des données spécifiques. La même année est ajoutée la notion de scénario (*script*) [SA75], une structure décrivant une séquence d'événements, comparable à un scénario de tournage de film. Les scripts représentent des situations de la vie quotidienne, stéréotypées, comme des séquences d'actions dans des contextes particuliers.

Par la suite, avec l'avènement des sciences du numérique, de nombreux projets de recherche se sont structurés et de nouvelles représentations plus techniques ont vu le jour.

Campagnes d'évaluation, projets et jalons

La recherche sur l'extraction d'événements dans du texte a été stimulée par une longue histoire qui a commencé avec les *Message Understanding Conferences* (*MUC*) [GS96] de 1987 à 1998 sous les auspices du gouvernement américain (*ARPA*/*DARPA* [1]). Les évaluations *MUC* ont été les premières à s'engager dans le développement de métriques et d'algorithmes pour soutenir la conception de technologies émergentes pour l'extraction d'information. Au milieu des années 1990, les évaluations *MUC* ont commencé à publier des corpus de données et des définitions de tâches, en plus de fournir un logiciel d'annotation entièrement automatisé. Les travaux se sont concentrés sur l'extraction de connaissances à partir de messages catégorisés (instructions militaires, dépêches de presse). On peut citer les tâches de reconnaissance d'entités nommées (*Named Entity Recognition*, *NER*) [Bor+98] de liaison (*Entity Linking*, *EL*), d'extraction de relations (*Relation Extraction*, *RE*) et d'extraction d'événements (*Event Extraction*, *EE*) [GS96]. Les jeux de données pour *MUC-3* et *MUC-4* sont disponibles. Les textes utilisés pour *MUC-6* [Sun95] et *MUC-7* [Chi98] sont protégés par des droits d'auteur spécifiques et ne sont disponibles que par le biais du *Linguistic Data Consortium* (*LDC*) [2]. Les premiers jeux de données *MUC* contenaient des dépêches d'actualités couvrant les activités terroristes en Amérique latine. *MUC-6* et *MUC-7* avaient un spectre plus large, se concentrant sur des articles de presse aux sujets génériques.

Avant la fin des campagnes *MUC* en 1996, le projet *Topic Detection and Tracking* (*TDT*) également parrainé par la *DARPA* s'est concentré sur des objectifs différents

1. *Defense Advanced Research Projects Agency*, agence du département de la Défense des États-Unis chargée de la recherche et du développement de nouvelles technologies destinées à un usage militaire.

2. `https://www.ldc.upenn.edu/` (archive sur `https://web.archive.org`)

et a duré jusqu'en 2001 [All02b]. Alors que les *MUC*s se focalisaient sur la détection d'événements dans des documents, *TDT* a adopté une approche différente et a donné une nouvelle définition des événements mentionnés dans les textes. Les trois corpus produits par le projet *TDT* [Lin02 ; Cie+02] contiennent plus d'un millier de thématiques issues de diverses sources telles que des dépêches et des transcriptions de programmes audio. Les contenus sont rédigés en anglais et en mandarin.

En 2002, émerge de la communauté l'idée de définir un langage de spécification riche pour ordonner les événements dans le temps. Il se nomme *TimeML*. Ce projet faisait partie du programme *AQUAINT*. Ce dernier est dédié à la recherche d'information dans des corpus de données temporelles structurés en différentes langues et formats [NIS10]. Il s'intéresse aux systèmes de réponse aux questions (*Questin Answering*, *QA*) [WJD03 ; Pus+03]. *TimeML* adopte une définition simple des événements qui met l'accent sur les relations temporelles entre les éléments de la phrase [Spr18]. Une ontologie temporelle spécifique est créée à partir de ces travaux [FZ02]. Le schéma d'annotation étend et poursuit les objectifs des systèmes de *QA* explorés lors des conférences *TREC* (*Text Retrieval Conferences*) [Voo99 ; Voo00]. Cette campagne a conduit à la création du corpus *TIMEBANK* composé de 300 articles annotés [GS03]. Il contient des articles d'actualité et des transcriptions de journaux télévisés.

Deux ans plus tard et de 2004 à 2008 dans la continuité de *MUC*, émerge le projet *Automated Content Extraction* (*ACE*) mené par le gouvernement américain et soutenu par le *National Institute of Standards and Technology* (*NIST*). Trois tâches d'extraction sont développées au sein de ce projet : l'extraction d'entités nommées (*ACE* 1 à 5), de relations entre les entités (*ACE* 3 à 5) et l'extraction d'événements (*ACE* 7) [Dod+04]. Une riche taxonomie a également été développée afin de classer les événements selon leur type : *Die* (mort), *Conflict* (conflit), etc. Le *Linguistic Data Consortium* développe quelques directives d'annotation d'événements, construit des corpus et fourni des ressources linguistiques pour soutenir le projet *ACE* [Lin05]. Les jeux de données comprennent des transcriptions d'émissions, des dépêches et articles de journaux en anglais, en mandarin et en arabe [Lin05].

À partir de 2008, les travaux engagés dans *ACE* se poursuivent au sein des *Text Analysis Conferences* (*TAC*) [3]. Les ateliers d'évaluation de *TAC* introduisent une tâche de suivi des événements. Elle propose un ensemble de projets axés sur l'extraction d'événements à partir de textes, avec l'identification des participants et des relations qui les unissent. Ces conférences organisées par le *NIST* ont traité une variété de tâches, dont certaines centrées sur la détection d'événement, de 2014 à 2017 [Agu+14 ; Son+15].

TempEval consistait en des tâches d'évaluation faisant partie de la *Semantic Evaluation* (*SemEval*) et des ateliers axés sur la sémantique numérique. Trois tâches [Ver+07 ; Ver+10 ; UzZ+13] se sont intéressées à l'identification et l'extraction des événements et des relations temporelles entre ceux-ci, de 2007 à 2013. Le langage d'annotation *TempEval* était une extension de *TimeML* et le corpus *TIMEBANK* a été utilisé. Les travaux menés au sein du projet *TempEval* consistaient à annoter les relations temporelles entre les termes exprimant le temps et les événements.

3. `https://tac.nist.gov/about` (archive sur `https://web.archive.org`)

Durant toutes ces années et dans la continuité des premiers travaux sur les ontologies liés aux événements émergeants engagés par *TimeML* [FZ02], de nombreux autres projets et axes de recherche ont choisi d'adopter une définition d'événement structurée par des ontologies [LHP01 ; Rai+07 ; RA07 ; Sch+09 ; vHag+11 ; GD19]. Dans ces projets, répartis sur une vingtaine d'années, différentes structures d'événements sont proposées, parfois spécifiques à certains types de recherches. En 2007, le *CIDOC Conceptual Reference Model* modélise les événements pour connecter des faits ensemble, former des histoires cohérentes et ainsi donner une représentation du monde liée et centrée sur les événements [DOS07]. La conception de *LODE* [STH09 ; Sha10] deux ans plus tard structure la représentation des événements et formalise des relations complexes dans les données : lieux, participants, temps, toutes ces données liées qui décrivent précisément les événements. *Simple Event Model* (*SEM*) [vHag+11], publié plus tard, simplifie cette représentation. En 2021, le *Records in Context Conceptual Model* (*RIC CM)* [Int21] définit également ce concept d'événement dans le cadre de l'archivage de documents.

La question des événements a également été abordée dans le contexte de la biologie, avec quelques efforts menés dans le cadre des campagnes *BioNLP* [Kim+09 ; MSM11 ; Rie+11 ; Ned+13]. Les tâches de recherche consistaient à extraire des événements de documents biomédicaux. Un autre concours, lancé en 2004 par l'*Informatics for Integrating Biology and the Bedside* (*i2b2*), encourageait le développement de techniques de traitement du langage pour l'extraction d'informations relatives aux médicaments à partir de dossiers médicaux. L'objectif était d'accélérer la traduction des résultats cliniques en nouveaux diagnostics et pronostics. Dans le même temps, le suivi de mentions d'événements épidémiologiques [Lej13 ; Lej+15] a apporté un nouvel éclairage sur la notion d'événement, ici spécifique au domaine médical et aux modélisations sanitaires.

Nous avons distingué plusieurs stratégies et tendances dans la définition des événements, ainsi que plusieurs campagnes d'évaluation, projets et concours différents. La notion d'événement est un concept ambigu [Spr18] et sa définition a beaucoup varié au fil des années, selon les domaines et champs d'investigation.

En raison de la variabilité de la définition des événements, nous différencions trois approches qui traitent les événements différemment. L'une les considère comme des constituants naturels du texte et un article de presse décrit nécessairement un événement. Une autre explore les liens sémantiques entre les événements et les entités qui y sont rattachées. Enfin, une dernière traite l'événement comme une structure ontologique qui organise la connaissance.

— **Événements à domaine ouvert**. Les documents traités sont des articles ou des brèves de presse, des documents qui rapportent authentiquement des événements. La représentation de ces derniers ne suit pas de schéma prédéfini ni aucune structure. Le postulat est que chacun des documents traite d'un événement. Décrire l'événement, c'est extraire des propriétés de ces textes pour en générer une représentation unique. Les projets liés à *TDT* [All02b], *EMM* [PSD08] ou *Event Registry* [Rup+16] exploitent ce type de représentation.
— **Événements spécifiques à un domaine**. Le niveau de traitement de l'événement n'est plus le document, mais la phrase. Chacune est analysée à la recherche

de constituants qui décrivent une action, des dates, des lieux, des participants et tout ce qui lie ces éléments entre eux. C'est la stratégie utilisée dans les campagnes d'évaluation *MUC*, *ACE*, *TimeML* et *TAC-KBP*. C'est l'échelle la plus fine et la plus petite pour représenter un événement dans un texte [FHM06].

— **Description ontologique d'événements**. Les événements sont représentés sous forme d'ontologies : des propriétés définissent l'action et tout ce qui s'y rapporte (les lieux, les entités participantes, les types de liens entre elles, les dates, etc.). Ces représentations sont riches, mais permettent de structurer le concept d'événement, facilitant sa définition formelle.

À ce stade, nous disposons d'une riche littérature scientifique donc l'objet est l'étude des événements et de leur propagation. En s'intéressant aux définitions adoptées dans ces projets puis aux représentations exploitées, nous cherchons à identifier les éléments que toutes ont en commun. Nous proposons une définition formelle adaptée au contexte de suivi de mentions d'événements décrits dans la presse.

2.1.1 Événements à domaine ouvert

Les événements que l'on catégorise dans des domaines « ouverts » sont faiblement structurés. Aucune organisation de la connaissance n'est supposée comme dans une ontologie. Ces représentations d'événements sont fortement liées au contenu manipulé : des articles de presse. Sur de tels supports, chaque article fait le récit d'un événement donné et donc tous les termes qui composent l'article décrivent cet événement. Cette formulation de « domaine ouvert » se retrouve dans la littérature [XW19] pour nommer ce type de représentation, à l'opposé du « domaine fermé », spécifique à des cas d'étude particuliers.

Dans ce domaine, nombreux sont les travaux qui depuis plus de vingt ans, c'est-à-dire depuis que le projet *Topic Detection and Tracking* a fondé la discipline, analysent l'évolution des événements – ici parfois nommés sujets – tant à la fois sur la presse [ALJ00 ; Pou+04 ; PSD08 ; RM12 ; Rup+16 ; LH17 ; CCG17 ; MBC19 ; Fan+21 ; ZXZ21 ; LBH21 ; SMM22] que sur les réseaux sociaux numériques et plus spécifiquement Twitter [PM10 ; POL10 ; JSS11 ; WL11 ; BNG11 ; Rit+12 ; BGC14 ; AK15 ; MBC19 ; Mot+21 ; Asg+21].

L'objectif premier de ces travaux est de classer un ensemble de documents et de former des histoires, des linéaires de documents organisant la connaissance à propos d'un événement donné. Les solutions peuvent traiter des flux de documents [RM12 ; MBC19 ; Mir+18] ou des bases figées d'articles de presse. Les prémisses du projet *TDT* l'inscrivent également dans la recherche de solutions multilingues à cette problématique. Les premiers essais furent sur des textes en anglais et en mandarin [Cie+02 ; CL02], de nombreux langages de l'Union européenne pour le projet *Europe Media Monitor* [Pou+04 ; Ste+04 ; PSD08 ; Ste+15] ou d'autres travaux postérieurs [RM12 ; Mir+18 ; LH20 ; SMM22]

Trois concepts se trouvent mêlés : celui de l'événement, de sujet (*topic* en anglais) et de l'histoire (*story*). La première est la plus essentielle. Le projet *TDT* donne de premières définitions de ces termes :

Définition 1. *« Un événement dans le contexte de TDT est quelque chose qui se produit à un endroit et un moment spécifiques associés à certaines actions spécifiques ». (an event in the TDT context is something that occurs at a specific place and time associated with some specific actions.) [Yan+00a].*

Plus tard, une autre définition est apportée, précisant le lien entre un événement et les éventuelles causes (ou prémisses) qui l'ont engendré et les conséquences que l'événement peut entraîner.

Définition 2. *« Un événement est une chose spécifique qui se produit à un moment et en un lieu précis, avec toutes les conditions préalables nécessaires et les conséquences inévitables ». ([an event is a] specific thing that happens at a specific time and place along with all necessary preconditions and unavoidable consequences) [Cie+02].*

À partir de ces définitions, on extrait un ensemble de propriétés basiques : un « quoi », l'action qui a eu lieu, la chose qui s'est déroulée. Un « où » et un « quand » ancrent l'événement dans son contexte spatio-temporel. Pour qu'une action soit événement, elle doit être définie dans le temps et dans l'espace. L'événement ne doit pas être isolé de ce qui l'a engendré ni de ses conséquences. Plus formellement, un événement est connecté à d'autres événements, ses prémisses et conséquences. L'histoire est intrinsèquement liée à l'événement.

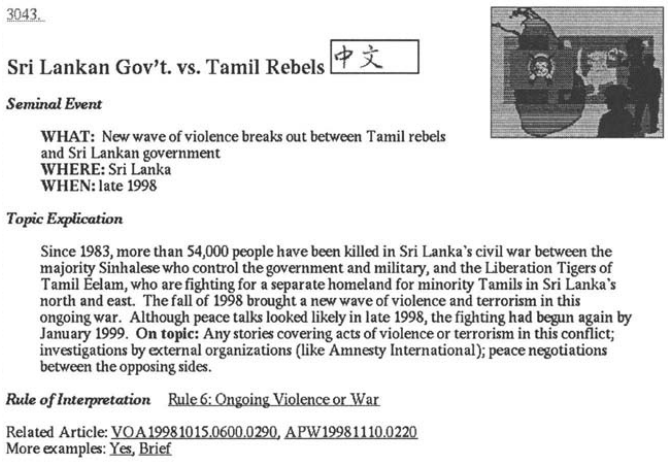
Définition 3. *« Une histoire est un segment d'information cohérent qui traite du même sujet et qui comprend deux ou plusieurs clauses déclaratives indépendantes sur un seul événement ». ([a story is a] topically cohesive segment of news that includes two or more declarative independent clauses about a single event [GD02].*

Une histoire débute dès lors que deux événements sont interconnectés. Si l'on fait le lien avec la définition 2, l'interconnexion s'établit par l'existence de prémisses et de conséquences à l'événement initial. Une histoire est dans ce cas composée d'une succession d'événements tous liés les uns aux autres. Le dernier concept est enfin celui de sujet (*topic*) qui englobe tous les autres pour former un ensemble cohérent.

Définition 4. *« Un sujet est défini comme un ensemble d'histoires qui sont fortement liées à un quelconque événement marquant initial ». (a topic is defined to be a set of news stories that are strongly related by some seminal real-world event [All02a]).*

Par exemple, un article intutilé *Résultats du référendum sur l'appartenance à l'Union européenne* publié par la *BBC* rapporte un événement qui s'est produit : un référendum, le 23 juin 2016, sur le territoire du Royaume-Uni. Cet événement s'inscrit dans un sujet plus large, celui du *Brexit* et forme la conclusion, peut-être le dernier événement de l'histoire dédiée au vote et à la campagne électorale. La figure 2.1 est un exemple de sujet présenté au sein du projet *TDT* au cours de la campagne d'annotations [Cie+02].

Cette définition formelle d'un événement est basée sur une interprétation plutôt linguistique et une expérience empirique du monde. Pour le projet *Europe Media Monitor* [Pou+04 ; PSD08], la définition des événements est identique à *TDT*. Les auteurs introduisent la notion de sous-événement [PSD08], c'est-à-dire d'événements inclus dans



3043.

Sri Lankan Gov't. vs. Tamil Rebels 中文

Seminal Event

WHAT: New wave of violence breaks out between Tamil rebels and Sri Lankan government
WHERE: Sri Lanka
WHEN: late 1998

Topic Explication

Since 1983, more than 54,000 people have been killed in Sri Lanka's civil war between the majority Sinhalese who control the government and military, and the Liberation Tigers of Tamil Eelam, who are fighting for a separate homeland for minority Tamils in Sri Lanka's north and east. The fall of 1998 brought a new wave of violence and terrorism in this ongoing war. Although peace talks looked likely in late 1998, the fighting had begun again by January 1999. **On topic:** Any stories covering acts of violence or terrorism in this conflict; investigations by external organizations (like Amnesty International); peace negotiations between the opposing sides.

Rule of Interpretation Rule 6: Ongoing Violence or War

Related Article: VOA19981015.0600.0290, APW19981110.0220
More examples: Yes, Brief

FIGURE 2.1 – Exemple de sujet d'un des corpus diffusés par le projet *TDT* [Cie+02].

d'autres, plus larges. La *bataille de la Marne*, incluse dans l'événement *Première Guerre mondiale*, en est un exemple. Au niveau numérique, d'autres points de vue, moins formels ont été adoptés. Dans la même lignée, les travaux de Mele et coll. [MBC17 ; MBC19] considèrent les notions d'événements et de sujet comme des concepts interchangeables. Ils utilisent des approches par modélisation de sujet (*topic modelling* en version originale) pour représenter les événements [BJY03 ; BL06], ce qui les pousse à fusionner ces deux concepts en un seul. Pour Laban et coll. [LH17 ; LBH21], les mentions d'événements dans des articles de presse peuvent se répandre sur des mois, comme exprimé par Pouliquen et coll. [PSD08 ; SPV09]. Les histoires modélisées dans *TDT* s'inscrivent dans le temps, contrairement aux sujets. Les liens de causalité lient les événements entre eux. Un événement en entraîne un autre, comme une réaction en chaîne, formant une histoire avec son début et sa fin. Pour Leban et coll. [Leb+14] aucune définition consensuelle d'événement dans la littérature n'existe : ils les décrivent comme « tout ce qui se passe de significatif dans le monde » (*any significant happening in the world*). Par conséquent, ils retirent les unités de temps et de lieu décrites dans le projet *TDT* et dans ses successeurs. Dans cette même veine, dans les travaux de Rupnik et coll. [Rup+16], les représentations d'événements (et non pas les événements eux-mêmes) ne sont que des groupes d'articles (*clusters*). Les événements sont « tout élément significatif dont les médias font état » (*any significant happening that is being reported in the media*). L'outil développé à la suite de ces recherches [Eve20] groupe les documents traitant d'un même événement et extrait des propriétés répondant aux interrogations quoi, quand, qui et où. La vision informatique qu'ils adoptent tient compte, à la différence de la définition qu'ils donnent, de ces éléments supplémentaires qui situent les événements dans le temps et dans l'espace. Et pour cause, deux ensembles distincts d'articles liés au même sujet peuvent ne pas rapporter le même événement s'ils évoluent dans des contextes spatio-temporels différents.

Approches et applications

Nombreuses sont les implémentations numériques de cette définition d'événement. Le document est lui-même la description de l'événement. Dans les approches initiales de ce problème, les documents sont convertis en vecteurs numériques, ce qui permet de calculer leurs similarités par paires [YPC98 ; Nal+04 ; MRŠ12 ; Rup+16 ; SW17 ; BPL18]. Dans certains cas, des pondérations *TF-IDF* [Sta+19 ; Mar+21] encodent les textes tandis que dans d'autres, les documents sont représentés sous forme de chaînes lexicales, des séquences de mots sémantiquement liées [SC01]. Ces représentations basées sur les termes du texte et leur analyse sont en difficulté face à des messages courts comme peuvent l'être les textes publiés sur Twitter [WL11 ; Rit+12 ; AK15 ; GLM16 ; Asg+21].

Les approches multilingues, apparues dans le projet *TDT*, sont basées sur des contenus traduits vers une langue pivot, choisie pour la grande disponibilité de ses ressources linguistiques. Les textes en mandarin étaient traduits par un processus automatique en anglais dans *TDT* et les résultats étaient médiocres [All02b]. La polysémie est un élément facteur de confusion dans ces représentations. Aujourd'hui, le problème est limité par les dernières avancées en traduction automatique [VBG18]. Les projets ultérieurs se sont concentrés sur l'analyse multilingue de documents [SPH02 ; SGJ11 ; RM12 ; MRŠ12]. Cette dernière est au cœur de travaux plus récents qui analysent des documents de presse rédigés dans des langues différentes. L'objectif est de construire des histoires à partir de représentations comparables entre toutes les langues [All02b ; Pou+04 ; Rup+16 ; Mir+18 ; LH20 ; SMM22]. Cette recherche s'est appuyée d'abord sur des données supplémentaires extraites du texte telles les entités nommées (noms de personnes, dates, lieux), des taxonomies, des unités ou des mots identiques dans toutes les langues (par exemple, *tsunami*). Les vectorisations issues de modèles d'apprentissages profonds comme *BERT* [Dev+19 ; RG19 ; RG20], *XL-NET* [Yan+20] pour ne citer qu'eux et basés sur des architectures de transformateurs (*Transformers* [KRS21]) fournissent des encodages multilingues, c'est-à-dire des vecteurs similaires pour des concepts identiques représentés dans des langues différentes.

L'extraction d'événements à domaine ouvert dans des articles repose sur une définition élémentaire. Dans ce contexte, l'objectif est la détection d'événements dans de grands corpus sans schéma prédéfini. Une autre approche basée sur l'analyse de phrases et l'extraction de termes spécifiques avec leurs rôles est possible. Ce sont les approches en « domaine fermé ».

2.1.2 Événements spécifiques à un domaine

Les représentations d'événements à « domaine fermé » [XW19] ou « spécifiques à un domaine » sont utilisées pour les événements pour lesquels une représentation sous forme de structure type existe. Des représentations ontologiques ou encore les modèles (*templates* en anglais) des projets *MUC* ou *ACE* sont des représentations de ce type. Un schéma d'événement reflète la structure de l'événement. Le cas des ontologies plus génériques est évoqué distinctement à la section suivante (sous-section 2.1.3, page 37).

Les travaux liés à cette thématique de recherche sont portés par une longue histoire

qui débute dans les années 1990 avec le projet *MUC* [CLH93 ; SC93 ; Sun95 ; Bor+98 ; Chi98] suivi quelques années après par le projet *ACE* [Dod+04 ; Lin05]. De ces derniers découlent d'autres tels *TAC-KBP* [Agu+14] ou la campagne d'évaluation qui a permis à la norme d'annotation *TimeML* d'émerger [Pus+03 ; Pus+04 ; PLS07 ; Pus+10 ; Cas+11 ; Bit+11]. Chacune porte en elle sa propre vision des événements et tous sont liés à la tâche de reconnaissance et de détection d'événements (*Event Detection and Recognition tasks*) [ST17].

L'objectif de ces approches est d'extraire les informations contextuelles des événements mentionnés dans des phrases. L'approche à domaine ouvert est très fortement liée au suivi des mentions d'événements, tandis que la recherche en domaine fermé est davantage associée à l'extraction d'information (*Information Retrieval*, *IR*). La première définition (5) donnée au concept d'événement est cryptique et vague puis fortement formalisée par la suite en des termes techniques, à la définition 6.

Définition 5. *« Les événements sont définis comme des "occurrences spécifiques", impliquant des "participants spécifiques" ». (Events [...] are defined as "specific occurrences", involving "specific participants") [Lin05 ; Agu+14].*

Définition 6. *« Les événements sont représentés par leurs attributs et participants. Ces derniers sont les entités ACE qui participent à l'événement. [...]. Un événement ACE peut avoir un certain nombre de participants, chacun caractérisé par le rôle qu'il joue dans l'événement (agent, objet, source, cible). Actuellement, les attributs des événements sont le type (détruire, créer, transférer, déplacer, interagir) et la modalité (réel, non réel) » [Dod+04].*

Au sein du projet *MUC-3*, la tâche *Scenario Template* (*ST*) se concentre sur l'extraction d'informations à propos des événements dans du texte. Elle traite également de l'association des actions (par exemple le lancement d'un véhicule aérien [CM98]) avec les entités nommées impliquées dans celles-ci. Dans ce cadre, un événement est considéré comme un ensemble de relations qui associe un acte et des participants dans un cadre spatio-temporel défini [GS96 ; ST17]. Un tel exemple d'annotation au sein du projet *MUC-3* est présenté en figure 2.2.

Chaque modélisation de scénario est spécifique à un domaine particulier. Dans le cas de la figure 2.2, le modèle contient 18 éléments à extraire. Ces éléments donnent une vision globale de l'événement, ici de type « attaque terroriste ». Chaque type d'événement dispose de son propre modèle d'extraction. En effet, un *auteur* perpétue une attaque, c'est un détail dénué de sens dans le cas d'un événement de type naissance, par exemple. Il n'existe pas d'*auteur* pour un événement de type « naissance ».

Plus tard, le programme *ACE* aborde les mêmes questions, mais avec une approche nouvelle qui hérite des définitions liées à *MUC*. Une tâche d'extraction d'événements (*Event Extraction*, *EE*) est définie dans ce cadre. Elle propose d'identifier des instances d'événements dans du texte et toutes les entités impliquées dans celles-ci. Une expression représente chaque événement, une phrase ou une portion de texte : le déclencheur de l'événement. Après classification du déclencheur en un type donné, des arguments

0. MESSAGE ID	TST1-MUC3-0080
1. TEMPLATE ID	1
2. DATE OF INCIDENT	03 APR 90
3. TYPE OF INCIDENT	KIDNAPPING
4. CATEGORY OF INCIDENT	TERRORIST ACT
5. PERPETRATOR: ID OF INDIV(S)	"THREE HEAVILY ARMED MEN"
6. PERPETRATOR: ID OF ORG(S)	"THE EXTRADITABLES" / "EXTRADITABLES"
7. PERPETRATOR: CONFIDENCE	CLAIMED OR ADMITTED: "THE EXTRADITABLES" / "EXTRADITABLES"
8. PHYSICAL TARGET: ID(S)	*
9. PHYSICAL TARGET: TOTAL NUM	*
10. PHYSICAL TARGET: TYPE(S)	*
11. HUMAN TARGET: ID(S)	"FEDERICO ESTRADA VELEZ" ("LIBERAL SENATOR" / "ANTIOQUIA DEPARTMENT LIBERAL PARTY LEADER" / "SENATOR" / "LIBERAL PARTY LEADER" / "PARTY LEADER")
12. HUMAN TARGET: TOTAL NUM	1
13. HUMAN TARGET: TYPE(S)	GOVERNMENT OFFICIAL / POLITICAL FIGURE: "FEDERICO ESTRADA VELEZ"
14. TARGET: FOREIGN NATION(S)	-
15. INSTRUMENT: TYPE(S)	*
16. LOCATION OF INCIDENT	COLOMBIA: MEDELLIN (CITY)
17. EFFECT ON PHYSICAL TARGET(S)	*
18. EFFECT ON HUMAN TARGET(S)	-

FIGURE 2.2 – Exemple de document pour le projet *MUC-3*, annoté d'un événement et de participants, sous la forme de *Scenario Template* [CLH93].

décrivant ce type d'événement sont recherchés. Ce sont des expressions, entités et informations spatio-temporelles impliquées dans l'événement, comme représenté dans la figure 2.3.

Dans *ACE*, un événement est une structure de données à propriétés multiples [Lin05] :

— **Mention de l'événement** : phrase ou expression dans lesquelles l'événement est contenu.

— **Déclencheur ou ancre de l'événement** : le mot qui exprime le plus clairement l'occurrence de l'événement. C'est un verbe simple ou à particule qui peut être étendu à des noms, des noms à particule, des pronoms et des adverbes. Ils sont classés en types et sous-types prédéfinis (par exemple, *Conflit* a deux sous-types, *Attaque* et *Manifestation*).

— **Argument d'événement** : une mention d'entité, une expression temporelle ou une valeur qui sert de participant ou d'attribut avec un rôle spécifique dans une mention d'événement. Ce sont souvent des entités nommées [Bor+98] de personnes, d'organisations ou de lieux.

— **Rôle de l'argument** : la relation entre un argument et l'événement. Chaque type d'événement est associé à certains arguments attendus qui ont des rôles (le type d'événement *Vie.Mariage* attend quatre arguments, de type *Personne, Temps* et *Lieu*) : les mariés ainsi que le lieu et la date du mariage.

L'exemple de la figure 2.3 est un événement de type *Attaque* comme l'indique le déclencheur « *attacks* ». Les arguments extraits sont « Groupe Al-Qaeda d'Oussama ben Laden » pour l'*attaquant*, type d'argument spécifique au type *Attaque* et « les États-Unis » pour le *lieu*.

Le schéma *ERE* (*Entities, Relations and Events*), développé dans le cadre du programme *DEFT* de la *DARPA* [Agu+14] simplifie les représentations de *ACE* [Son+15].

```
<event ID="APW_ENG_20030520.0757-EV8" TYPE="Conflict" SUBTYPE="Attack" MODALITY="Other"
  <event_argument REFID="APW_ENG_20030520.0757-E18" ROLE="Attacker"/>
  <event_argument REFID="APW_ENG_20030520.0757-E9" ROLE="Place"/>
  <event_mention ID="APW_ENG_20030520.0757-EV8-1">
    <extent>
      <charseq START="1392" END="1477">Osama bin Laden's Al-Qaeda group possibly
launching fresh attacks in the United States</charseq>
    </extent>
    <ldc_scope>
      <charseq START="1305" END="1516">Earlier this week, Saudi and U.S. officials said they had new
intelligence pointing to Osama bin Laden's Al-Qaeda group possibly
launching fresh attacks in the United States or against American
interests overseas</charseq>
    </ldc_scope>
    <anchor>
      <charseq START="1450" END="1456">attacks</charseq>
    </anchor>
    <event_mention_argument REFID="APW_ENG_20030520.0757-E18-42" ROLE="Attacker">
      <extent>
        <charseq START="1392" END="1423">Osama bin Laden's Al-Qaeda group</charseq>
      </extent>
    </event_mention_argument>
    <event_mention_argument REFID="APW_ENG_20030520.0757-E9-44" ROLE="Place">
      <extent>
        <charseq START="1461" END="1477">the United States</charseq>
      </extent>
    </event_mention_argument>
  </event_mention>
</event>
```

FIGURE 2.3 – Exemple d'annotation d'un événement au sein du corpus *ACE* publié en 2005 [Lin05]

.

Plus tard, l'approche au sein de *TAC-KBP* passe à une nouvelle représentation plus riche, *Rich-ERE* [Son+15]. Dans cette nouvelle structure de données, un type d'événement et onze sous-types sont ajoutés à la représentation originale. Alors que *Light-ERE* possède un modèle formel de coréférence d'événements (même agents, mêmes patients, même heure et même lieu) et se limite aux mentions d'événements véridiques, *Rich-ERE* fournit une définition plus simple. Chaque mention d'événement possède un attribut *realis*, utilisé pour classer les mentions d'événement avec des valeurs possibles : *Actuel* (pour un événement réel qui survient), *Générique* (événement générique, habituel) et *Autre* (événement futur, hypothétique, nié, incertain, etc.) [Bie+16 ; Gha+18 ; Son+18].

Le projet *TimeML* [Pus+04] traite de l'identification temporelle des événements, de l'horodatage, de l'ordonnancement, du raisonnement avec des expressions temporelles et de la persistance des événements [Pus+04]. *TimeML* a conduit à la création d'une norme d'annotation axée sur les relations temporelles entre deux événements (*par ex. malade* et *mort*) ou entre un événement et une expression temporelle (*par ex.* les termes *naissance* et *il y a deux mois*). La définition des événements adoptée dans *TimeML* ne limite pas son applicabilité à d'autres langues et l'annotation sémantique, d'abord adaptée à la langue anglaise, est devenue une norme internationale, *ISO-TimeML* [Pus+10]. Le corpus *TIMEBANK* met en œuvre la norme *TimeML* à partir de publications de nombreux médias [GS03]. Il contient 300 textes, dont 50 proviennent de transcriptions d'émissions *ACE* et 49 de dépêches *ACE*. Le reste est issu du corpus *PropBank*, qui se compose de textes du Wall Street Journal [GS03]. Des auteurs publient des corpus localisés dans d'autres langues comme en français [Bit+11], en italien [Cas+11] ou coréen [Jeo+16 ; LJC19].

Approches et applications

En domaine fermé, les principaux défis reposent sur l'annotation des différents éléments : la mention, le déclencheur et les arguments. Les travaux dédiés à l'extraction d'événements [Bor18] sont répartis en différentes catégories : ceux basés sur des patrons (*patterns*) [Ril95 ; Yan+00b], sur de l'apprentissage automatique basé sur les caractéristiques du texte (*feature-based* en anglais) [Hon+11 ; LJH13 ; Bro+15] et les approches basées sur une architecture d'apprentissage automatique et profond [Yan+00b ; NG15 ; FQL18 ; Bor+20b ; BD21].

Les premières, basées sur des patrons, consistaient en des prédicats, des déclencheurs d'événements et des contraintes dans le contexte syntaxique local. Elles comprenaient également un riche ensemble de caractéristiques lexicales *ad hoc* (mots composés, lemmes, synonymes, marques grammaticales (*Part-of-Speech*), etc.) et de caractéristiques sémantiques (*WordNet* [Mil95] ou répertoires géographiques) pour identifier les rôles de chaque argument. Les systèmes d'extraction d'événements actuels sont basés sur des architectures d'apprentissage profond. Plus généralement ce sont deux modèles : les réseaux neuronaux convolutifs (*CNN*), venant du domaine du traitement des images et les réseaux neuronaux récurrents (*RNN*), généralement mieux adaptés au caractère séquentiel des textes. Ces méthodes traitent généralement chaque mot comme un déclencheur d'événement potentiel [NG15 ; Che+15]. Dans un second temps, les réseaux de neurones récurrents bidirectionnels (*Bi-RNNs*) [LJH13 ; NCG16 ; JY16] ont remplacé les *CNN*. Les approches récentes exploitent soit les réseaux génératifs [Hon+18], soit le paradigme de l'apprentissage par renforcement [ZJS19]. Enfin, Yang et coll. [Yan+19] tentent une approche basée sur le modèle préentraîné *BERT* [Dev+19].

Dans un contexte d'extraction d'événements à domaine fermé, plusieurs autres solutions ont également été abordées. Par exemple, dans le domaine épidémiologique, le système *DAnIEL* [Bri+13 ; Lej+15] accélère la détection de maladies émergentes à partir de sources de presse multilingues [tea11]. Pour fonctionner, le système connaissait des noms de maladies et identifiait les articles pertinents et les mentions d'événements. Rachelle Sprugnoli [Spr18] aborde la question de la détection et de la catégorisation des événements dans les textes historiques.

2.1.3 Descriptions ontologiques d'événements

Les descriptions ontologiques d'événements sont liées aux représentations à domaine fermé. Elles en sont un sous-ensemble. Cependant, ces représentations ont été également associées au champ de recherche *Topic Detection and Tracking* [Liu+20], explorant la construction d'histoires dans la presse à partir de représentations ontologiques d'événements.

Depuis le début des années 2000, de nombreuses définitions du concept d'événement sont adoptées. Le projet *ABC* [LHP01] visait à rendre interopérables des concepts de différents domaines et a donné une définition spécifique des événements (7). Dans cette représentation, des liens de causalité associent les événements, rejoignant la vision philosophique de causalité kantienne [Kan81].

Définition 7. *« Un événement marque une transition d'un état à l'autre. Les événements sont toujours ancrés dans le temps ». (An event marks a transition from one state to another. Events always have time properties) [LHP01].*

Par la suite, quelques auteurs donnent une autre définition (8) liée au domaine musical [Rai+07 ; RA07]. On y trouve comme toujours l'ancrage de l'événement dans un contexte spatio-temporel, élément incontournable. Cette notion de changement d'état se retrouve par ailleurs dans d'autres travaux de la littérature à ce sujet [KIF07 ; STH09 ; Sha13] et plus précocement dans les travaux de Bach [Bac86].

Définition 8. *« Une classification arbitraire d'une région d'espace-temps, par un agent cognitif. Un événement peut avoir des agents qui participent activement, des facteurs passifs, des produits et une localisation dans l'espace et le temps ». (An arbitrary classification of a space/time region, by a cognitive agent. An event may have actively participating agents, passive factors, products, and a location in space/time) [RA07].*

Enfin, la troisième définition (9) d'importance que l'on peut relever dans ce domaine est celle diffusée au sein du projet *EventKG* [Got+18 ; GD19]. Son but est de fournir une base de connaissance des événements pour les analyser.

Définition 9. *« Les événements [...] sont des actions du monde réel d'importance sociétale, par exemple des conflits militaires, des tournois sportifs et des élections politiques. En particulier, nous considérons les événements, les entités qu'ils impliquent et les relations temporelles - c'est-à-dire les relations du monde réel entre les événements et les entités valables sur une période ». Events considered in this work are real-world happenings of societal importance, with examples including military conflicts, sports tournaments and political elections. In particular, we consider events, entities they involve and temporal relations - i.e. real-world relations between events and entities valid over time [GD19].*

Pour synthétiser ces trois définitions, on retrouve évidemment l'ambiguïté qui est propre au concept d'événement dans les domaines ouverts et fermés. L'ancrage spatio-temporel, le lien avec des concepts exprimant les lieux et les temporalités sont nécessaires. Des acteurs ou plus généralement des participants aux événements sont mentionnés. Au sein du projet *CIDOC* [DOS07], l'objectif est de connecter des faits avérés en une représentation cohérente des histoires de presse. Elle ne sont dans ce cas qu'une succession d'événements dont la connaissance est structurée par une ontologie. Ryan Shaw [STH09 ; Sha10] est à l'origine d'un état de l'art de ces ontologies d'événements spécifiquement orientées vers les projets de Web sémantique et de la spécification de *LODE*. Celle-ci est une synthèse de ces ontologies d'événements. Le modèle d'événement nommé sobrement *F* [Ans+09] s'intéresse quant à lui plutôt à la modélisation des interactions entre les événements, notamment les liens de causalité et de corrélation.

Sur un plan plus générique, le concept d'événement existe aussi sans être défini dans de nombreuses bases de connaissances comme *DBPedia* [Aue+07], *YAGO2* [Hof+13] ou plus récemment Wikidata [VK14]. Toutes sont liées entre elles et le choix d'une base

de connaissance doit faire l'objet d'une étude préliminaire à chaque tâche [Fär+17]. *DBPedia* est le premier graphe de connaissance issu de l'encyclopédie Wikipedia et peuplé à partir des boîtes d'informations des en-têtes (*infobox*). *YAGO2* dispose des mêmes caractéristiques, il est bâti sur Wikipédia, mais agrémenté de concepts issus de *WordNet* [Mil95] et de *GeoNames*[4].

Pour répondre à l'absence de graphe de connaissance spécifiquement dédié aux événements, Gottschalk et coll. publient le graphe *EventKG* [GD19]. Il implémente l'ontologie *Simple Event Model* [vHag+11]. Au sein de celle-ci, l'événement est l'élément central, typé. Il est lié à des acteurs, des lieux et est ancré temporellement selon son type. Des contraintes supplémentaires s'appliquent sur les liens entre événements et participants, comme le rôle du lien ou sa durée. Le rôle de ces entités est d'informer sur qui, quoi, quand et où l'événement se produit [XW19].

Approches et applications

Une tendance pour le suivi d'informations dans la presse consiste à extraire la connaissance sémantique depuis des sources de données afin de connecter les événements aux articles de presse. La représentation ontologique des événements est un moyen d'obtenir des informations structurées utilisables dans d'autres contextes. La question de l'espace-temps est centrale dès l'origine où plusieurs travaux utilisent des expressions temporelles et spatiales pour améliorer la détection des événements dans du texte [MAS03 ; MAS04]. Le projet *NewsReader* [Vos+05 ; Vos+16 ; Age+16] utilise la représentation *SEM* pour extraire des informations élémentaires sur les événements : ce qu'il se passe, quand et avec quels participants. Des ontologies participent aux mécanismes de suivi de mentions d'événements par la classification des concepts (lieux, dates) [Liu+20]. Les représentations de ces concepts associés sont également utilisées afin de comparer des propriétés d'événements et de calculer des similarités entre elles [Mak03].

Un autre objectif des représentations ontologiques est leur utilisation dans des bibliothèques numériques en fournissant, par exemple, des moteurs de recherche sémantiques ou basés sur les événements pour explorer les nouvelles historiques [VT11 ; Sha13 ; FH17 ; WJY22]. Ces derniers sont utilisés pour rechercher des documents évoquant des événements du monde réel et les visualiser chronologiquement. Des applications spécifiques de *Simple Event Model* ont été réalisées sur le domaine maritime, montrant la plasticité de cette représentation, adaptée à la fois à des événements spécifiques et à des événements plus généraux, comme dans *EventKG* [GD19]. Les représentations ontologiques d'événements sont aussi utilisées pour décrire et enrichir des archives documentaires. C'est ce dont est chargé notamment la représentation *Records in Context Conceptual Model* (*RIC CM*) [Int21].

2.1.4 Synthèse

Dans cette première section, nous avons proposé une synthèse des connaissances sur les événements. Nous avons exploré les définitions adoptées dans trois disciplines : à

4. `https://www.geonames.org/` (archive sur `https://web.archive.org`)

domaine ouvert et associées au projet *TDT*, à domaine spécifique, dans la lignée de *MUC* et *ACE* et l'autre discipline manipulant des ontologies. En définitive, nous proposons une synthèse du concept d'événement (définition 10) que nous utiliserons par la suite. Cette synthèse reflète les différents points de vue des travaux de recherche consacrés au suivi d'événements.

Définition 10. *Un événement est une action qui peut traduire un changement d'état. Ce dernier, induit par un autre, entraînera lui-même des conséquences. Pour être un événement, une action s'inscrit dans son propre cadre spatio-temporel et demeure bornée par ce cadre. Deux actions identiques se déroulant dans une temporalité ou une localisation différentes ne seront pas les mêmes événements. Tout événement fait intervenir des agents, actifs ou non dans l'action. Ces derniers sont les participants et sont représentés par des entités nommées (personnes, organisations géopolitiques, etc.) qui ont un rôle spécifique au regard de l'action. Les événements peuvent être typés, ce qui implique des rôles spécifiques aux agents (par exemple, dans le cas d'un événement de type « décès », l'identité du mort est attendue).*

Cette définition s'adapte bien à un large groupe d'événements comme les catastrophes naturelles, les élections, les actions politiques, culturelles (festivals, concerts, sorties au cinéma), les rencontres sportives, etc. Tous ces événements sont susceptibles d'être mentionnés dans les médias et donc faire l'objet d'un suivi spécifique pour retracer le parcours de l'information. Différentes implémentations techniques sont possibles, comme nous l'avons montré et certaines s'appliquent bien au suivi de mentions d'événements dans la presse.

2.2 Suivi d'événements à partir de documents de presse

La question du suivi des événements se pose lorsque l'on cherche à comprendre un fait historique ou actuel dans toute sa profondeur et ses subtilités. C'est, parmi d'autres, un travail réalisé par les historiens et historiennes amateurs ou universitaires lors de leurs recherches historiques. Pour le grand public et les événements actuels, des outils comme *Google News*[5] permettent d'organiser une veille informationnelle à l'échelle individuelle. L'outil trie les informations selon les centres d'intérêts de l'utilisateur ou de l'utilisatrice. Dans certains secteurs marchands notamment, une veille de ce type est essentielle : un événement quelconque peut avoir un retentissement global sur des marchés ou des activités professionnelles. Que ce soit pour obtenir un avantage concurrentiel en identifiant des signaux faibles ou pour réagir sur des marchés financiers [Bee+13 ; BGK15 ; Le 22], la connaissance des événements qui se produisent à un instant donné et l'identification de leurs causes et conséquences sont l'un des enjeux des économies modernes. Le secteur bancaire est l'un de ceux participant à la conception de ces systèmes [LH20].

La question du suivi des événements dans la presse est à l'origine des projets que nous avons mentionnés à la section précédente. Grâce à eux, nous avons analysé la définition

5. `https://news.google.com`

d'événements dans la presse. Pour brièvement rappeler ces projets, en premier lieu *TDT* [All02b] ouvre la discipline, définit les différentes tâches et formalise les problématiques scientifiques. Viennent ensuite, sans volonté d'être exhaustif, les projets liés à l'*Europe Media Monitor* [Pou+04], puis les travaux produits par des acteurs variés aux besoins divers à partir des années 2010. C'est dans cette continuité que vont émerger les projets *Event Registry* [RM12] ou *newsLens* [LH17]. Tous deux ont pour objectif de construire des histoires d'événements à partir d'articles de presse collectés automatiquement. Ces travaux reposent tous sur le même fonctionnement : les documents, généralement des articles de presse, sont groupés ensemble par divers procédés en *clusters*. Chacun d'eux est le reflet d'un unique événement dans la presse : tous les articles qui le composent décrivent le même événement, ses causes et ses conséquences. Une analyse de ces groupes par des mécaniques de *Question Answering* (*QA*) permet d'obtenir des descriptions sur ce qui se passe, quand, où, etc. [WJY22] et donc de qualifier les événements.

Hormis ceux-là, d'autres travaux sont conduits en ordonnant les articles de presse au sein de graphes orientés. Chaque nœud représente un article et les arcs matérialisent les connexions entre ces articles. Ces liens peuvent avoir différentes significations : ordre chronologique, similarité entre documents, plagiat partiel ou total, provenance géographique, etc. [AA05 ; RBS11 ; Zha+16 ; LS17]. Ces travaux s'inscrivent davantage dans une problématique de suivi des dynamiques de diffusion des informations, ici dans des documents textuels. Des thématiques identiques à *TDT* sont explorées, comme pour déterminer l'origine d'une information avant qu'elle ne se propage [PTV12]. Ces recherches sont toutes liées à la viralité de la diffusion de l'information, s'intéressant surtout à la propagation et à la rapidité des échanges d'informations dans la presse. Une grande partie de cette littérature s'intéresse d'abord à la blogosphère puis aux réseaux sociaux numériques, supports principaux des informations virales depuis leur utilisation massive au début des années 2010. [VVP08 ; Wan+11 ; Zar+17]. Les travaux les plus récents ont quant à eux dans l'objectif plus actuel de combattre la diffusion des fausses informations [NNT12 ; Gra+21].

Enfin, une troisième catégorie de travaux est associée à celle des moteurs de recherche sémantiques. Pour répondre aux questions basiques sur les événements, certaines recherches scientifiques se basent sur des requêtes types issues des textes décrivant les événements [Gan+20 ; WJY21 ; Wan+21a ; WJY22]. Ces outils sont, dans ces situations, utilisés pour analyser des corpus massifs de documents historiques, afin d'en extraire de la connaissance et notamment celle liée aux événements [Sha13]. Ils se basent sur des descriptions textuelles d'événements à partir desquelles les principaux termes sont utilisés pour forger une requête. L'objectif de cette requête est d'obtenir les documents rapportant un événement particulier puis d'utiliser, comme précédemment mentionné, des techniques de *Question Answering* pour obtenir davantage de détails sur les événements décrits.

Pour cette section, nous proposons d'analyser ces procédés de suivi d'événements. D'abord, nous nous focaliserons sur la représentation des documents dans des formats numériques. Ensuite, nous détaillerons les algorithmes de construction d'histoires journalistiques. Dès les premières heures des projets liés au programme *TDT*, la question du

multilinguisme était posée. Quatre ans après le début du projet *Europe Media Monitor*, le système développé était en mesure d'analyser jusque 40 langues. Il pouvait établir des connexions entre des histoires rédigées dans 19 langues différentes [PSD08]. Le multilinguisme est à la base des enjeux associés à cette recherche. Nous présentons donc une partie de l'état de l'art dédiée à cet aspect particulier. Nous conclurons par un tableau récapitulatif des différentes méthodes employables pour suivre les événements rapportés par la presse.

2.2.1 Représentation des documents

Traiter des données, ce que font les algorithmes, présuppose que celles-ci soient préparées. Cette préparation a différentes formes : les articles et textes sont vectorisés et des propriétés sont extraites comme les dates et entités nommées mentionnées.

Les premiers travaux portés par le projet *TDT* utilisaient une vectorisation des documents par la méthode de pondération de termes *TF-IDF* [All02b]. Cette dernière signifie *Term Frequency, Inverse Document Frequency*. C'est une méthode de pondération qui évalue l'importance des termes relativement à un corpus de plus grande envergure. Nous reparlerons de cette méthode plus en détail à la section 4.1 lorsque nous présenterons les procédés de vectorisation que nous utilisons pour nos implémentations. Ce procédé d'encodage des textes est utilisé très largement dans la littérature [Pou+04 ; POL10 ; Shi+18 ; Mir+18 ; Huy+19 ; LH20 ; Jia+21 ; Mar+21 ; LLY21 ; SMM22 ; Zha+22]. Associé à une mesure de distance entre vecteurs telle la similarité cosinus, la comparaison des encodages, donc des documents est possible. Lorsque la similarité entre deux vecteurs est proche d'un, les termes encodés sont probablement semblables, ce qui signifie que les documents contiennent quasiment les mêmes termes. Il a par ailleurs été montré que la distance cosinus est efficace, comparé à la mesure euclidienne, pour distinguer des documents représentés par des vectorisations de ce type [Hua08]. Cependant, cette approche repose sur le mécanisme de « sac de mots » (*bag of words* en anglais). Les mots sont considérés ensemble sans tenir compte de leur ordre d'apparition dans les phrases ou de certaines nuances comme l'ironie [VLH18 ; Gha+20], le sarcasme [Sar+20 ; Fre+22] ou le ton. Ces derniers peuvent modifier intégralement la signification d'un texte pourtant rédigé dans les mêmes termes. La vectorisation par n-grammes contribue à réduire ce phénomène [BPL18]. Les termes du texte ne sont plus isolés, mais pris ensemble avec un nombre $n-1$ de voisins proches, formant des tuples de n termes. En intégrant du contexte, variable selon la valeur de n, l'effet négatif du sac de mots se réduit. L'autre problème posé par les vectorisations *TF-IDF* est que les vecteurs sont creux. Principalement constitués de zéros, leur taille est celle du vocabulaire de l'ensemble des documents du corpus. Seuls les termes présents dans le document encodé sont pondérés. Chaque texte ne contient qu'un sous-ensemble du vocabulaire d'entrée, ce qui explique la présence de valeurs nulles. Les vectorisations par *TF-IDF* ne permettent pas de comparer des textes rédigés dans des langues différentes. Le texte de chaque terme est pondéré par rapport à un corpus qui est représentatif de cette langue. C'est-à-dire qu'une vectorisation d'un document en anglais n'a de sens que pour être comparée à d'autres vectorisations dans la même langue, et basées sur le même vocabulaire d'entrée. Plutôt que de seulement

utiliser un vecteur par document, certains auteurs [Mir+18 ; LH20 ; SMM22] proposent d'encoder séparément les jetons, les lemmes et les entités nommées de chaque titre et corps d'article. À la place d'un unique vecteur qui décrit l'article, ils en traitent neuf, représentant plus finement le texte.

L'apparition de vectorisations d'un autre type au milieu des années 2010 a ouvert la voie à la résolution de la problématique multilingue. Au début de la décennie émergent des projets dont l'objectif est d'aligner des vecteurs dans un même espace s'ils représentent les mêmes concepts [Hua+15 ; Amm+16]. Dans le même temps, au sein des laboratoires de recherche de Google inc. apparaît l'idée de représentation distribuée de texte. Le mécanisme repose sur des réseaux de neurones simples qui encodent les mots et les documents [LM14]. Ces travaux ont initié le mouvement de plongement lexical (*word embedding* en anglais) qui consiste à représenter des phrases dans des vecteurs de taille fixe qui encapsulent et capturent le contexte. Les premiers travaux ont mené à la publication de *word2vec* [LM14] et *doc2vec*, ce dernier adapté à l'encodage de documents. Puis ont été publiés les modèles *Glove* [PSM14] ou *PolyGlot* [APS14]. Ce dernier est entraîné sur les 170 éditions de Wikipédia contenant plus de 10 000 articles, permettant d'obtenir des représentations vectorielles dans de nombreuses langues. L'apparition à la fin de la décennie de l'architecture transformateur (*Transformers* en version originale) [Dev+19 ; Yan+21] associée aux modèles linguistiques à base de masque (*Masked Model Language*, *MML*) renouvelle l'encodage et le décodage de documents. Ces modèles produits par *MML* demandent à être affinés pour fonctionner sur les tâches spécifiques comme la similarité sémantique [RG19 ; LH20]. Les modèles *BERT* [Dev+19] ou *USE* [Cer+18] sont capables d'encoder des textes en des vecteurs multilingues, c'est-à-dire que les encodages sont similaires pour des concepts identiques entre les langues [PSG19]. Pour capturer le contexte des phrases et non des termes, des travaux plus récents [RG19 ; RG20] ont abouti à la création de *S-BERT*. Ce modèle créé des vecteurs de phrases (*S* vaut pour *sentence*, phrase en anglais), comme *USE*, comparables par paires pour en déterminer la similarité sémantique entre les deux. Ces vecteurs ont été utilisés pour du suivi d'événements mentionnés dans une seule langue [SL20 ; LLY21 ; Ai21] ou dans plusieurs. Dans ce dernier cas, l'encodage fourni par ces modèles denses est un moyen d'identifier les articles de presse qui rapportent les mêmes événements, mais rédigés dans des langues différentes [LH20 ; SMM22]. Outre *S-BERT* [RG19], *USE* [Cer+18], les encodeurs *multilingual USE* ou [Chi+19], *Laser* [AS19] sont agnostiques aux langues. Ils génèrent des représentations alignées entre les différentes langues. Ces modèles reposent sur le principe d'apprentissage par transfert. Des modèles sont entraînés sur des langues fortement dotées en ressources pour résoudre d'autres tâches dans d'autres langues faiblement dotées. Staykovski et coll. [Sta+19], pour améliorer l'efficacité des algorithmes développés par Miranda et coll. [Mir+18] utilisent conjointement des représentations issues de pondérations *TF-IDF* et des représentations denses encodées avec *BERT* ou *doc2vec*.

La question du multilinguisme s'est posée dès l'origine. Tracer le parcours des informations dans une unique langue ne suffit pas. Estimer sa propagation dans différentes langues et à travers différents pays apparaît comme le seul moyen d'avoir une informa-

tion exhaustive et objective. Avant l'apparition des vectorisations denses fournies par les modèles présentés précédemment, d'autres systèmes sont proposés. En premier lieu, au sein de *TDT*, la question de la traduction automatique assistée par ordinateur est posée [ALJ00]. Les textes en mandarins étaient traduits vers l'anglais et analysés dans cette langue. L'exploitation de langues pivots [RM12], fortement dotées en ressources linguistiques a aussi été utilisé pour déterminer la similarité interlingue entre documents. Deux documents représentés dans des langues peu dotées en ressources sont comparés à la langue pivot au lieu de l'être entre eux. Une hypothèse formulée précocement [Pou+04 ; RM12] considère que les entités nommées et certains mots spécifiques (par exemple : *tsunami*) rédigés de la même façon dans toutes les langues sont bénéfiques. Au sein du projet *Europe Media Monitor*, une liste de ces derniers est construite pour identifier les documents contenant les mêmes termes. Les noms de lieux sont automatiquement connectés à des références uniques, grâce à la liaison des entités (*Entity Linking*) avec une base géographique. Enfin les termes sont comparés dans un modèle de classification multilingue basé sur les classes d'*EuroVoc* [SPH02]. C'est un corpus fournissant des traductions validées pour 6 000 classes dans 22 langues de l'Union européenne. Ce corpus de termes est utilisé pour identifier les similarités textuelles entre documents rédigés dans des langues différentes.

Dans les travaux liés à la recherche sémantique de documents, dont la portée est plus confidentielle [Gan+20 ; WJY21 ; WJY22], rien ne mentionne un prétraitement du texte. Une requête est forgée pour un moteur de recherche, en l'occurrence *ElasticSearch*, qui renvoie les documents les plus pertinents. Si la requête est une description d'un événement, par le titre d'un article comme c'est le cas ici, l'ensemble des documents obtenus seront en lien avec cet événement. Les termes de la recherche quant à eux sont prétraités. Les noms, adjectifs, verbes sont extraits du texte décrivant l'événement et étendus à tous leurs synonymes à l'aide de *WordNet* [Mil95].

Enfin, une dernière représentation de documents est basée sur des techniques de modélisation de sujets (*topic modelling*) [LCB12 ; MBC17 ; MBC19]. Une modélisation par une approche similaire à *Latent Dirichlet Allocation* (*LDA*) [BJY03], mais dynamique, c'est-à-dire évoluant dans le temps [BL06] est utilisée. Les articles sont représentés selon les sujet qu'ils décrivent. L'approche dynamique est gérée en faisant glisser une fenêtre de sept jours sur l'ensemble des données.

Par conséquent, la question du temps est primordiale dans l'analyse de la presse. La description des événements s'inscrit dans un contexte temporel borné, comme nous l'avons vu à la section 2.1. Sa gestion diffère selon les implémentations en deux possibilités. La première revient à traiter les articles de presse dans des fenêtres temporelles [Pou+04 ; LH17 ; MBC17 ; MBC19 ; LH20]. De taille variable, de 24 h [Pou+04] à 6 ou 7 jours [LH17 ; MBC19 ; LH20 ; LBH21], les documents publiés sur cette période sont groupés et traités ensemble. La matérialisation du temps qui s'écoule se fait en passant à la fenêtre suivante. Ce fonctionnement limite la portée des événements dans le temps. Un événement ne pourra pas durer plusieurs semaines. Une autre technique compare les horodatages des documents avec un score de similarité. Le temps est traité comme n'importe quelle information [Rup+16 ; Mir+18 ; SMM22]. Le temps est aussi un moyen

d'ordonner l'information et d'identifier les instants de contagion, c'est-à-dire, de diffusion de l'information dans un graphe [RBS11 ; Zha+16 ; Zar+17].

2.2.2 Algorithmes de suivi d'événements

La problématique à laquelle s'attellent les algorithmes de suivi d'événements s'assimile à de la classification. Chaque article se place dans au plus une classe ou potentiellement aucune. Le nombre de classes, donc d'événements est inconnu par avance. Il évolue au cours du temps. À chaque fois qu'assigner une classe à un article est impossible, une nouvelle est créée, avec ce document comme initialisation de la classe. Chaque classe est représentée par l'ensemble des documents qui la compose et par conséquent, chaque classe décrit un événement.

En domaine ouvert, dans la lignée des travaux associés à *TDT*, le nombre de projets et d'algorithmes est conséquent, que ce soit pour trier des documents de presse ou des informations publiées sur les réseaux sociaux numériques comme Twitter [WL11 ; Rit+12 ; BGC14 ; GF15 ; Asg+21]. Nous allons décrire dans cette section les principaux algorithmes que nous pouvons utiliser dans le cadre de ce travail, en mettant en avant ce qui les différencie et les confond.

L'algorithme de suivi développé dans le cadre de l'*Europe Media Monitor* [SPH02 ; Pou+04 ; Pou+06 ; PSD08 ; SPV09] fonctionne en deux étapes : la construction des *clusters* puis des histoires. Des *clusters* d'articles sont créés chaque jour à l'aide d'un algorithme hiérarchique agglomératif. Regrouper les articles chaque jour est un moyen de limiter la quantité d'information manipulée [GC09]. Les documents sont encodés par *TF-IDF*, comme pour certains des algorithmes de suivis proposés par *TDT* [SL99]. Les histoires sont composées de *clusters* qui s'enchaînent dans le temps. Ils sont liés entre eux dans la limite de ceux existant dans les sept derniers jours [Pou+04 ; LH17 ; LH20 ; MBC19]. Les *clusters* sont liés si leurs représentations *TF-IDF* sont similaires, par calcul de similarité cosinus, à au moins 50%. Des *clusters* qui ne sont pas associés à d'autres dans cette fenêtre de sept jours restent seuls : ils ne s'inscrivent dans aucune histoire. Cette stratégie est également mise en place au sein du projet *newsLens* [LH17 ; LBH21]. Dans ce dernier, le temps est représenté au sein de fenêtres de sept jours. Elles sont glissantes et, avec un recouvrement de 50 %, elles assurent que des histoires peuvent durer plus de sept jours au total. Les recherches autour de *newsLens* [Leb+14], de Staykovski et coll. [Sta+19] et de Linger et coll. [LH20] intègrent les mêmes paramètres temporels, se basant sur des fenêtres glissantes à recouvrement. Le même mécanisme de construction de *clusters* puis d'histoire est proposé. Les *clusters* d'articles de presse sont créés à l'aide de matrices de similarités *TF-IDF* (entre tous les documents publiés dans la même fenêtre temporelle). Sur celle-ci, ils appliquent un algorithme d'identification de communautés, Louvain [Blo+08], pour identifier les articles similaires, représentés sous forme de communautés de documents. La création des histoires (des liaisons entre les *clusters*) se fait suivant trois principes : création de lien, fusion ou division. Le premier se produit lorsque les *clusters* doivent être liés parce qu'associés au même sujet qui se développe. La fusion intervient lorsque les *clusters* sont identiques entre plusieurs fenêtres et la division lorsque les sujets évoluent en deux événements distincts. Les travaux de Linger et

coll. [LH20] ajoutent le suivi multilingue à ces travaux, en utilisant des représentations vectorielles denses, comme nous l'avons évoqué dans la section précédente.

Pour la question du multilinguisme, nous l'avons vu, au sein de *TDT*, les articles en anglais sont les seuls traités. Ceux rédigés dans une autre langue sont traduits vers l'anglais et analysés comme n'importe quel autre document [ALJ00]. L'*Europe Media Monitor* [Pou+04], utilise des propriétés additionnelles pour identifier les similarités entre documents rédigés dans différentes langues : mots rédigés de façon identique, corpus *EuroVoc* [SPH02]. Un score de similarité est produit entre les différentes histoires et *clusters* pour les lier entre eux, quelle que soit la langue de rédaction. Pour le projet *Event Registry* [Leb+14 ; Rup+16], les auteurs se basent sur la notion de langue pivot [RM12]. C'est une langue intermédiaire, riche en ressources linguistiques et en contenu. Grâce à cette dernière sont établies des connexions entre des termes contenus dans les documents et un ensemble de documents. Les histoires et *clusters* sont comparés en utilisant un algorithme de K-Moyennes [Mac67] ou CL-LSI [Dee+90]. Des liens sont créés entre documents issus de *clusters* de différentes langues. Plus le nombre de liens est élevé entre ceux-ci, plus la probabilité qu'ils traitent de la même histoire augmente.

Miranda et coll. [Mir+18 ; SMM22] proposent une alternative pour gérer à le temps et les *clusters*. Leur algorithme est similaire, dans sa conception, à celui proposé par Pouliquen et coll. [Pou+04] pour l'*Europe Media Monitor*. Dans leur proposition, les *clusters* ne sont pas issus d'articles publiés sur une même fenêtre de temps, mais construits au fil de l'eau. D'abord, les articles de presse sont représentés par *TF-IDF* et groupés sous la forme d'un flux. Le nombre de *clusters* n'est pas fixe, il évolue constamment dans le temps. Cela implique que, contrairement aux autres travaux, le temps est une caractéristique quelconque. Ici, le temps s'exprime comme la différence temporelle entre la publication de l'article et le *cluster* qui représente l'événement. Plus cette durée est faible, plus haute est la probabilité que le document y soit rattaché. Les *clusters* sont construits avec l'aide de l'algorithme *SVM-Rank* [Joa06] qui détermine si un article doit être associé à un *cluster* existant. Linger et coll. préfèrent la régression logistique, dont les implémentations sont plus largement disponibles [LH20]. Avec ce dernier et Rupnik et coll. [Rup+16] l'association multilingue se produit dans un second temps. Les *clusters* monolingues sont liés dans les différentes langues grâce à une représentation vectorielle qui projette les documents rédigés dans des langues différentes dans le même espace (avec *BERT* [Dev+19] par exemple). Un processus d'apprentissage sert à trouver un seuil de similarité pour déterminer si les histoires en langues différentes sont identiques ou non.

L'originalité des mécanismes proposés par Mele et coll. [MBC17 ; MBC19] provient de la représentation qu'elle donne aux événements. Dans ces travaux, la modélisation de sujets [BJY03 ; BL06] définit les événements dans une fenêtre sans recouvrement de sept jours. Sur Twitter également [LCB12], l'approche par *topic modelling* a montré son utilité. Ensuite, un algorithme de K-Moyennes ou de *LDA* hiérarchique (*hLDA* [Gri+03]) groupe ces thèmes entre eux pour former des histoires à partir des documents de presse. Dans cette lignée de travaux, peu d'auteurs proposent des représentations alternatives à *TF-IDF* [Sta+19 ; LLY21] pour encoder les documents.

Dans le cas du moteur de recherche d'événements que nous avons brièvement évoqué,

les requêtes, une fois forgées, servent à collecter des documents en rapport avec les événements. Les auteurs [Wan+21a ; WJY22] limitent la recherche à 100 documents. Leur objectif est de décrire les événements, non de collecter tous les articles qui le mentionnent. Cette limitation n'a pas de sens dans notre cas. Un événement peut être décrit avec 10 ou 1000 documents. Limiter artificiellement la récupération des résultats aurait un effet délétère. Les articles retournés sont ordonnés selon le calcul probabiliste *Okapi BM25* [Rob+97 ; SPH02].

On peut également citer, dans une catégorie à part, les approches développées dans le cadre de projet d'exploration de presse [SSG14 ; Khr+15 ; Ves+17]. Celles-ci se basent sur la réutilisation de texte entre les documents. Constatant que l'information se diffuse aussi par le biais de reprise de textes, repérer ces réutilisations est un moyen d'identifier la propagation des mentions d'événements. Cette approche ne dispose pas de la précision sémantique que peuvent apporter les techniques basées sur la vectorisation que nous venons de décrire, mais elles ont été mises en pratique dans certaines publications dédiées au suivi d'événements dans la presse historique [Oiv+19].

Algorithme	Multilingue	Représentation des documents	Supervisé	Algorithme de base	Gestion du temps	Code disponible
TDT [All02b]	✓	*TF-IDF*, traduction vers l'anglais	✓	*Clustering* Agglomératif [SL99], K-Moyennes [LSS02]	Différence temporelle [LSS02]	
Europe Media Monitor [Pou+04]	✓	*TF-IDF*, *Eurovoc*, mots tels *tsunami*	✓	*Clustering* Agglomératif, spécifique	Fenêtres de 7 jours	
ClusStream [AY06 ; AY10]				Spécifique	Différence temporelle	
Suivi sur Twitter [PM10 ; Asg+21]		*TF-IDF*, noms propres			Différence temporelle	
Event Registry [Rup+16]	✓	*TF-IDF*	✓	K-Moyennes, CL-LSI, CCA	Différence temporelle	
newsLens [LH17]		*TF-IDF*	✓	Louvain [Blo+08]	Fenêtres de 7 jours, recouvrement 50%	
Suivi multisources [MBC19]		DTM [BL06]		K-Moyennes, LDA hiérarchique [Gri+03]	Fenêtres de 7 jours	
Streaming News [Mir+18 ; SMM22]	✓	*TF-IDF*, vecteurs dense	✓	*SVM-Rank* [Joa06]	Différence temporelle	Partiel
Suivi d'événements par lots [LH20]	✓	*TF-IDF*, vecteurs denses *S-BERT* [RG19]	✓	Louvain [Blo+08], Régression logistique	Fenêtres de 7 jours, recouvrement 50%	
Moteur de recherche [WJY22]		Brute		Moteur de recherche *ElasticSearch*, *Okapi BM25* [Rob+97]	Aucune	

TABLEAU 2.1 – Synthèse de quelques algorithmes cités capables d'opérer un suivi des événements mentionnés dans la presse.

2.2.3 Synthèse

Nous avons vu dans cette section différents algorithmes utilisés depuis vingt ans pour réaliser du suivi d'événements mentionnés dans des textes. En l'état actuel de nos connaissances, rien n'est spécifique aux documents historiques. Nous l'étudions dans cet

ouvrage. Nous retenons également deux méthodes parce qu'elles sont dissemblables, mais répondent à la même problématique. Les algorithmes développés dans le cadre du suivi d'événements dans des domaines ouverts (comme nous l'avons mentionné à la section précédente, la section 2.1) monopolisent la littérature scientifique à ce sujet. Les approches et techniques exploitant des graphes sont assez similaires à l'approche générique en domaine ouvert. Ces approches par graphes étudient la viralité de la propagation des informations plutôt que la modélisation d'histoires de presse. L'autre voie portée par la recherche sur la sémantique des données paraît prometteuse. De ces deux idées, nous retenons celle du moteur de recherche sémantique comme une piste pour identifier des documents relatant des événements historiques.

Nous avons également évoqué la question centrale de la représentation et de l'encodage de l'information pour le traitement numérique. L'apparition d'architectures d'apprentissage profond révolutionne les processus basés sur des pondérations *TF-IDF*. Le plongement numérique n'est pas l'unique solution à ce problème. D'autres approches sont basées sur la conservation des textes bruts, sans aucune modification, et utilisent des systèmes d'indexation textuels. De tels techniques, lorsqu'exploitées conjointement à des moteurs de recherche, tracent les événements dans les documents. Dans les deux cas, la représentation *TF-IDF* ou par texte brut est dépendante de la langue utilisée pour la rédaction du contenu. À ce niveau, le multilinguisme est complexe et repose sur des moyens détournés, comme les langues pivots. Pour simplifier cette tâche, les encodeurs sont entraînés sur des corpus multilingues et vectorisent des documents rédigés quelle que soit la langue. Capables de représenter de façon similaire des concepts exprimés dans des langues différentes, ils s'inscrivent dans la problématique de suivi d'événements au sein de corpus multilingues.

Nous proposons d'implémenter, dans le cadre de ces travaux, deux méthodes parmi les algorithmes et procédés décrits au tableau 2.1. Nous faisons le choix d'utiliser l'un des algorithmes de l'état de l'art pour du suivi en domaine ouvert. Ce processus est décrit et évalué au chapitre 4. Nous reprenons l'idée du moteur de recherche pour l'étendre et l'évaluer sur la recherche d'événements mentionnés dans du texte. C'est le sujet du chapitre 5. En ce qui concerne les vectorisations utilisées pour les algorithmes de suivi, nous proposons d'évaluer les pondérations *TF-IDF* ainsi que les vectorisations denses obtenues par apprentissage profond. Bien que des comparaisons aient été réalisées sur de la presse récente, son application à la presse historique est à explorer.

Nous avons analysé la notion d'événement puis nous avons décrit les différentes stratégies de la communauté scientifique pour reconstruire les histoires des événements mentionnés en presse écrite. Pour parachever cet état de l'art, la section suivante décrit les corpus que nous pouvons exploiter pour suivre des événements dans des articles de presse.

2.3 Jeux de données pour le suivi d'événements

2.3.1 Sélection des jeux de données expérimentaux

Historiquement, les premiers travaux scientifiques dédiés à la recherche et au regroupement d'événements dans des textes de presse datent de la fin du XX^e siècle, dans les années 1990. De ces projets précurseurs est né un nouveau champ de recherche scientifique. Avec les travaux relatifs à *MUC* [CLH93] et *TDT* [ALJ00] sont publiés des jeux de données largement utilisés par la suite à des fins expérimentales.

— *Message Understanding Conference* [CLH93]. Pour rappel, l'objet du projet *MUC* est de retrouver automatiquement, dans des documents, des réponses à des questions basiques permettant de définir l'action : quel est le type de document, quand l'événement a-t-il eu lieu, etc. Chaque jeu de données publié durant les campagnes (MUC-1, MUC-2, etc.) est spécifique à un domaine particulier. Pour MUC-3, ce sont des messages militaires.

— *Topic Detection and Tracking (TDT)* [ALJ00]. Les données sont des articles de presse nativement numériques, rédigés en anglais et en mandarin. Les annotations concernent des sujets macroscopiques, tels la finance, le sport, etc. Elles sont spécifiques à des événements [Cie+02] que l'on peut décrire en répondant aux questions « quoi », « qui » et « quand ». Trois jeux de données différents ont été publiés, TDT-1 pour le projet pilote puis TDT-2 et TDT-3. Ces derniers sont davantage utilisés pour des expériences d'identification et de suivi de sujets mentionnés dans la presse [Sta+19].

Dans les années suivantes, c'est le projet *Europe Media Monitor* (*EMM*) [Pou+04] qui reprend les acquis de *TDT*. Le jeu de données utilisé n'est pas publié et ne pouvait pas l'être compte tenu des restrictions liées au droit d'auteur [6]. Au début des années 2010, de nouveaux acteurs apparaissent. Le plus notable est le groupe de recherche qui conduit le projet *Event Registry*. Ce travail s'inscrit dans la lignée de *TDT* et de l'*EMM*. Ces équipes de recherche se focalisent sur le suivi de mentions d'événements dans des articles de presse, quelle que soit la langue de rédaction [Leb+14]. Ils publient en 2016 un jeu de données de plusieurs dizaines de milliers de documents [Rup+16] et annoté de milliers d'événements. Ces données ont permis des améliorations notables dans le développement d'algorithmes de suivi de mentions d'événements multilingues [Mir+18 ; Sta+19 ; LH20 ; SMM22]. D'autres travaux plus récents [MBC17 ; MBC19] se sont intéressés au suivi des mentions d'événements dans la presse par la fouille de sujet et des techniques de *topic mining*. Les auteurs ont partagé leur jeu de données prétraité. Les documents sources sont inaccessibles et nous ne pouvons l'exploiter dans le cadre de cette thèse.

Pour s'inscrire dans la continuité des travaux précédents sur la discipline, *Event Registry* est sélectionné pour les expériences présentées dans ce document. Aucun autre jeu de données n'est à notre connaissance utilisable dans le même contexte que peut l'être *Event Registry*. Il est également public, ce qui répond à l'une de nos contraintes : utiliser des données accessibles à tous pour faciliter la reproduction de nos résultats.

Néanmoins, un autre corpus contenant de documents de presse historique, *NewsEye*

6. Cette affirmation est issue d'une correspondance personnelle avec les auteurs principaux de l'article mentionné.

[New18], fait l'objet d'expériences de détection et d'extraction d'événements [Bor+22b]. Nous verrons plus avant, au chapitre 3, qu'il n'est pas utilisable dans notre contexte pour des raisons qui lui sont spécifiques. Les données annotées par des événements à sujets macroscopiques comme *TDT* [Cie+02] ou *Reuters* [05 ; SL18] ne conviennent pas non plus. Les événements de ces corpus ne sont pas conformes à la définition que nous avons proposée en section 2.1.

À la fin de la même décennie apparaît un nouveau virus, le SARS-Cov2 [Ben+20], dont la diffusion devient pandémique au début de l'année 2020. En parallèle, le volume d'informations de presse diffusé sur les réseaux sociaux numériques à ce sujet est croissant. La massification des données échangées sur les supports numériques marque la décennie : sites Web, réseaux sociaux, messageries instantanées. C'est l'avènement des données massives et de leur analyse. De façon concomitante et par voie de conséquence, le monde est submergé d'échanges de fausses informations relatives à la diffusion de ce virus. L'Organisation mondiale de la santé (OMS) décrète, dans les jours qui suivent la déclaration pandémique de la maladie Covid-19, l'émergence du phénomène d'« infodémie » [Org20a ; Org20b]. Ce mot-valise, composé de « information » et « épidémie », décrit le phénomène qui se joue ici : des informations fausses circulent de façon virale et incontrôlée sur les réseaux sociaux numériques. Les publications sont analysées et les faits vérifiés par des communautés de journalistes et de volontaires tout au long de la phase aiguë de la pandémie. Des sites Internet tels Factuel de l'Agence France-Presse (AFP) [Age22], *Snopes* [Sno22] ou *PolitiFact* [Pol22] référencent de fausses informations. Ils produisent des analyses détaillées des faits ainsi qu'une collection de liens vers des publications épinglées pour leur contenu. Chaque visiteur ou visiteuse est dès lors capable de déterminer si l'information qu'il consulte est véridique ou non.

C'est en marge de ces conflits informationnels que des projets de recherche sont initiés pour améliorer la détection d'informations fausses ou malhonnêtes sur les réseaux sociaux numériques [Pat+21 ; Ban+21 ; Wan+21b]. Pour créer les jeux de données, les publications sont croisées avec les plates-formes de vérification de faits. Chaque publication sur un réseau social est rattachée à un fait documenté et numéroté sur les plates-formes. Chaque identifiant de fait correspond à un événement précis et vérifié. En 2020 et 2021, des corpus sont partagés pour aider à détecter les fausses informations liées au Covid-19 sur les réseaux sociaux numériques. Certains sont multilingues, MM-Covid [Li+20], FakeCovid [SN20], contiennent des milliards de documents [IQO21 ; Ban+22] ou encore des informations de localisation [QIO20].

Ce type de jeux de données est exploitable pour suivre les mentions d'événements dans la presse. Certains corpus différencient deux types de publications : les réactions des utilisateurs et utilisatrices et les faits. Ce sont les partages de faits qui sont liés aux plates-formes de vérification évoquées précédemment. Chaque publication est associée à un événement (un fait rapporté par les organes de presse) et annotée par un identifiant unique renvoyant à cet événement. Sur les réseaux sociaux, les contenus partagés sont en général des brèves journalistiques rédigées ou non par des journalistes et des organes de presse. Compte tenu de leur longueur et du fait qu'elles peuvent répondre aux questions de base (quoi, quand, qui, comment, pourquoi…) certaines publications sur Twitter

peuvent être considérées comme des brèves de presse. Les internautes peuvent y ajouter des commentaires et interagir. Les corpus utilisés pour la détection de fausses informations intègrent donc des mentions d'événements, à travers les brèves journalistiques, ainsi que des identifiants qui les associent à des événements. Parce que cette construction répond à notre problématique, nous avons sélectionné deux jeux de données liés au Covid-19 pour mener nos futures expériences : *CoAID* et *FibVid*. Dans ces deux jeux de données, nous isolons les tweets de brèves de presse en retirant les réactions d'utilisateurs. Chaque publication est associée à un identifiant d'événement, ce qui permet de grouper tous les articles évoquant un fait particulier.

— *CoAID* [CL20]. C'est un jeu de données adapté à la détection de fausses informations liées au Covid-19 et publiées sur Twitter. Il est collecté du 1er décembre 2019 au 1er septembre 2020, durant la phase aiguë de la pandémie. Il contient une grande diversité de tweets, dont des brèves de presse et des réactions d'utilisateurs et d'utilisatrices (commentaires variés en rapport avec chaque publication). Il contient seulement des documents rédigés en anglais. Le tableau 2.2 en présente un extrait.

— *FibVid* [Kim+21]. Ce jeu de données est dédié à l'étude des fausses informations liées au Covid-19 sur Twitter. Les tweets sont collectés du 1er février 2020 jusqu'à la fin de l'année. Ils sont classés selon leur niveau de véracité et selon qu'ils sont en rapport ou non avec la pandémie du Covid-19. *FibVid*, contrairement à *CoAID* est constitué en majorité de réactions utilisateurs, en comparaison avec les publications de presse. Le contenu est seulement rédigé en anglais. Le tableau 2.3 en présente un extrait.

#1	HAPPENING NOW : the Senate Appropriations panel passed a measure that I introduced with @SenatorStefano that would implement a statewide moment of silence in schools, observing September 11th. After the unanimous vote, SB869 moves to the full Senate for consideration. @PASenateGOP
#2	Contact tracing is receiving the most media attention. But we should be talking more about surveillance to improve early detection of potential COVID-19 outbreaks, protect vulnerable populations, and use testing resources smartly.
#3	"The American Academy of Pediatrics strongly advocates that all policy considerations for the coming school year should start with a goal of having students physically present in schools."

TABLEAU 2.2 – Exemples de tweets publiés dans *CoAID*, sélectionnés au hasard.

Dans cette section, nous avons sélectionné trois jeux de données : *Event Registry*, *CoAID* et *FibVid*. Ils répondent aux impératifs de suivi d'événements. D'abord, ils contiennent exclusivement des données de presse. Ensuite, qu'il s'agisse d'articles complets ou seulement de brèves, ce sont des textes rédigés par des auteurs qui manient le style singulier de la presse. Chaque document contient également une annotation d'événement dont la granularité est fine et en phase avec la définition d'événement que nous proposons en section 2.1. Des métadonnées, les dates de publications et les sources sont incluses. Pour mieux comprendre les corpus sélectionnés et identifier de potentiels biais,

#1	Joe Biden will again politicize coronavirus today. But his record on pandemics is one of incompetence. During the 2009 swine flu outbreak, Biden made reckless comments unsupported by science the experts. The Obama Admin had to clean up his mess ; apologize for his ineptitude.
#2	On the direction of Nancy Pelosi, portraits of 4 Speaker of the House who served in the Confederacy have been removed from public view placed into storage in Congress. They are : Robert Hunter (18391841) Howell Cobb (18491851) James Orr (18571859) Charles Crisp (18911895)
#3	The White House is launching a communications plan across multiple federal agencies that focuses on accusing Beijing of orchestrating a "coverup" and creating a global pandemic, according to two U.S. officials and a cable obtained by The Daily Beast.

TABLEAU 2.3 – Exemples de tweets publiés dans *FibVid*, sélectionnés au hasard.

une analyse exploratoire est nécessaire.

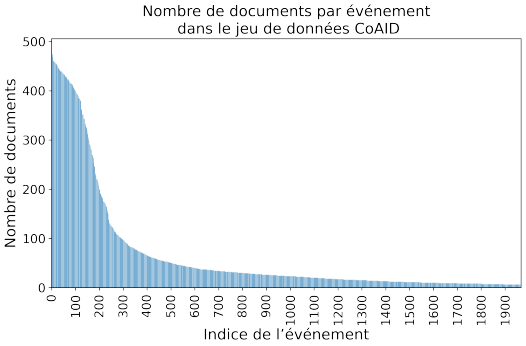
2.3.2 Analyse exploratoire des données

Un jeu de données collecté en sources ouvertes et annoté manuellement ou assisté par ordinateur comporte nécessairement à la fois biais et erreurs. Ces défauts sont multiples et dépendent des données collectées. Dans le cas présent ce sont des articles de presse. Chacun est associé à un identifiant qui le rattache à un événement spécifique. Tous les documents rattachés au même identifiant décrivent le même événement duquel découle une histoire : c'est l'histoire de l'événement raconté dans la presse.

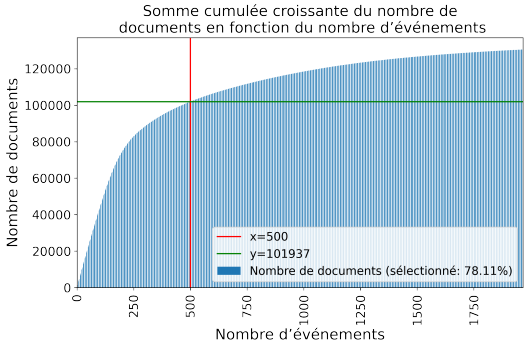
Une limite importante des annotations de ce type réside dans une incapacité à traiter manuellement les gros volumes de documents. Considérons la consigne d'annotation suivante : *« lisez le document ci-dessous et rattachez-le à l'un des x événements connus. S'il s'agit d'un événement inconnu jusqu'alors, créez-lui un identifiant »*. Ce type d'annotation implique une connaissance fine du jeu de données et de l'ensemble des événements connus à un instant t donné. Dans ce cas de figure, il est possible que deux documents soient rattachés à deux identifiants différents alors qu'ils traitent d'un même événement. Pour ce type d'annotation, la date de publication, l'état mental de l'annotateur ou annotatrice, son niveau d'expertise (candide ou au contraire spécialiste), la longueur des documents peuvent entraîner des biais variés. Ils sont liés par exemple à la sélection des documents au sein des jeux de données, à la répartition temporelle des articles publiés, à la longueur des documents ou encore au caractère multilingue de certains événements décrits. Seule une analyse exploratoire précise permet de comprendre la structure des données, ses qualités ainsi que les défauts qui peuvent altérer nos conclusions si nous les utilisons. Nous publions les codes sources de l'analyse présentée ici sur Internet, en accès libre [Ber22a].

Analyse et filtrage de jeux de données pour le suivi d'événements

Les caractéristiques d'*Event Registry*, de *CoAID* et de *FibVid* les distinguent. Les langues utilisées pour la rédaction des articles, le nombre d'événements global, la quantité



(a) Nombre de documents par événement.



(b) Somme cumulée croissante du nombre de documents par événement.

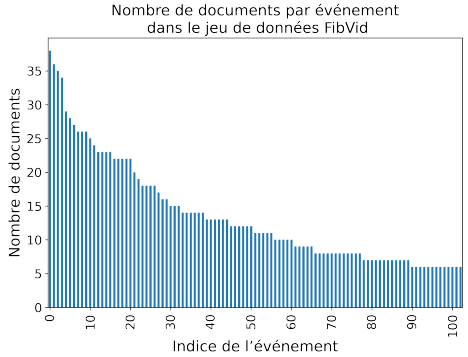
FIGURE 2.4 – Aperçu des quantités de documents par événement dans CoAID.

de documents ou encore la taille des événements sont différents.

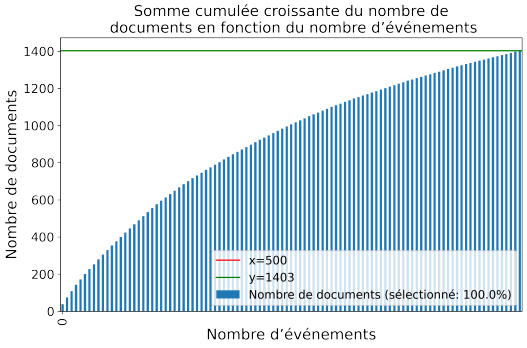
CoAID et *FibVid* sont créés pour répondre à des objectifs de recherche autres que ceux étudiés dans cette thèse. Ici, l'objet d'étude principal est l'événement, représenté par des documents, les articles de presse. Les figures 2.4 et 2.5 présentent le nombre de documents des corpus, en fonction de chaque événement. Plus la colonne est haute, plus un événement sera décrit par un nombre élevé de documents. Les seuls événements qui contiennent au moins cinq documents sont conservés, les autres éliminés. Ce seuil est choisi arbitrairement d'après le jeu de données *Event Registry*, dans lequel rares sont les événements décrits par moins de cinq documents. Dans les jeux de données *TDT* 2 et 3, cette limite est de quatre documents [Cie+02]. De plus, en dessous de ce seuil, la problématique relève plutôt de la détection de signal faible. Un événement évoqué dans un nombre limité de documents peut représenter un signal d'émergence d'une nouvelle information. Une petite quantité de documents décrivant un événement rend également impossible sa description détaillée, par manque de données.

Event Registry est utilisé dans la littérature scientifique pour mener des expériences de détection et de suivi d'événements dans la presse. Les auteurs de l'étude originale [Rup+16] ont agrégé des articles sur une période couvrant les années 2014 et 2015 à partir de sources variées publiées en trois langues : anglais, allemand et espagnol. Le jeu de données est utilisable tel quel. Pour garantir la bonne comparaison de nos travaux avec les expériences passées et futures, ne pas l'altérer est primordial. À titre d'information, ce sont 33 807 articles de presse et 1 467 événements que contient *Event Registry*. Une éventuelle mise de côté des événements contenant moins de cinq articles réduirait le nombre de documents à 33 006 (-801) et le nombre d'événements à 1 305 (-162, en moyenne 4,9 documents par événement). Cela confirme l'absence de petits événements de moins de cinq documents dans *Event Registry*. Les événements qui seraient éliminés contiennent en moyenne presque cinq documents. Comme expliqué précédemment, nous ne retirons aucun document d'*Event Registry*.

Pour *CoAID*, présenté en figure 2.4, le nombre d'événements est remarquable : il y



(a) Nombre de documents par événement.



(b) Somme cumulée croissante du nombre de documents par événement.

FIGURE 2.5 – Aperçu des quantités de documents par événement dans FibVid.

a 2 702 événements différents pour 132 403 documents. Parmi ceux-là, 734 événements contiennent au maximum cinq tweets et sont écartés. Ce sont 1 896 documents qui sont mis de côté par ce procédé. Dans un second temps et spécifiquement pour *CoAID*, nous opérons une seconde réduction, cette fois du nombre d'événements. D'après la courbe dessinée en figure 2.4a, le nombre de documents par événement décroît très fortement jusqu'environ cinq cents événements. Cette forte décroissance est très limitée au-delà de cinq cents événements et forme un coude. Cette limite du nombre d'événements est choisie pour ne sélectionner qu'une quantité restreinte d'événements. Après réduction, 80 % de la totalité des documents du corpus original est conservé, comme présenté en figure 2.4b. Les données restantes sont utilisées pour préparer les modèles de pondération *TF-IDF* de Tweets décrits à la section 4.1.

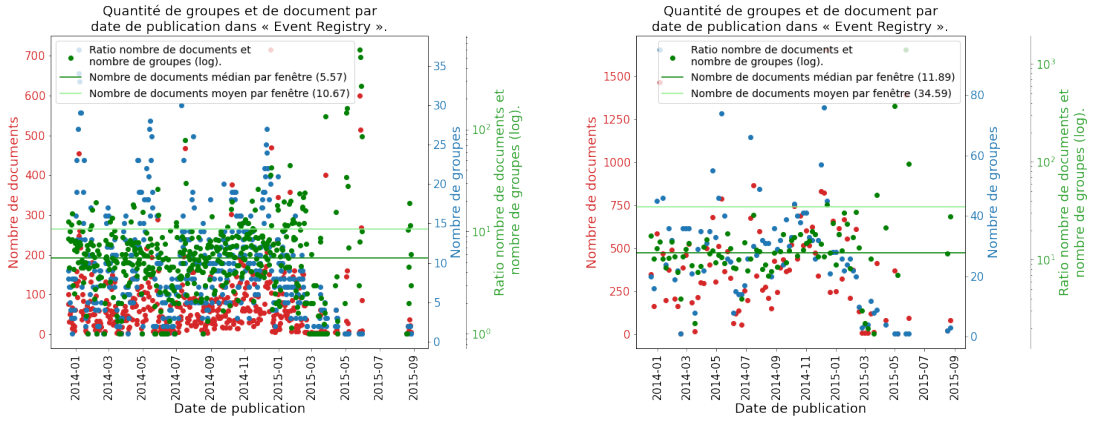
Dans *FibVid* présenté en figure 2.5, il y a 1 774 tweets et 295 événements au total. *FibVid* est un jeu de données de tweets contenant en majorité des réactions d'utilisateurs et d'utilisatrices. Le nombre de tweets de brèves est faible en comparaison au nombre de réactions. Ces dernières ne sont pas adaptées à notre objet d'étude : ce ne sont pas des partages de documents de presse. Les réactions sont écartées. Contrairement à *CoAID*, *FibVid* contient des tweets liés au Covid-19, mais pas seulement : d'autres sujets y sont évoqués. L'élimination des plus petits événements entraîne la suppression de 371 tweets et de 103 événements. Aucune autre réduction liée au nombre d'événements n'est réalisée devant la faible quantité de données restantes.

Ce sont trois jeux de données aux caractéristiques variées que nous avons sélectionnés. D'abord l'un des mètres étalons, largement utilisé : *Event Registry*. Ensuite, nous exploitons les jeux de données *CoAID* et *FibVid* en les adaptant à des problématiques de suivi d'événements, en éliminant les plus petits événements et en réduisant la quantité de données de *CoAID*.

Répartition des documents dans le temps

Event Registry est composé d'articles et de leurs titres, *CoAID* et *FibVid* de brèves de presse partagées sur les réseaux sociaux. L'action de publication est inscrite dans un temps donné. Les documents nativement numériques sont publiés avec une précision de l'ordre de la minute et l'horodatage est en général précisé au sein même de l'article. Le lectorat sait à quel moment de la journée l'information a été partagée. L'échelle de temps est fine : une information publiée le matin à dix heures peut devenir obsolète et être invalidée dans l'après-midi. La presse écrite imprimée est publiée selon un rythme au mieux quotidien. La publication des articles se fait sur des échelles de temps plus grandes, à date près et non à la minute. La notion de temps de publication est donc différente entre ces deux supports, le papier et le numérique. Le cas particulier du temps dans la diffusion des documents historiques sera étudié en section 3.1. Tous les jeux de données sélectionnés sont nativement numériques et l'horodatage des publications a une précision à la minute près.

Les figures 2.6, 2.7 et 2.8 présentent les répartitions du nombre de documents et d'événements dans le temps, respectivement sur l'axe des ordonnées et des abscisses. Dans un premier cas, l'échelle de temps est fixée au jour. Dans l'autre, des fenêtres de sept jours englobent les articles publiés sur une semaine, comme cela pourrait être le cas avec la publication d'un hebdomadaire, en opposition à un quotidien. Notons que cette fenêtre de sept jours est celle utilisée par les auteurs du projet *newsLens* [LH17 ; LBH21] dans leur algorithme de suivi de mentions d'événements. Elle est supposée représenter une bonne division permettant d'obtenir des groupes de documents cohérents pour la reconstruction d'histoire d'événements dans la presse.



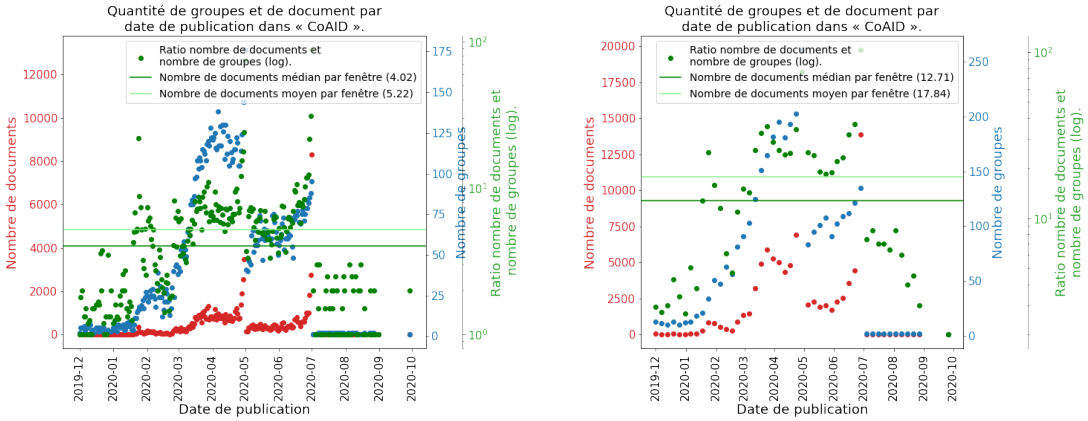
(a) Pour chaque jour.

(b) Pour chaque fenêtre de six jours.

FIGURE 2.6 – Répartition du nombre de documents et du nombre d'événements en fonction du temps dans *Event Registry*.

Pour *Event Registry* présenté en figure 2.6, la répartition dans le temps est assez équilibrée : la publication est continue sur l'ensemble des dates couvertes, à l'exception

des mois de juin à août 2015. On peut retrouver dans cette homogénéité un rythme de publication (indiqué par une quantité de documents sur une plage donnée) propre à la presse écrite imprimée. Dans celle-ci, le nombre de pages et la mise en page sont relativement stables. Le volume d'articles publié peut varier, mais uniquement par la réduction ou l'augmentation de la taille des articles : deux textes courts peuvent prendre la place d'un long. Il est donc raisonnable de s'attendre à une répartition uniforme dans le temps du nombre d'articles lors du traitement de documents de presse imprimés et numérisés. Le nombre d'articles (en bleu sur la figure) est parfois très important par rapport au nombre de documents (en rouge). Le ratio (en vert), parfois élevé entre ces deux informations, rend compte d'un faible nombre d'événements pour un nombre important d'articles. Si ce constat s'applique à chaque plage temporelle (jour ou semaine), le jeu de données présente un biais majeur : il devient possible de regrouper les documents en considérant que chaque jour ou semaine représente un unique événement. Cet indicateur de ratio marque la diversité des événements à chaque plage temporelle. En moyenne, ce sont dix documents par événement pour un intervalle quotidien, trente-quatre pour un intervalle de semaines. Les valeurs de médianes sont respectivement de 5,6 et de onze.



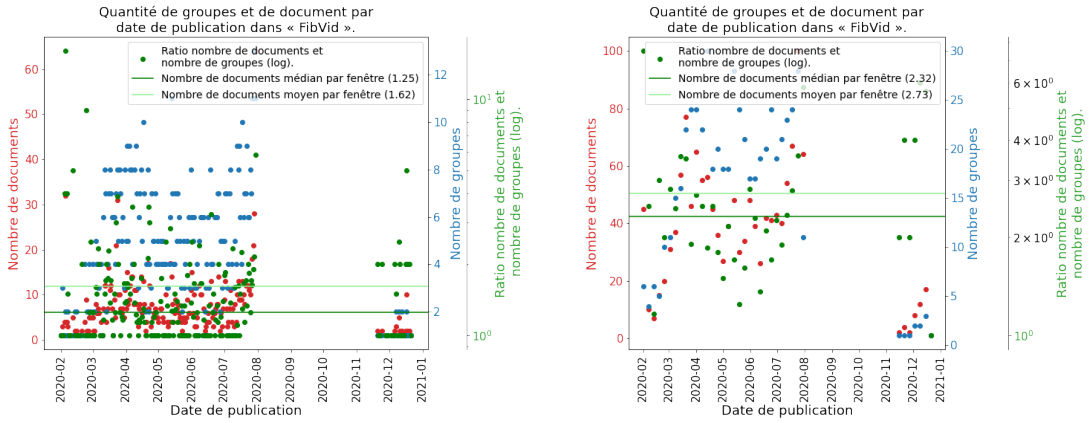
(a) Pour chaque jour.

(b) Pour chaque fenêtre de six jours.

FIGURE 2.7 – Répartition du nombre de documents et du nombre d'événements en fonction du temps dans *CoAID*.

L'analyse des contenus des tweets de *CoAID* montre une autre répartition, présentée à la figure 2.7. Les créateurs du corpus *Event Registry* semblent avoir prêté attention à la réduction de ce biais. Cependant, l'adaptation des données *CoAID* à d'autres fins que la détection de fausses informations peut mener à divers effets de bord. Tout d'abord, la répartition des publications dans le temps est homogène, il n'y a pas de vide : des articles sont publiés tout au long de la période étudiée. La quantité des publications est au contraire inégale dans le temps. Deux pics de publications sont visibles à la fin du mois de mai puis durant le mois de juillet 2020. Bien que le nombre d'articles soit parfois très élevé, le nombre d'événements rapportés dans ces articles a tendance à légèrement lisser ce phénomène. Le nombre d'articles moyen par événement est de cinq par jour et de dix-

huit par semaine avec des médianes à respectivement quatre articles et douze. Comme évoqué précédemment dans la section 2.3.2, le nombre total d'événements rapportés dans *CoAID* est peu élevé, mais couvre de grandes plages temporelles. Après sélection des cinq cents plus gros événements en nombre d'articles, on constate que pour certains jours, ce sont jusque 125 événements traités (mois d'avril et de mai 2020). C'est un quart de tous les événements rapportés.



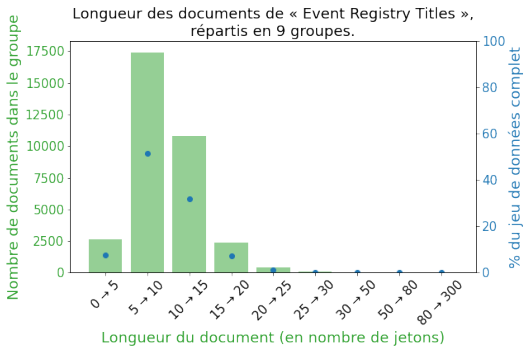
(a) Pour chaque jour.

(b) Pour chaque fenêtre de six jours.

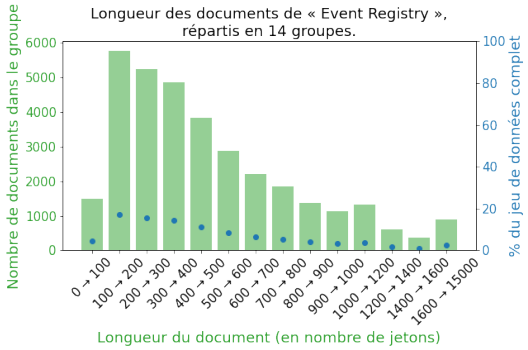
FIGURE 2.8 – Répartition du nombre de documents et du nombre d'événements en fonction du temps dans *FibVid*.

Pour *FibVid* présenté en figure 2.8, le nombre global de documents est faible et les publications ne sont pas uniformément réparties sur l'année. Les données ont cessé d'être collectées entre les mois d'août et décembre 2021. Il y a pour certaines dates et sur une échelle de jours et non de semaines, un nombre de publications qui est parfois égal au nombre d'événements mentionnés. Sur ces données, grouper les documents suivant les jours et non le contenu peut donner des résultats pertinents par rapport à notre objectif : reconstruire les chronologies des événements. Le nombre moyen de documents par jour et par événement est légèrement supérieur à un, de même pour la médiane. Le regroupement des articles sur des intervalles de semaines limite légèrement ce phénomène, en augmentant le nombre de documents par événement à 2,3 et 2,7 pour la médiane.

Dans tous les cas, cette répartition par intervalles de temps, jour ou semaine ne tient pas compte du nombre global d'événements. Pour *Event Registry*, ce sont 1 467 événements différents qui sont décrits dans l'ensemble des données, avec une répartition assez homogène dans le temps. Pour *CoAID*, le nombre d'événements traité est limité artificiellement à cinq cents, éliminant les plus petits clusters répondant mal à la problématique d'identification d'événements. Nous avons fait le constat que sur certains jours, une grande proportion d'événements est décrite dans les documents, jusque 25 % du total. Pour *FibVid* enfin, c'est le faible nombre de documents et d'événements qui constitue le biais le plus important par rapport à la répartition temporelle des données.



(a) Nombre de documents du corpus en fonction du nombre de termes du titre.



(b) Nombre de documents du corpus en fonction du nombre de termes de l'article.

FIGURE 2.9 – Nombre de documents du corpus *Event Registry* en fonction du nombre de termes qu'ils contiennent.

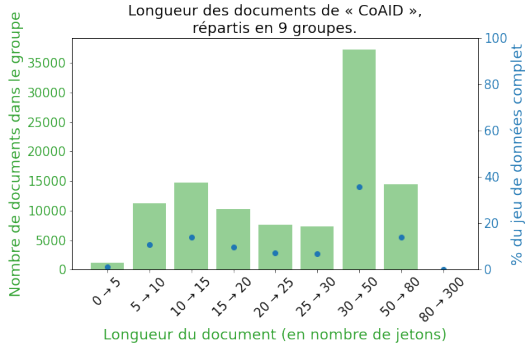
Longueur des documents

Les textes, leur qualité et leur quantité sont déterminants pour en extraire les événements mentionnés, les classer et les ordonner. En termes de longueur de contenu, les jeux de données *CoAID* et *FibVid* sont comparables. Dans les deux cas, ce sont des tweets partageant des articles de presse. La taille maximale est de fait limitée à 280 caractères [Rn17], en vigueur depuis 2017 et durant les différentes périodes de collecte des données. Pour *Event Registry*, les titres et les articles sont présents. Alors que les titres sont nécessairement plus courts, les articles présentent des longueurs variables, pour certains, plus de dix mille mots.

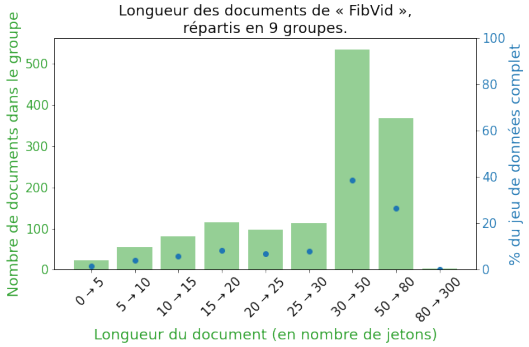
Par longueur du texte, nous devons comprendre le nombre de mots ou de termes reconnus par les logiciels de segmentation automatique de texte. Nous utilisons le logiciel `spaCy` [Sof22] et une implémentation dédiée [Ber21e] qui extrait le nombre de termes, à la fois en anglais, pour les trois jeux de données et en espagnol et en allemand pour les documents rédigés dans ces langues au sein d'*Event Registry*.

Event Registry contient deux types de textes : les titres des articles et leurs corps. Les titres sont en général très courts : près de 95 % des titres contiennent moins de vingt termes et près de 50 % des titres sont rédigés avec entre cinq et dix termes seulement. La longueur des textes des articles est plus variable. La quasi-intégralité des articles contient plus de cent termes et 50 % des articles sont écrits avec entre cent et quatre cents termes. Au-delà, le nombre d'articles plus longs décroît fortement et rares sont les documents rédigés avec davantage de mille termes.

Les profils de longueur des tweets de *CoAID* et *FibVid*, présentés respectivement en figure 2.10a et figure 2.10b sont similaires. La majorité des tweets ont une longueur comprise entre trente et quatre-vingts termes. Contrairement aux titres d'*Event Registry*, les textes de *CoAID* et *FibVid* sont bien plus longs et la différence nette : rares sont ceux d'*Event Registry* avec plus de trente termes. Une première différence de forme distingue



(a) Nombre de documents du corpus *CoAID* en fonction du nombre de termes du texte.



(b) Nombre de documents du corpus *FibVid* en fonction du nombre de termes du texte.

FIGURE 2.10 – Nombre de documents de corpus de tweets en fonction du nombre de termes du texte.

les textes publiés dans *CoAID* et *FibVid* et les titres des articles d'*Event Registry*.

Événements multilingues

Event Registry est le seul jeu de données contenant des articles rédigés dans plusieurs langues. Certains événements sont décrits à la fois en anglais, en allemand et en espagnol. Les événements identiques sont rattachés au même identifiant unique. Le tableau 2.4 synthétise les nombres de documents en fonction des langues de rédaction et du nombre d'événements. La somme des événements en anglais, allemand et espagnol est supérieure à la somme totale des événements : la différence représente le nombre d'événements multilingues. Pour la suite, on qualifiera de bilingue un événement rapporté dans deux langues différentes et de trilingue un événement rapporté dans trois langues.

Parmi les 1 467 événements, la figure 2.11 présente le nombre de langues dans lesquelles les événements sont décrits. La majorité des événements (1 100, 74 %) sont rédigés dans une seule et unique langue. Au sein d'*Event Registry*, 374 (26 %) des événements sont bilingues et 6 (< 1 %) sont trilingues. Pour les événements bilingues, 151 (39 %) le sont en anglais et allemand, 75 (19 %) en allemand et en espagnol et 166 (42 %) en anglais et en espagnol. Ces proportions, non homogènes, peuvent biaiser les résultats d'un algorithme qui analyse les événements multilingues dans le but de rapprocher les événements identiques et décrits dans des langues différentes. Les figures 2.12, 2.13 et 2.14 rendent compte des quantités de documents décrivant les événements bilingues. La dernière, la figure 2.15 présente le nombre de documents dans chaque langue pour les six événements trilingues.

Pour chaque paire de langues, les couples anglais - allemand, allemand - espagnol et anglais - espagnol, respectivement représentés par les figures 2.12, 2.13 et 2.14, deux phénomènes indépendants sont notables. Le nombre de documents des événements bilingues est systématiquement supérieur en anglais comparé à l'allemand ou à l'espagnol : il se

Langue	Documents	Événements	Ratio docs./événement
Indifférenciée	33 807	1 467	22,84
Allemand	6 144	490	12,41
Anglais	20 959	808	25,71
Espagnol	6 704	554	11,86

TABLEAU 2.4 – Le nombre de documents et d'événements d'*Event Registry*.

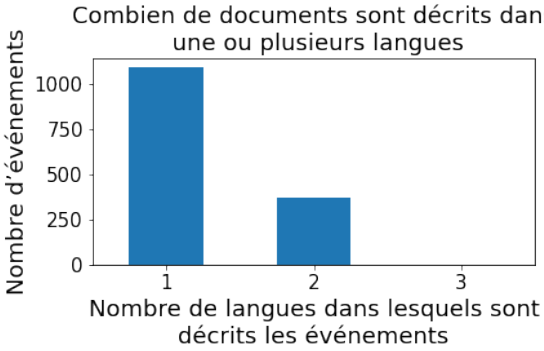


FIGURE 2.11 – Nombre de langues dans lesquelles les événements sont décrits.

situe entre 37 et 40 documents en moyenne par événement contre 15 pour les deux autres langues. Le couple allemand et espagnol montre une autre tendance. Les événements bilingues rapportés dans ces deux langues contiennent en moyenne un nombre identique de 19 documents. Du point de vue du traitement multilingue, les langues allemandes et espagnoles sont comparables. Un biais existe, conséquence de la surreprésentation de l'anglais dans les événements bilingues, lié à un nombre de documents par événement plus élevé.

Six événements sont trilingues. C'est peu en proportion de l'ensemble des événements (< 1 %) d'*Event Registry*. Pour les événements trilingues, une disparité d'équilibre entre les langues existe, visible dans la figure 2.15. Quatre d'entre eux sont associés à des documents en allemand (indices 2 à 5) dont la quantité est insignifiante comparé à ceux rédigés en espagnol ou en anglais. Pour les deux autres événements, d'indices zéro et un, la présence de l'allemand est notable et équilibrée avec l'espagnol. L'anglais l'emporte toujours comme étant la première langue contributrice, en nombre de documents, aux événements multilingues.

Les événements multilingues d'*Event Registry* sont certes peu nombreux au regard de l'ensemble des événements, mais les annotations d'événements fournies ne sont pas non plus dépourvues d'erreurs ou de défauts.

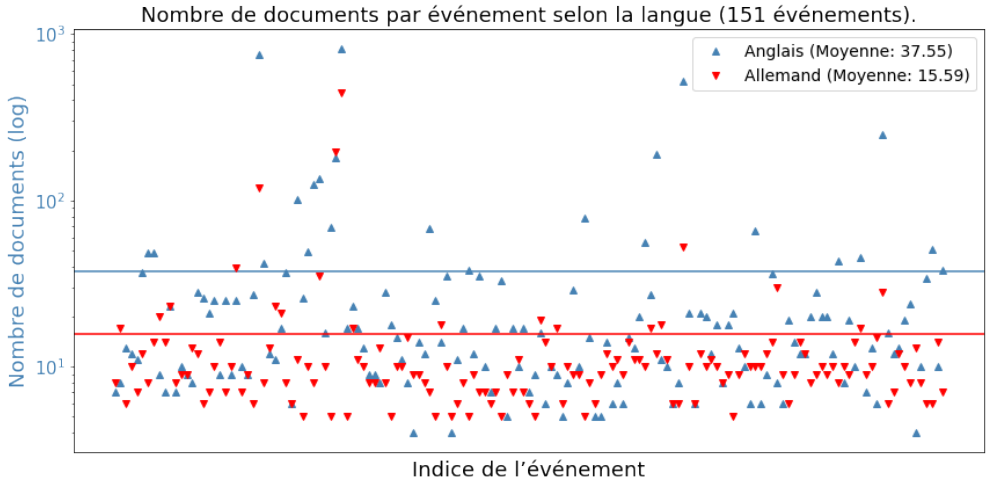


FIGURE 2.12 – Pour chacun des événements décrits en anglais et en allemand, le nombre de documents associés à l'événement.

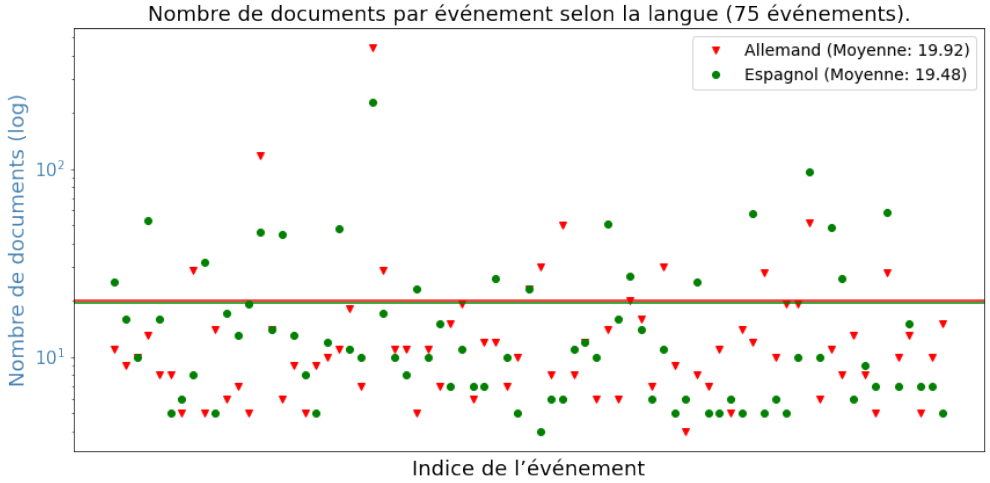


FIGURE 2.13 – Pour chacun des événements décrits en allemand et en espagnol, le nombre de documents associés à l'événement.

Erreurs d'annotations ou défauts des documents

Les trois jeux de données sont annotés avec des indices d'événements selon deux processus indépendants. L'annotation de *CoAID* et *FibVid* repose sur les sites de vérification de faits. Les identifiants des faits sont des numéros uniques et chacun est considéré comme un événement. Pour *Event Registry* les auteurs mentionnent que les événements ont été annotés manuellement [Rup+16]. Certains événements contiennent des erreurs d'annotation, que l'on peut classer en différentes catégories et dont certaines

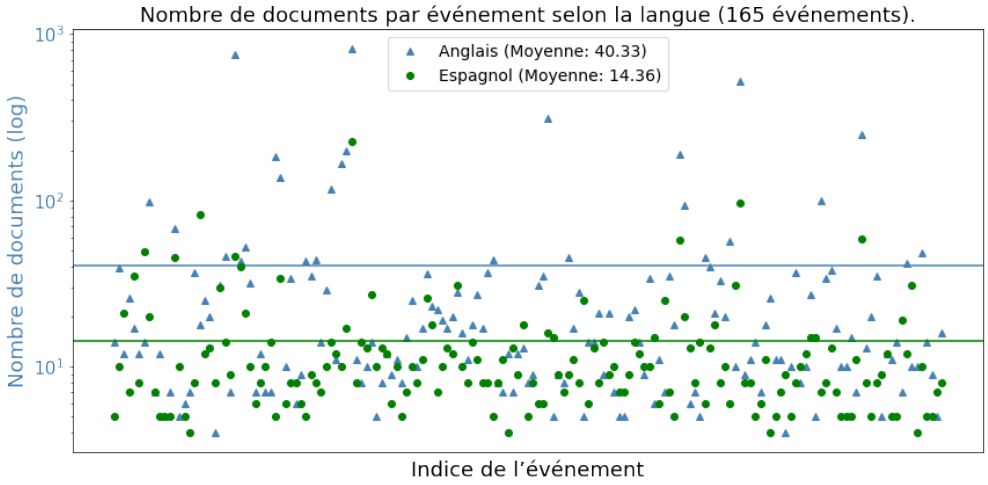


FIGURE 2.14 – Pour chacun des événements décrits en anglais et en espagnol, le nombre de documents associés à l'événement.

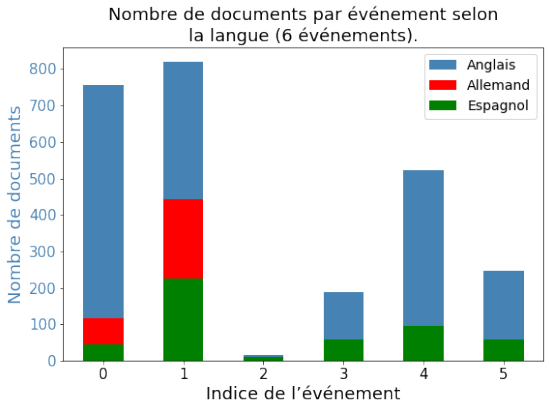


FIGURE 2.15 – Pour les six événements trilingues, le nombre de documents de chaque langue.

sont présentées dans le tableau 2.5. C'est aussi la nature des données collectées qui peut poser question, avec parfois la présence de doublons. Nous nous attachons ici à présenter quelques défauts d'annotation, qui ne sont pas représentatifs de l'ensemble des données. Une analyse quantitative serait pourtant nécessaire afin d'identifier les biais potentiels qui pourraient y être liés.

Un premier phénomène s'observe relativement fréquemment : les documents doublons. Il s'agit de documents associés au même événement, provenant parfois de sources différentes, avec des identifiants différents mais au contenu presque identique. Des exemples sont présentés dans le tableau 2.6. L'événement choisi est le discours de l'état de l'Union de Barack Obama de 2014. La liste des articles présentés dans ce tableau n'est pas ex-

Type	Description	Événements concernés
1	Doublons de contenu (réplication de contenu d'agence de presse)	1196, 1233, 1197
2	Article(s) sans rapport avec le reste du groupe	619, 1155
3	Contenu rédigé dans une autre langue que celle annoncée	76, 77, 1315, 987

TABLEAU 2.5 – Typologie de quelques défauts des données et des annotations d'*Event Registry*.

haustive : d'autres correspondent dans *Event Registry* à cet événement. Trois articles ont un titre identique. La distance d'édition entre les deux premiers corps de texte est faible (537) par rapport à leur longueur. Ils sont longs de respectivement 8 932 et 8 498 caractères. Le premier est une version enrichie et allongée du second. Les doublons sont régulièrement présents dans les données. Nous émettons l'hypothèse qu'ils sont une conséquence du travail des agences de presse. Ces acteurs fournissent à des journaux, sites de presse en ligne ou autres des brèves qui sont publiées et partagées avec peu de modification. En France, c'est l'Agence France-Presse qui fournit par exemple une partie des contenus publiés par la rédaction de France TV Info, Médiapart ou Sud-Ouest, pour ne citer qu'eux. Ce phénomène de doublon est également présent dans les jeux de données *CoAID* et *FibVid* qui contiennent principalement ce type de données.

Identifiant	Titre	Texte
3944828	Obama vows to flex presidential powers in speech	WASHINGTON (AP) – Seeking to energize his slu...
3944803	Obama vows to flex presidential powers in speech	WASHINGTON (AP) - Seeking to energize his slu...
3943806	Obama vows to flex presidential powers in speech	Seeking to energize his sluggish second term...
3940407	KEY POINTS : State Of The Union Address	- President Barack Obama said in his State of...

TABLEAU 2.6 – Doublons apparents de titre et de texte pour des documents différents (type 1).

Un autre type de cas concerne les intrus, comme le montre le tableau 2.7. L'événement rapporté est l'édition 2014 du Dakar en Amérique du Sud. Alors que tous les documents traitent de cet événement, un intrus évoque l'importation du blé dans les minoteries brésiliennes. Ce document contient des mots clefs communs avec les documents rapportant réellement l'événement. La présence de noms de pays sud-américains est une explication possible de l'erreur d'annotation. Les intrus sont cependant rares sur l'ensemble des événements d'*Event Registry* que nous avons parcourus pour cette analyse. Le cas échéant, c'est au maximum deux intrus qui sont détectés par événement.

Le troisième type de cas est lié au multilinguisme, déjà évoqué en section 2.3.2. Il peut exister une différence entre la langue indiquée dans les données et la langue effective du titre ou du texte du document. Le tableau 2.8 présente le cas de documents indiqués

Identifiant	Titre	Texte
1588500	Sousa and Barreda won prelude stage of dakar	Sousa war von Rosario nach San Luis der Schnel...
1880245	Bolivians won't block Dakar Rally on salt flats	LA PAZ, Bolivia (AP) An Aymara Indian group sa...
1903823	**Brazil's mills turn to U.S. wheat on Argentine...**	**Brazilian flour mills are importing nearly all...**

TABLEAU 2.7 – Au milieu des autres documents, l'un d'entre eux évoque un sujet complètement différent (type 2).

comme rédigés en allemand mais dont les titres sont dans les faits rédigés en anglais.

Identifiant	Langue	Titre	Texte
11841483	Allemand	European Court of Justice : Germany loses dispu...	Luxemburg. Deutschland darf seine Grenzwerte...
11895691	Allemand	Kadenbach : european rules protect from toxic c...	SPÖ-Europaabgeordnete begrüßt Urteil des Euro...
12110592	Allemand	Heavy metal in toys : eu's borders valid	AFP/Adrian Dennis - Deutschland muss EU-Schwer...

TABLEAU 2.8 – La langue annoncée du document, l'allemand n'est pas forcément celle utilisée, comme ici dans le titre (type 3).

Des erreurs d'annotations ou de construction des jeux de données influent nécessairement sur la qualité des documents, quelle que soit l'expérience menée. Dans le cas d'algorithmes d'apprentissage, une division en ensembles d'entraînement et de test peut être un important vecteur d'erreurs.

Division des données en jeux d'entraînement et de test

Les algorithmes basés sur des méthodes d'apprentissage ont besoin de données d'entraînement et de validation pour déterminer et valider un modèle algorithmique qui répond à l'objectif choisi. Pour qu'*Event Registry* soit utilisé avec de tels algorithmes, Miranda et coll. [Mir+18] l'ont fragmenté en un ensemble d'entraînement et un autre de test. Les données sont divisées suivant les événements, puisqu'aucun événement n'est présent à la fois dans le jeu d'entraînement et de test. Plus tard, Linger et coll. [LH20] ont proposé une division supplémentaire du jeu d'entraînement en deux ensembles de taille égale, un pour l'entraînement et un pour le développement. La division y est également réalisée par rapport aux événements. Les auteurs ont divisé la liste des identifiants d'événements selon les proportions désirées et ensuite formé les jeux de données en collectant tous les documents rattachés aux événements retenus. Rupnik et coll. [Rup+16] ont publié le corpus *Event Registry* puis Miranda et coll. [Mir+18] une division d'entraînement et de test. Ils sont utilisés par d'autres auteurs sur la même thématique [Sta+19 ; LH20]. C'est de cette répartition que nous utilisons dans le cadre de cette thèse pour permettre une comparaison des résultats obtenus avec les expériences passées.

Nous proposons d'utiliser les corpus *CoAID* et *FibVid* dans un cadre expérimental différent de ce pour quoi ils sont conçus et partagés. Aucune division d'entraînement et

de test n'est disponible à ce jour, elle doit être réalisée manuellement. La répartition suivie est de deux tiers de documents pour les données d'entraînement et d'un tiers pour les données de test. Une contrainte supplémentaire est de ne pas répartir un même événement dans les deux sous-ensembles, contrainte déjà évoquée pour *Event Registry*. Une implémentation est proposée [Ber21g]. Le tableau 2.9 synthétise ces répartitions. L'équilibrage prioritaire est réalisé sur les documents, de sorte que leur nombre est équilibré autour de 60 % - 70 % pour les données d'entraînement et le reste pour les données de test. Les trois divisions de jeux de données intègrent cette contrainte. La répartition en nombre d'événements est similaire pour *Event Registry* et *CoAID* avec des proportions de trois quarts pour l'entraînement et un quart pour le test. Pour *FibVid*, il y a autant d'événements dans l'ensemble d'entraînement et de test, même si le nombre de documents quant à lui est divisé comme souhaité.

Jeux de données	Type	Documents	Événements	Ratio événements	Ratio documents
Event Registry	Entraînement	20 803	1108	74,5%	61,5%
	Test	13 004	381	25,5%	38,5%
CoAID	Entraînement	72 045	375	75%	69%
	Test	32 100	125	25%	31%
FibVid	Entraînement	988	51	51,5%	71%
	Test	402	52	49,5%	29%

TABLEAU 2.9 – Statistiques de répartition des jeux de données en entraînement et en test.

2.3.3 Synthèse

Nous avons présenté ici les trois jeux de données utilisés dans la suite de ce document. Ils permettent de répondre à la problématique du suivi de mentions d'événements présentée en section 1.1. Quelques jeux de données existent pour le suivi de mentions d'événements dans des documents de presse numérique. Il s'agit souvent d'une vision d'événement davantage associée à l'idée de sujet, évoquée dès les débuts du programme *TDT*. Cette vision, très large, classe tous les événements relatifs au sport dans une catégorie éponyme, de la même façon que ceux relatifs à la politique, à la finance, etc. Cette représentation peu bornée ne correspond pas à la définition retenue ici (définition n° 10) : celle d'événements spécifiques, ancrés dans une unité de temps et d'espace et décrivant des actions faisant intervenir un ou plusieurs participants. Le jeu de données *Event Registry* est construit pour répondre à cette problématique. Pour contrer cette limite évidente liée au manque de données, nous proposons d'étendre les usages des corpus utilisés pour la détection de fausses informations sur Twitter, liés à des faits, ou à des événements précis. *CoAID* et *FibVid* sont sélectionnés dans ce but. Loin d'être parfaits, ils présentent certaines caractéristiques similaires à *Event Registry* parce qu'ils contiennent des publications de brèves de presse sur les réseaux sociaux numériques.

Tous trois présentent des biais, plus ou moins importants, liés à leur type, au mécanisme ou à la date de la collecte, ou encore à leur contenu. Certains sont traités dans

cette section, telle l'influence de la sélection des documents, la longueur du contenu textuel, la répartition des événements multilingues ou les erreurs et défauts d'annotation. Le tableau 2.10 présente quelques statistiques quant aux langues, nombres de documents et d'événements des données sélectionnées.

Jeu de données	Type	Langue	Docs.	Longueur de titre		Longueur de texte		Événements	Taille d'événement	
				$\bar{x}$	σ	$\bar{x}$	σ		$\bar{x}$	σ
Event Registry	Entraînement	Indif.	20803	57	20	2408	2020	1108	19	29
		Allemand	4043	55	20	1999	1541	377	11	7
		Anglais	12233	56	19	2638	2287	593	21	32
		Espagnol	4527	62	20	2151	1467	416	11	9
	Test	Indif.	13004	58	20	3120	3097	381	34	99
		Allemand	2101	56	21	3222	3545	118	18	45
		Anglais	8726	58	19	3265	3146	222	39	89
		Espagnol	2177	62	19	2440	2225	149	15	21
CoAID	Entraînement	Anglais	72045	nd	nd	155	84	375	192	146
	Test	Anglais	32100			194	79	125	257	163
FibVid	Entraînement	Anglais	72045			155	84	375	192	146
	Test	Anglais	32100			194	79	125	257	163

TABLEAU 2.10 – Statistiques synthétiques des documents des jeux de données sélectionnés.

2.4 Conclusion et positionnement

Le suivi de mentions d'événements dans la presse est porté par une longue histoire qui débute il y a plus de vingt ans, lorsque de premiers projets sont développés. Nous avons parcouru dans ce chapitre les questionnements et enjeux liés à cette question. Le concept même d'événement porte une forte ambiguïté. Nous avons proposé une définition à la fois en langage naturel et plus formelle de cette notion. Nous l'utilisons dans le reste de cet ouvrage. Bien que des différences existent entre les projets de recherche ou même les domaines (ouvert, fermé, ontologique), des points communs demeurent entre toutes ces représentations. Nous en donnons une définition à la sous-section 2.1.4 : un événement est « une action ancrée dans le temps, dans l'espace et qui fait intervenir des participants ». Des événements s'incluent dans d'autres lorsque cela est nécessaire, et ils sont toujours liés à leur contexte d'action. Celui-ci est représenté par les événements qui ont eu lieu avant et qui l'ont engendré, et ceux qui vont découler de cette action. Nous nous inscrivons donc dans un premier temps dans la lignée des suivis d'événements en domaine ouvert.

Ensuite, nous avons présenté les algorithmes capables d'identifier, de représenter et de définir des histoires d'événements à partir d'articles. La représentation de chaque document, les informations à exploiter de ceux-ci sont les premiers enjeux de recherche dans ce domaine. Nous avons présenté les stratégies de plongement numérique, basées sur des vectorisations de documents. Les pondérations par calcul de *TF-IDF* ainsi que les vectorisations par apprentissage profond remplissent ce rôle. Les vecteurs obtenus décrivent le document et son contenu. Les dates des publications, les entités nommées ou les langues contextualisent les documents, et donc les événements. Cette approche de suivi considère que chaque article décrit un événement. Par conséquent, l'encodage du

document de presse décrit l'événement. La représentation des histoires dans des graphes utilise des approches relativement similaires, en comparant les textes des articles et en établissant des connexions entre les nœuds (donc entre les articles) basées sur des propriétés supplémentaires (date, origine, etc.). Ce travail de recherche s'inscrit d'abord dans cette continuité. Notre problématique présentée en Introduction (page 16) est de détecter et suivre des mentions d'événements dans la presse historique. Nous avons montré dans cet état de l'art que c'est une tâche aujourd'hui appliquée assez largement à la presse récente. Le présent document se consacre à l'analyse des algorithmes et méthodes de suivi en présence de documents historiques. Comme mentionné à la section 2.2, nous utiliserons pour cela deux méthodes différentes pour évaluer le suivi d'événements : une à base d'algorithmes issus du domaine *TDT* et une fonctionnant tel un moteur de recherche. Nous l'avons évoqué, les vectorisations des documents influeront nécessairement sur les résultats de certains de ces algorithmes. Nous étudierons ces processus au chapitre 4 en nous focalisant d'abord sur l'influence des représentations vectorielles. Dans le cas du moteur de recherche, comme nous l'avons indiqué dans ce chapitre, les vectorisations semblent superflues. Ce procédé décrit au chapitre 5 est moins consommateur de ressources et d'énergie. En effet, il n'impose pas d'apprentissage machine ou de calculs préparatoires comme des vectorisations. Nous estimons qu'il s'agit d'une première piste pour mettre en pratique la sobriété évoquée en introduction de ce document.

Nous devons évaluer nos méthodes et ce n'est possible qu'avec des données de presse annotées d'événements. Nous avons, à la section 2.3, analysé trois jeux de données. La problématique principale que nous soulevons est celle du choix des données. Par rapport à l'objectif de ce projet de recherche, peu de corpus de documents existent. Nous en avons sélectionné un, *Event Registry*, utilisé par la communauté scientifique et deux autres, *CoAID* et *FibVid*, que nous exploitons dans un contexte différent de ce pour quoi ils sont prévus. Par analogie, ces deux derniers étant comparables à des corpus de brèves de presse, nous proposons de les utiliser comme tels. Nous avons également analysé les répartitions des événements, le multilinguisme d'*Event Registry* ainsi que la taille et la variété des textes. Par cette analyse, nous avons identifié quelques biais pouvant affecter des résultats d'algorithmes opérant des suivis d'événements.

Les événements définis, les algorithmes de suivi décrits et les jeux de données d'évaluation trouvés, nous devons nous intéresser aux spécificités des documents historiques. Comme nous l'avons déjà évoqué, peu de travaux se spécialisent sur cette problématique. La faible disponibilité de ces ressources historiques en est un facteur d'explication. Dans le chapitre suivant, nous proposons d'étudier en détail les problématiques liées aux documents de presse numérisés. Nous proposons une méthodologie pour évaluer les algorithmes en présence de documents dégradés, comme le seraient de vrais articles historiques numérisés.

Chapitre 3

Manipuler la presse historique

Sommaire

Au chapitre 2, nous avons principalement évoqué les travaux relatifs aux événements rapportés dans la presse récente, les travaux scientifiques se focalisant davantage sur ce domaine. Les données sont facilement disponibles au travers d'APIs ou par extraction de contenu des pages Web. Peu de travaux traitent de ces mêmes problématiques appliquées aux documents historiques. Nous l'avons vu en introduction à la page 16, les données historiques de presse font depuis quelques décennies l'objet de campagnes de numérisation au sein de nombreuses bibliothèques de conservation dans le monde. La mission principale de ces bibliothèques est de sauvegarder et conserver des documents fragiles et parfois rares. Les documents historiques sont rares du fait de tirages anciens, et fragiles entre autres à cause de l'altération du papier, dans le cas de documents de ce type. Les données sont généralement accessibles sous la forme d'images accompagnées de nombreuses métadonnées et du texte reconnu dans l'image par l'ordinateur. Ces numérisations sont dédiées avant tout au grand public pour faciliter la diffusion des contenus et des connaissances. Les publics d'historiens et d'historiennes ou d'humanités numériques les utilisent également et de nombreux projets de recherche sont développés dans ce but, nous les verrons ci-après.

Nous évoquerons d'abord dans ce chapitre les différentes spécificités rencontrées lors de la manipulation de documents historiques numérisés (section 3.1). Certains défauts sont constants et dépendent des conditions de conservation des journaux ou d'acquisition numérique. Parmi l'ensemble des projets de la dernière décennie dédiés à l'analyse de la presse historique, nous présenterons plus en détail l'un d'entre eux en section 3.2. Il a notamment produit un corpus de données que nous envisageons d'utiliser pour suivre des mentions d'événements dans des documents historiques. Nous présentons une analyse succincte de ces données dans la même section. Nous verrons pourquoi il n'est pas, en l'état, possible de l'exploiter pleinement par manque d'annotations adaptées. Pour résoudre cette difficulté, un processus d'annotation d'événements est proposé en section 3.3, synthétisant les différents travaux réalisés par le passé. Les données historiques de presse annotées pour le suivi d'événements ne sont pas disponibles. Nous proposons en section 3.4 une stratégie de création artificielle de documents historiques à partir de presse récente. Enfin, nous faisons face aux questions de reproductibilité de recherche liées aux données. Certains algorithmes et outils utilisés pour l'extraction de caractéristiques du texte (entités, événements, etc.) se basent sur des modèles d'apprentissage ou des données qui ne sont pas toujours conservées correctement. Nous introduirons une solution d'archivage, à la section 3.5, de documents historiques numérisés préanalysés et enrichis. L'objectif de ce travail est de rationaliser et simplifier les efforts de recherche en limitant les tâches redondantes de préparation des données.

3.1 Spécificités des documents de presse historique

La consommation de la presse en 2022 n'a aucune commune mesure avec celle des décennies passées. L'introduction des technologies de l'information et de la communication dans la diffusion des informations fait diminuer régulièrement la part de journaux consommés au format papier. À la fin de l'année 2021, 68 % [ACP21 ; New+21] des

consommations de presse se font en ligne, au détriment du support papier. Ce chiffre a peu évolué depuis les années 2012 et 2015 [New12 ; NAK15].

Les documents de presse numérique font l'objet de nombreux travaux, comme nous l'avons illustré au chapitre 2. Afin de comprendre les enjeux liés à l'analyse des documents historiques numérisés, il convient d'en connaître les particularités, à la fois relatives au support et au contenu.

3.1.1 Les dégâts propres aux documents historiques numérisés

Les documents historiques de presse se placent en opposition aux documents de presse récents et généralement nativement numériques, c'est-à-dire disponibles sous forme directement interprétable par un programme informatique. L'ensemble des documents rédigés à l'ère prénumérique fait, dans un premier temps, partie de cet ensemble. Cette période s'arrête au début des années 2000, à partir de laquelle les organismes de presse commencent à diffuser massivement leurs contenus sur des sites Internet. Cette démarche concorde avec l'émergence du Web, né une dizaine d'années plus tôt. Rentre de fait en ligne de compte l'ensemble des documents de presse publiés sur un support physique tels un journal ou un magazine. La page physique sur laquelle est imprimé du texte et des images constitue la source de l'information traitée dans cette thèse.

Les documents sont numérisés par un processus qui opère une transition du document physique préhensible vers un document numérique. Cette transition est rendue possible grâce à des outils qui opèrent à différents niveaux :

— Le **scanneur** représente un document sous la forme d'une image, donc d'une composition de pixels à partir d'une page de journal par exemple. Ce processus est sensible à la qualité du document physique fourni.
— L'outil de **segmentation de texte** identifie les composants originels du document reconnaissables dans l'image, les paragraphes, les titres, les figures, etc.
— Le logiciel de **reconnaissance optique de caractères** (*OCR*) analyse les pixels des portions de l'image identifiés par la segmentation pour en extraire le texte. Il fournit une interprétation textuelle des pixels de l'image qu'il a la tâche de reconnaître.

À chaque étape du processus de traitement, des erreurs dommageables aux suivantes surviennent nécessairement. Chacune est sensible à différents niveaux d'erreurs.

Le scanneur est sensible au matériau d'origine : l'altération du papier, ses tâches, pliures, trous, l'effacement de l'encre du texte, etc. Sur la figure 3.1a, on constate des dégradations sur les marges du journal et du papier manque. Des tâches se trouvent au milieu du document et la colonne de droite de l'image est plus sombre qu'ailleurs. La figure 3.1b, quant à elle, met en avant les risques de pliures et d'altérations non pas du document, mais liés à l'utilisation de l'appareil de numérisation : le papier est gondolé. Le texte est discontinu et semble écrit sous forme de vaguelettes. Pour certains mots, la fonte est grasse, rendant difficile la lecture, même pour un humain.

Les défauts propres aux documents historiques numérisés sont connus et ont fait l'objet de nombreuses études [Ale+12 ; Chi+17 ; Jou+17 ; Mut+18]. Nous distinguons

Cinq Centimes

ERNEST VAUGHAN

L'AURORE

Littéraire. Artistique. Sociale

ERNEST VAUGHAN

J'Accuse...!

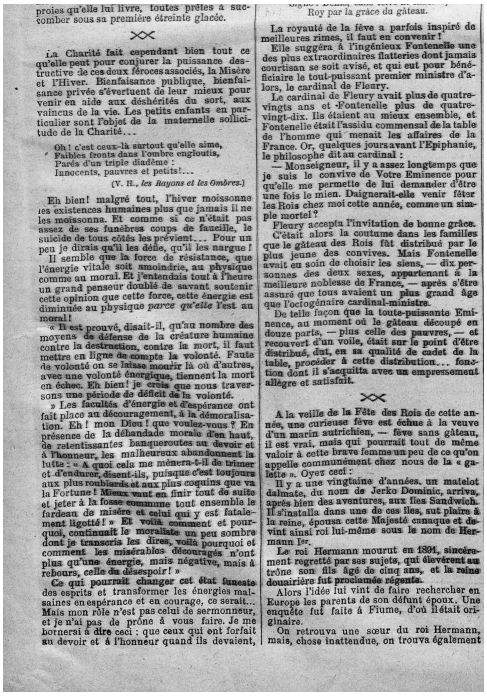
LETTRE AU PRÉSIDENT DE LA RÉPUBLIQUE

Par ÉMILE ZOLA

LETTRE
A M. FÉLIX FAURE
Président de la République

Source gallica.bnf.fr / Bibliothèque nationale de France

(a) Numérisation d'un exemplaire célèbre du journal L'Aurore, contenant l'article d'Émile Zola relatif à l'affaire Dreyfus.
Source : `gallica.bnf.fr`/BnF



La Charité fait cependant bien tout ce qu'elle peut pour conjurer la puissance destructive de ces deux féroces associés, la Misère et l'Hiver. Bienfaisance publique, bienfaisance privée s'évertuent de leur mieux pour venir en aide aux déshérités du sort, aux vaincus de la vie. Les petits enfants en particulier sont l'objet de la maternelle sollicitude de la Charité...

Oh ! c'est ceux-là surtout qu'elle aime,
Faibles fronts dans l'ombre engloutis,
Purés d'un triple diadème :
Innocents, pauvres et petits !...
(V. H., *les Rayons et les Ombres.*)

Eh bien ! malgré tout, l'hiver moissonne les existences humaines plus que jamais il ne les moissonna. Et comme si ce n'était pas assez de ses funèbres coups de faucille, le suicide de tous côtés les prévient... Pour un peu je dirais qu'il les dicte, qu'il les nargue !

Il semble que la force de résistance, que l'énergie vitale soit amoindrie, au physique comme au moral. Et j'entendais tout à l'heure un grand penseur doublé de savant soutenir cette opinion que cette force, cette énergie est diminuée au physique *parce qu'elle l'est au moral !*

« Il est prouvé, disait-il, qu'au nombre des moyens de défense de la créature humaine contre la destruction, contre la mort, il faut mettre en ligne de compte la volonté. Faute de volonté on se laisse mourir là où d'autres, avec une volonté énergique, tiennent la mort en échec. Eh bien ! je crois que nous traversons une période de déficit de la volonté.

» Les facultés d'énergie et d'espérance ont fait place au découragement, à la démoralisation. Eh ! mon Dieu ! que voulez-vous ? En présence de la débandade morale d'en haut, de ratentissantes banqueroutes au devoir et à l'honneur, les malheureux abandonnent la lutte : « A quoi cela me mènera-t-il de trimer et d'endurer, disent-ils, puisque c'est toujours aux plus roublards et aux plus coquins que va la Fortune ! Mieux vaut en finir tout de suite et jeter à la fosse commune tout ensemble le fardeau de misère et celui qui y est fatalement ligotté ! » Et voilà comment et pourquoi, continuait le moraliste un peu sombre dont je transcris les dires, voilà pourquoi et comment les misérables découragés n'ont plus qu'une énergie, mais négative, mais à rebours, celle du désespoir ! »

Ce qui pourrait changer cet état funeste des esprits et transformer les énergies malsaines en espérance et en courage, ce serait... Mais mon rôle n'est pas celui de sermonneur, et je n'ai pas de prône à vous faire. Je me bornerai à dire ceci : que ceux qui ont forfait au devoir et à l'honneur quand ils devaient,

Roy par la grâce du gâteau.

La royauté de la fève a parfois inspiré de meilleures rimes, il faut en convenir !

Elle suggéra à l'ingénieux Fontenelle une des plus extraordinaires flatteries dont jamais courtisan se soit avisé, et qui eut pour bénéficiaire le tout-puissant premier ministre d'alors, le cardinal de Fleury.

Le cardinal de Fleury avait plus de quatre-vingts ans et Fontenelle plus de quatre-vingt-dix. Ils étaient au mieux ensemble, et Fontenelle était l'assidu commensal de la table de l'homme qui menait les affaires de la France. Or, quelques jours avant l'Épiphanie, le philosophe dit au cardinal :

— Monseigneur, il y a assez longtemps que je suis le convive de Votre Eminence pour qu'elle me permette de lui demander d'être une fois le mien. Daignerait-elle venir fêter les Rois chez moi cette année, comme un simple mortel ?

Fleury accepta l'invitation de bonne grâce.

C'était alors la coutume dans les familles que le gâteau des Rois fût distribué par le plus jeune des convives. Mais Fontenelle avait eu soin de choisir les siens, — dix personnes des deux sexes, appartenant à la meilleure noblesse de France, — après s'être assuré que tous avaient un plus grand âge que l'octogénaire cardinal-ministre.

De telle façon que la toute-puissante Eminence, au moment où le gâteau découpé en douze parts, — plus celle des pauvres, — se recouvert d'un voile, était sur le point d'être distribué, dut, en sa qualité de cadet de la table, procéder à cette distribution... fonction dont il s'acquitta avec un empressement allégre et satisfait.

A la veille de la Fête des Rois de cette année, une curieuse fève est échue à la veuve d'un marin autrichien, — fève sans gâteau, il est vrai, mais qui pourrait tout de même valoir à cette brave femme un peu de ce qu'on appelle communément chez nous de la « galette ». Oyez ceci :

Il y a une vingtaine d'années, un matelot dalmate, du nom de Jerko Dominic, arriva, après bien des aventures, aux îles Sandwich. Il s'installa dans une de ces îles, sut plaire à la reine, épousa cette Majesté canaque et devint ainsi roi lui-même sous le nom de Hermann Ier.

Le roi Hermann mourut en 1891, sincèrement regretté par ses sujets, qui élevèrent au trône son fils âgé de cinq ans, et la reine douairière fut proclamée régente.

Alors l'idée lui vint de faire rechercher en Europe les parents de son défunt époux. Une enquête fut faite à Fiume, d'où il était originaire.

On retrouva une sœur du roi Hermann, mais, chose inattendue, on trouva également

(b) Numérisation de la seconde page du Petit Journal illustré du dimanche 21 janvier 1893, coin inférieur gauche.
Source : `gallica.bnf.fr`/BnF

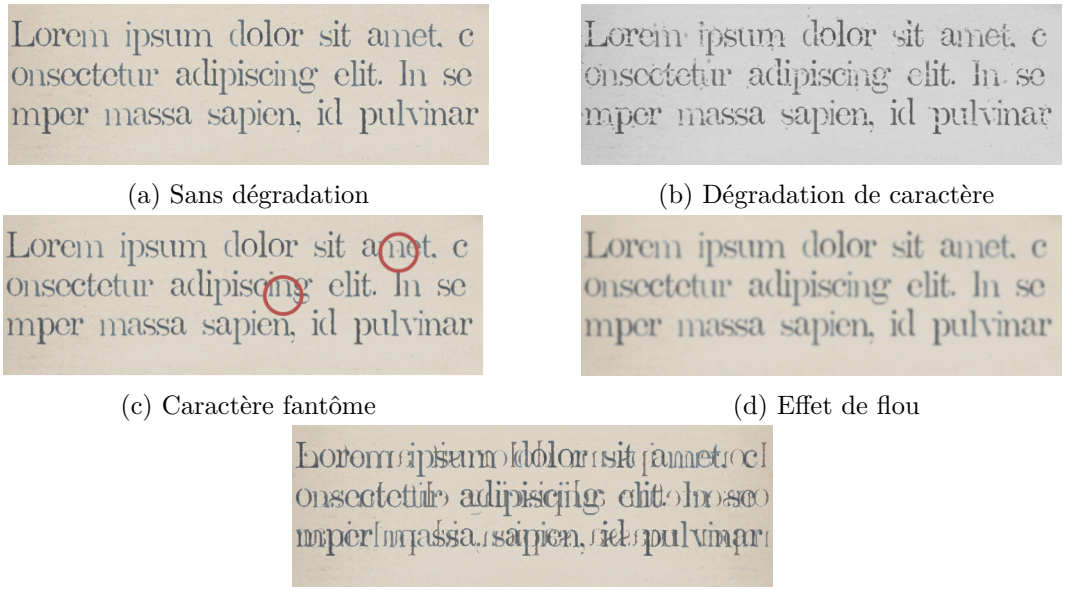
FIGURE 3.1 – Deux exemples de documents de presse historique numérisés.

différentes erreurs majeures qui ont été décrites comme ayant le plus d'impact sur le reste du processus [Lin+19] :

- La **dégradation de caractère** est une atteinte due à l'âge du document, à sa manipulation ou son impression. Elle entraîne la présence de points d'encre aléatoires, de tâches et limite la capacité de l'outil de reconnaissance des caractères ;
- l'introduction de **caractères fantômes**, conséquence de l'érosion de l'encre sur le support papier. Des tâches peuvent apparaître lors de la numérisation.
- l'apparition, par **transparence**, de la page située au verso de celle numérisée, sur les documents dont les deux côtés sont imprimés ;
- l'effet de **flou** lié à un problème de focale du dispositif d'acquisition.

La figure 3.2 présente les différentes dégradations couramment retrouvées sur les images numérisées de journaux anciens.

L'outil de segmentation textuelle traite l'image à la recherche de blocs individuels composant la page de journal. Ces blocs sont nommés segments et forment des unités de contenu, généralement du texte ou des images. Différentes techniques sont utilisées. La plus classique repose sur la détection des divisions entre segments par binarisation de l'image, puis par la recherche d'un seuil de séparation [WWC82]. Des travaux sur des



(a) Sans dégradation

(b) Dégradation de caractère

(c) Caractère fantôme

(d) Effet de flou

(e) Effet de transparence

FIGURE 3.2 – Les dégradations courantes visibles dans des documents historiques numérisés.

documents numérisés montrent l'intérêt des réseaux convolutifs [Mei+17], des chaînes de Markov [NNC19], des systèmes à base de règles [AG22] ou des réseaux de neurones de graphes [MWL22] pour extraire les articles de presse. La segmentation des pages numérisées de journaux est un enjeu de la recherche liée à la numérisation des contenus historiques : l'objectif est de reconstruire des articles à partir des différents segments de texte extraits par les logiciels. La sursegmentation du texte est le phénomène le plus généralement observé dans des textes numérisés segmentés automatiquement. Les articles sont divisés en petites unités et ne sont pas complets. Les articles peuvent être séparés au sein des pages par des images, des séparateurs visuels ou des sauts de colonne, comme cela est visible dans la figure 3.1b. Les logiciels de segmentation de texte sont adaptés à la reconnaissance de paragraphes ou de figures, mais pas d'articles de presse. La reconnaissance d'un article de presse à partir de segments de texte est une étape de post-traitement qui regroupe et réordonne les segments en un ensemble cohérent : celui d'un article complet. Chaque journal dispose de sa propre mise en page. Néanmoins, le format à colonnes multiples est commun aux éditeurs de journaux et de magazines. La consultation des archives numériques de presse témoigne que cette pratique a peu évolué depuis le début de l'imprimerie. Depuis le début des années 2010, des algorithmes [Pal+12 ; Mic+21 ; Zhu+22] sont capables de repérer de la continuité entre segments au sein d'une même colonne et entre différentes colonnes, participant à reconstruire des articles à partir de segments de texte individuels.

Enfin, un logiciel de reconnaissance optique de caractères analyse les segments détectés et extraits de l'image numérisée. Il est sensible à la qualité de l'image. Sa capacité

à reconnaître correctement les formes et donc les graphèmes est diminuée si l'image est celle d'un document dégradé. Il génère une séquence de caractères et conserve la mise en forme du texte : sauts de ligne, espaces surnuméraires, etc. La sensibilité de l'*OCR* entraîne trois types d'erreurs potentielles. Les erreurs d'insertion surviennent lorsque l'outil reconnaît de façon erronée un groupe de pixels comme un caractère alors qu'il n'existe pas. Une tâche peut être à l'origine d'une insertion. Les erreurs de substitution surviennent lorsqu'un caractère est remplacé par un autre dans un terme. Une tâche ou un trou peuvent expliquer une substitution de caractère. Lorsque l'*OCR* échoue à détecter un caractère pourtant présent dans le texte d'origine, il s'agit d'une erreur de suppression. À l'origine définie pour réaliser des calculs de similarités dans des mots binaires par celui qui lui donnera son nom, la distance de Levenshtein [Lev66] est l'une des métriques utilisée pour l'évaluation des outils *OCR* [Ham+19 ; Lin+20]. L'évaluation consiste à calculer la distance entre le texte extrait par l'*OCR* et la vérité terrain, connue [MS02]. Cette métrique est appelée taux d'erreur de caractères « *Character Error Rate* » (*CER*) pour laquelle une variante existe au niveau des mots : « *Word Error Rate* » (*WER*). L'amélioration des sorties d'*OCR* est un champ de recherche actif où la post-correction est une solution fonctionnelle par utilisation de dictionnaires ou d'autres techniques plus avancées [Ale+12 ; Chi+17 ; Ngu+19 ; Ngu+20 ; HHD20].

3.1.2 La différence de style et de contenu

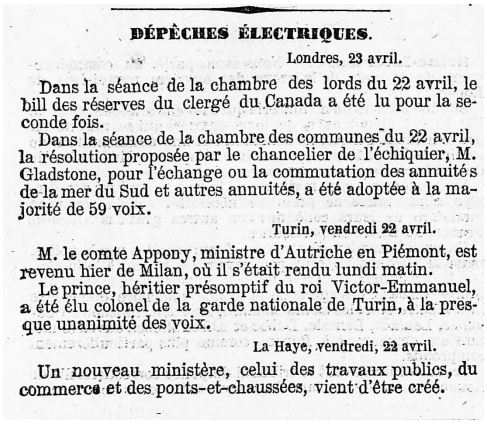
Les différences entre les documents de presse récents et anciens ne résident pas seulement dans leur forme. Le fond, le texte même de l'article est le marqueur d'une époque. N'importe quel lecteur ou n'importe quelle lectrice avisé différencie un texte rédigé dans cette décennie d'un autre écrit il y un siècle. D'abord, le lexique utilisé, à la fois dans la vie courante et dans le style journalistique a connu des évolutions, concomitantes aux changements linguistiques propres à une langue vivante [Bru81 ; Bru95]. Les syntaxes, formulations et expressions soumises à interprétation peuvent également avoir changé subtilement de signification.

Les noms de lieux connaissent aussi des évolutions liées à la géopolitique. En 1900, l'Empire russe et l'Empire ottoman existaient au même titre que l'Empire d'Autriche-Hongrie. Le Royaume-Uni que l'on connaît en 2022 était appelé, jusqu'en 1927, Royaume-Uni de Grande-Bretagne et d'Irlande. Les noms de personnes changent également, comme en témoigne l'évolution du prénom Anne, autrefois prénom masculin en France, il est devenu exclusivement féminin au cours du XX^{e} siècle. Ces changements sont liés à l'évolution des langues et des situations politiques ou culturelles. Des outils développés au début du XXI^{e} siècle entraînés sur des données récentes peuvent éprouver des difficultés à reconnaître correctement des entités jamais rencontrées auparavant.

Les types d'articles, éditoriaux, brèves de presse, reportages ou analyses sont communs à la presse récente et ancienne. L'éditorial le plus célèbre de la fin du XIX^{e} siècle est peut-être la prise de position d'Émile Zola sur l'Affaire Dreyfus. Ce dernier met en accusation nombre de membres du gouvernement et de notables militaires français. Ce document est présenté en figure 3.1a.

3.1.3 La brièveté du texte et la dépêche télégraphique

C'est au milieu du XIX^e^ siècle que la diffusion de l'information connaît un essor lié à l'apparition du télégraphe. À cette époque, des lignes de communication commencent à relier des zones géographiques lointaines [Bol19]. Format journalistique à part entière, ces messages courts relayés par différentes agences de presse, Havas à Paris, Reuters à Londres ou encore Wolff à Berlin permettent la diffusion des informations à travers l'Europe. Lisa Bolz [Bol19] est, à notre connaissance, à l'origine de la seule étude portant sur les dépêches télégraphiques. D'après ses travaux, les télégrammes, en tant que vecteurs de l'information, permettent un renouveau des pratiques journalistiques de l'époque, tant sur le contenu diffusé que sur la forme des faits rapportés. Deux dépêches télégraphiques sont présentées dans la figure 3.3 ci-après.



DÉPÊCHES ÉLECTRIQUES.

Londres, 23 avril.

Dans la séance de la chambre des lords du 22 avril, le bill des réserves du clergé du Canada a été lu pour la seconde fois.

Dans la séance de la chambre des communes du 22 avril, la résolution proposée par le chancelier de l'échiquier, M. Gladstone, pour l'échange ou la commutation des annuités de la mer du Sud et autres annuités, a été adoptée à la majorité de 59 voix.

Turin, vendredi 22 avril.

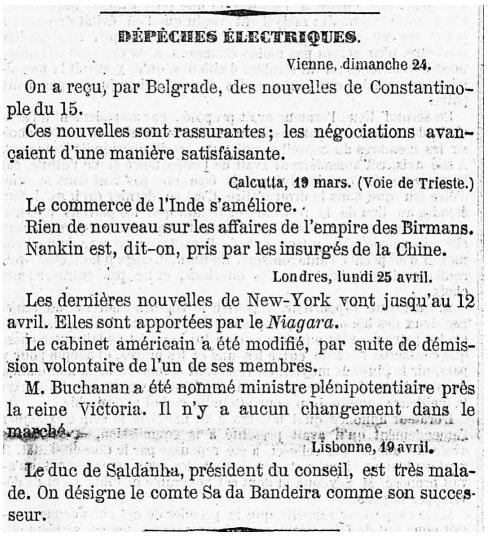
M. le comte Appony, ministre d'Autriche en Piémont, est revenu hier de Milan, où il s'était rendu lundi matin.

Le prince, héritier présomptif du roi Victor-Emmanuel, a été élu colonel de la garde nationale de Turin, à la presque unanimité des voix.

La Haye, vendredi, 22 avril.

Un nouveau ministère, celui des travaux publics, du commerce et des ponts-et-chaussées, vient d'être créé.

(a) Dépêches publiées dans La Presse, le 24 avril.
Source : gallica.bnf.fr/BnF



DÉPÊCHES ÉLECTRIQUES.

Vienne, dimanche 24.

On a reçu, par Belgrade, des nouvelles de Constantinople du 15.

Ces nouvelles sont rassurantes; les négociations avançaient d'une manière satisfaisante.

Calcutta, 19 mars. (Voie de Trieste.)

Le commerce de l'Inde s'améliore.

Rien de nouveau sur les affaires de l'empire des Birmans.

Nankin est, dit-on, pris par les insurgés de la Chine.

Londres, lundi 25 avril.

Les dernières nouvelles de New-York vont jusqu'au 12 avril. Elles sont apportées par le *Niagara*.

Le cabinet américain a été modifié, par suite de démission volontaire de l'un de ses membres.

M. Buchanan a été nommé ministre plénipotentiaire près la reine Victoria. Il n'y a aucun changement dans le marché.

Lisbonne, 19 avril.

Le duc de Saldanha, président du conseil, est très malade. On désigne le comte Sa da Bandeira comme son successeur.

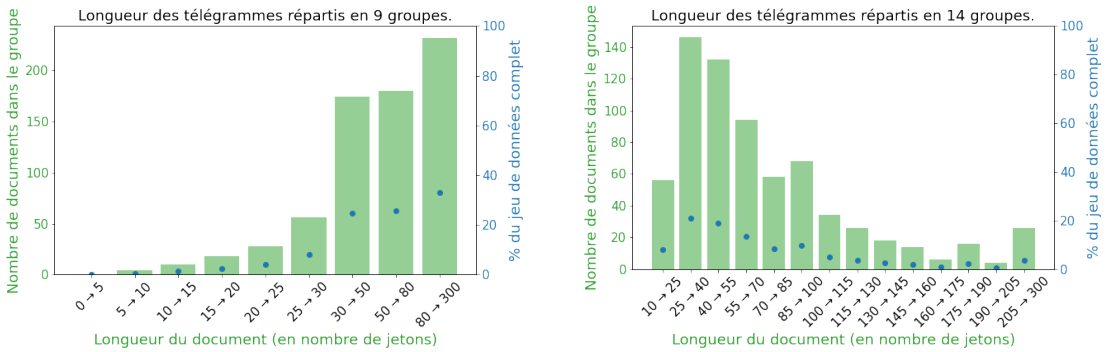
(b) Dépêches publiées dans La Presse, le 26 avril.
Source : gallica.bnf.fr/BnF

FIGURE 3.3 – Deux dépêches télégraphiques publiées dans le journal La Presse, en avril 1853

Les dépêches télégraphiques se caractérisent, en tant que documents historiques textuels, par leur faible longueur et la concision du propos. Il s'agit à l'époque de relever le défi de la transmission des faits observés ou rapportés en un minimum de caractères possibles. La transmission de l'information est à l'origine de coûts importants qui influent sur la quantité de données envoyée par les correspondants, comme le décrit Lisa Bolz dans sa thèse [Bol19, p. 181] :

> *Le prix élevé des télégrammes influe sur l'écriture des dépêches. La transmission de dépêches sur de longues distances reste onéreuse. Elles sont par conséquent courtes et condensées au maximum [...].*

En annexe de ce travail de thèse est publié un corpus de retranscriptions manuelles de dépêches télégraphiques. Il contient 30 télégrammes publiés en 1850, 133 en 1860 et 195 en 1870. La longueur des documents est détaillée en figure 3.4 et permet une comparaison avec l'analyse présentée en section 2.3.2. On note que le profil des longueurs des dépêches télégraphiques est similaire à celui des jeux de données *CoAID* et *FibVid* qui contiennent des brèves journalistiques. La longueur moyenne des documents est de 81 termes (écart-type de 83 termes, médiane à 58,5) avec respectivement un premier quartile à 38 et un troisième à 94 termes. En proportion, moins de 5 % des documents ont une longueur supérieure à 200 termes. Ces données fournissent une indication quant à la longueur des dépêches, sans toutefois permettre une analyse approfondie par manque de données.



(a) Longueur des télégrammes, avec les mêmes intervalles que celles de la section 2.3.2.

(b) Longueur des télégrammes, détaillée.

FIGURE 3.4 – Longueur des télégrammes, en nombre de termes, fournis par Lisa Bolz [Bol19].

3.1.4 La temporalité dans la diffusion de l'information

L'étude des dépêches télégraphiques présentée dans la section précédente soulève la problématique des délais de transmission des actualités. Depuis l'apparition du télégramme, la vitesse et le volume des informations échangées dans le monde sont en augmentation constante. Du milieu du XIX^{e} siècle au milieu du XX^{e} siècle, le développement télégraphique a connecté le monde et accéléré les échanges [21b]. Néanmoins, l'accélération ne signifie pas l'instantanéité des communications. L'analyse de certaines dépêches télégraphiques montre qu'un délai plus ou moins long s'écoule entre l'émission d'une information et sa publication dans un journal. Chacune des deux dépêches télégraphiques présentées en figure 3.3 est caractérisée par son lieu et sa date d'émission. Nous remarquons qu'il peut s'écouler plusieurs jours voire semaines après un événement pour que l'information soit publiée, comme dans la figure 3.3b. Comme suggéré précédemment [Bol19], la distance et le nombre de relais peuvent être des facteurs, toutes choses étant égales par ailleurs, d'explication des délais de transmission importants.

Pour illustrer cette hypothèse, les nouvelles rapportées le 24 avril, c'est-à-dire les

dépêches de la figure 3.3a datent de la veille ou du 22 avril. Les informations proviennent de correspondants et de correspondantes de villes européennes, Londres (Angleterre), La Haye (Pays-Bas) et Turin (Italie). Concernant les dépêches du 26 avril, présentées dans la figure 3.3b, la plus ancienne a été émise plus d'un mois plus tôt, depuis Calcutta (Inde) et a transité par les relais de Trieste, aujourd'hui ville italienne.

La plus courte durée qui peut s'écouler entre la survenue d'un événement et sa mention par les organes de presse est au minimum d'une journée. Ce délai est fonction de la périodicité du journal, de s'il s'agit d'un quotidien du soir ou du matin, de la réactivité de sa rédaction, etc.

3.2 Exploration et analyse de la presse historique

Les documents de presse font l'objet de campagnes de préservation et d'archivage dans le monde entier [New18 ; ZK20]. Ils constituent une part intégrante de l'héritage culturel de nos civilisations, de leurs histoires et des tourments qu'elles ont connus. Les journaux rapportent les événements politiques, sociaux, culturels, parmi bien d'autres, dans tous les pays et toutes les langues du monde [New18]. Grâce aux technologies de l'information, les grandes bibliothèques de conservation numérisent et archivent leurs fonds documentaires afin d'en préserver le contenu et de simplifier leur diffusion. En France, et pour ne citer qu'elle, la Bibliothèque Nationale de France conserve et publie une archive numérisée contenant plusieurs millions de documents de différentes natures : livres, revues, manuscrits, presse...

L'accès à ces ressources est critique pour mener des recherches historiques dans de multiples disciplines telles la sociologie ou l'histoire. Cependant, la numérisation des documents ne suffit pas, ils doivent être analysés. Différents projets avec cet objectif ont vu le jour depuis le début des campagnes de numérisation, en 1990. Ces projets sont toujours interdisciplinaires et impliquent des bibliothèques, des historiens, des chercheurs et chercheuses en humanités numériques, des scientifiques spécialistes du numérique, de l'analyse de données ou des linguistes. Les archives historiographiques sont le matériau de base de nombreuses recherches historiques. D'un autre côté, les technologies de l'information et de la communication apportent des capacités d'analyse massives.

Des projets comme *Europeana NewsPapers*[1] ou *Chronicling America*[2] ont pour ambition de sauvegarder et d'agréger les journaux historiques d'Europe et des États-Unis. Tous les deux soutenus par des organisations publiques, la Commission européenne pour le premier et la *Library of Congress* pour le second, ils publient à eux deux des millions de pages de documents numérisés. *Europeana NewsPapers* met à disposition dix-huit millions de pages de journaux issus de bibliothèques nationales européennes dont deux millions sont analysées et prétraitées. La bibliothèque du Congrès américain quant à elle diffuse cent cinquante mille numéros, de 1690 jusqu'à nos jours. Publiés aux États-Unis, ces journaux sont le reflet des évolutions des langues et des attitudes de l'opinion face à l'information. Dans les deux cas, leurs objectifs sont d'améliorer la navigation dans les

1. `http://www.europeana-newspapers.eu/` (archivé sur `https://web.archive.org`)
2. `https://chroniclingamerica.loc.gov/` (archivé sur `https://web.archive.org`)

corpus numériques de presse historique et de fournir des outils dédiés à l'analyse et à la fouille de textes.

Sur le plan de l'analyse de documents historiques, les projets *Impresso*[3], *Living With Machines*[4], *NewsEye*[5], *Numapresse*[6] ou *Oceanic Exchanges*[7] forment l'état de l'art. *Impresso* est un projet multidisciplinaire suisse et luxembourgeois dédié à la surveillance et à l'analyse des médias historiques. Il concentre six cent mille titres de presse tirées des bibliothèques nationales du Luxembourg et de Suisse. Les membres du projet explorent les archives historiques par des techniques de modélisation de sujet et le suivi de la diffusion des informations par analyse de la réutilisation de texte. Nous avons évoqué cette stratégie à l'état de l'art (section 2.2, page 40). Ils étudient les tensions sociales liées à l'idée d'intégration européenne entre la fin du XIX^e siècle et 1950 en Suisse. Les questions d'enrichissement des données, de visualisation et d'analyse sont un commun à tous ces projets. La réutilisation de texte est utilisée dans le cadre des projets *Oceanic Exchanges* et *Numapresse* [Oce17]. Le principe repose sur l'observation que nous avons faite précédemment, en sous-section 2.3.2 : certains textes sont dupliqués. Oiva et coll. [Oiv+19] ont analysé la façon dont l'assassinat de Nikolay Bobrikov est traité en identifiant les portions de texte réutilisées entre différents articles. Cette technique n'est pas explorée dans le cadre de cette thèse, bien qu'elle puisse mener à des résultats prometteurs pour tracer des événements [SCD13 ; Sal+21]. Au Royaume-Uni, le projet *Living With Machines* s'intéresse à la manière dont la presse rend compte de la révolution industrielle. Les journaux, publiés entre 1780 et 1914 sont issus de la *British Library*. En plus des jeux de données et outils de visualisation de données, ils publient des modèles de langue (comme ceux que nous avons évoqués en section 2.2) adaptés au traitement de textes historiques. Le projet *NewsEye* s'intéresse quant à plusieurs cas dont l'analyse du genre via l'expression des droits des femmes dans la presse. Ces études de cas permettent de valider les outils développés. Ce projet implique des bibliothèques, des centres de recherche en humanités numériques et en informatique. Ce travail interdisciplinaire mêle des enjeux scientifiques variés : amélioration des outils d'analyse des documents historiques, réduction des erreurs dues aux défauts mentionnés dans la section 3.1, création d'outils d'analyse multilingue de corpus historiques ou encore création d'outils de recherche pour les humanités numériques. Pour permettre ces travaux, la Bibliothèque Nationale de France, la Bibliothèque Nationale d'Autriche (*Österreichische Nationalbibliothek*) et la Bibiothèque Nationale de Finlande (*Kansalliskirjasto*) partagent une partie de leur fonds numérisé de documents de presse. Nous appellerons dans la suite de ce document « corpus *NewsEye* » l'ensemble des quinze millions de documents que forme la mise en commun des collections de presses anciennes de ces trois bibliothèques nationales européennes. Les traitements sur ces documents sont réalisés avec l'état de l'art actuel de la segmentation de documents, d'*OCR* ou d'extraction d'entités. Il est également disponible publiquement via la plate-forme du projet [JD20]. Pour ces raisons, ce corpus peut

3. `https://impresso-project.ch/` (archivé sur `https://web.archive.org`)
4. `https://livingwithmachines.ac.uk` (archivé sur `https://web.archive.org`)
5. `https://newseye.eu` (archivé sur `https://web.archive.org`)
6. `https://www.numapresse.org/` (archivé sur `https://web.archive.org`)
7. `https://oceanicexchanges.org/` (archivé sur `https://web.archive.org`)

répondre à notre problématique. Nous l'analysons dans la section suivante pour savoir s'il contient des événements et des annotations d'événements que nous pouvons utiliser.

3.2.1 Les corpus de presse du projet *NewsEye*

Le corpus *NewsEye* est un recueil de pages de journaux numérisées du même type que celui présenté en figure 3.1a. Il est issu de l'agrégation de fonds de publications en français, allemand, finnois, suédois et anglais rédigés durant un siècle, de 1850 à 1950. Pour chaque langue, ce sont plusieurs titres qui constituent le corpus, certains sont apparus durant la période, d'autres ont cessé d'exister. L'ensemble des périodiques référencés dans *NewsEye*, de leur première publication jusqu'à leur disparition est schématisé sur la frise chronologique en figure 3.5. Les dates de première publication et de disparition des journaux sont indicatives : rien n'indique que tous les exemplaires sont disponibles dans le corpus *NewsEye*. Ce sont six périodiques en français, quatre en allemand, huit en finnois et trois en suédois qui sont conservés. La répartition de ces publications est inégale dans le temps, bien qu'il existe systématiquement au moins un périodique pour chaque année sur le siècle étudié. Chaque jour de publication, différents journaux, rédigés chacun dans leur langue, rapportent les mêmes événements. Cette spécificité est un atout qui fait du corpus *NewsEye* un jeu de données adapté à l'étude des événements et des documents dans un contexte multilingue.

La recherche en informatique structurée autour des projets comme *NewsEye* ou *Impresso* a permis des progrès significatifs en reconnaissance optique de caractère [AC18 ; Mut+18 ; Ham+19 ; Lin+19 ; Ngu+19 ; HHD20], en extraction et liaison d'entités nommées avec des bases de connaissances [LMD20 ; Lin+20 ; Bor+20a ; Bor+20c ; Ehr+20 ; Ehr+22 ; Bor+22a] ou en segmentation d'articles sur des supports de presse numérisés [Mic+21]. En parallèle, des progrès sont réalisés sur la détection et l'extraction d'événements dans les articles de presse [Bor+20b ; BMD21 ; Mut+21a ; Mut+21b]. Néanmoins, les représentations d'événements adoptées privilégient une définition sémantiquement très structurée d'un événement [Bor+22b], avec des concepts se rapprochant du formalisme adopté dans *ACE* ou *ERE*. Conformément aux conclusions avancées dans l'état de l'art de la section 2.3, ces représentations structurées et formelles des événements s'adaptent mal à la problématique étudiée ici. La définition d'événement et sa représentation sont plutôt issues des mouvements à « domaine ouvert » dont sont issus les projets *Europe Media Monitor* [Pou+04] ou *Event Registry* [Rup+16].

Ces données sont toutes librement accessibles et manipulables à travers la plate-forme en ligne disponible à l'adresse `https://platform.newseye.eu` [JD20]. Le document présenté en figure 3.6 est un segment de texte de ce corpus, pour lequel une transcription réalisée par *OCR* est fournie. On peut y voir les entités nommées identifiées et parfois liées à des bases de connaissances telles que Wikipédia.

3.2.2 Analyse exploratoire de données de presse historique

NewsEye un bon candidat pour déterminer comment détecter et suivre les mentions d'événements dans des documents historiques. Il contient des journaux historiques pu-

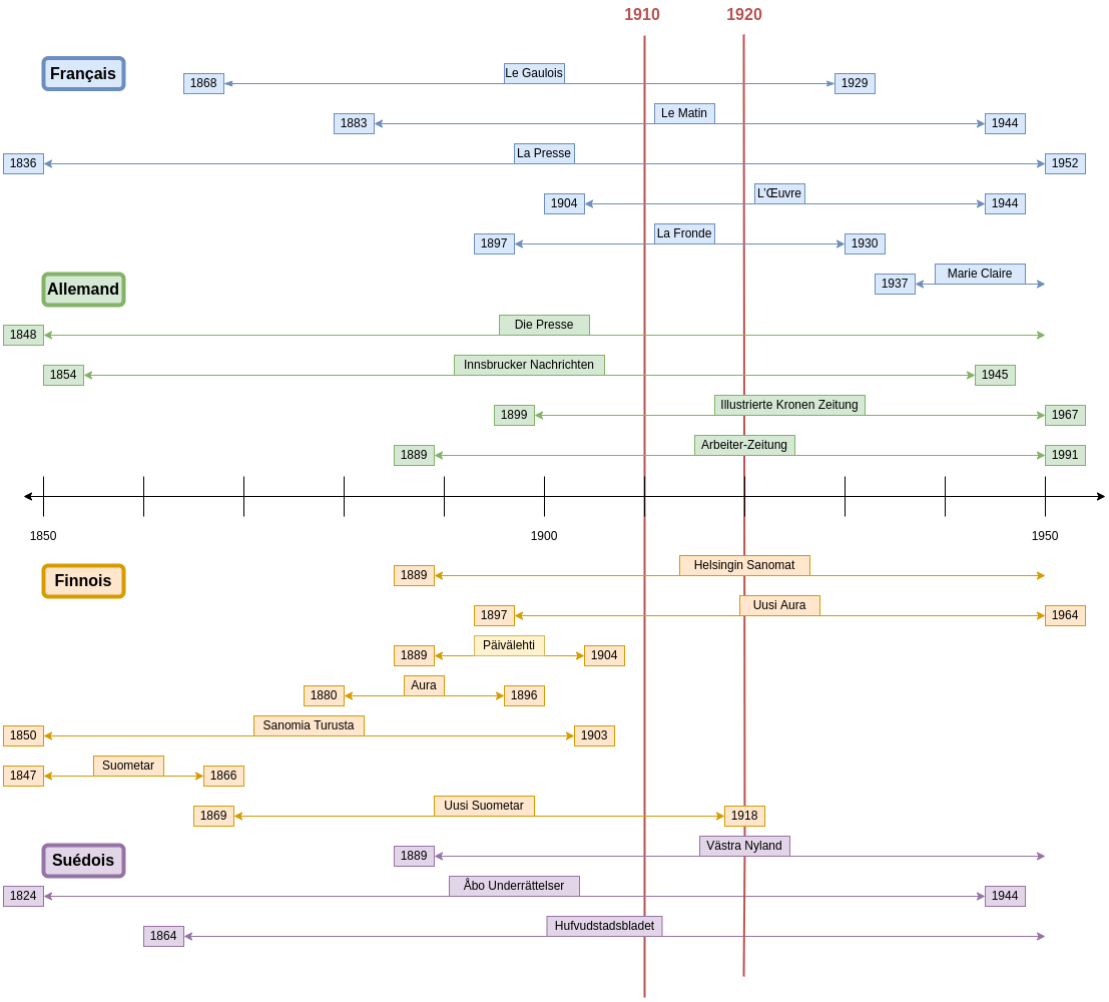
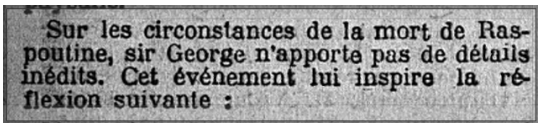


FIGURE 3.5 – Noms et périodes de publication des périodiques du corpus *NewsEye*.

bliés sur une centaine d'années dans plusieurs langues. Nous produisons ici une analyse exploratoire visant à comprendre comment sont organisées ces données et ce qu'elles contiennent. Le code produisant cette analyse est librement disponible sur Internet [Ber22b]. Compte tenu du volume des données (environ 1 To), l'analyse se focalise sur une période qui débute en 1910 et s'achève en 1920 (pour environ 50 Go de documents). C'est une période légèrement plus longue qui est choisie, au sein du projet *NewsEye*, pour étudier les événements dédiés au droit de vote des femmes et à la journée internationale des droits des femmes [Bor+22b].

D'après la figure 3.5, c'est sur cette période qu'est publié en simultané le maximum de journaux différents et dans toutes les langues. Sur cette décennie, il y a au total plus de 24 millions de segments de texte reconnus (dont près de 10,3 en allemand, 7 en finnois, 3,7 en suédois et 3,2 en français). Dans cette masse de documents, chaque segment


Sur les circonstances de la mort de Raspoutine, sir George n'apporte pas de détails inédits. Cet événement lui inspire la réflexion suivante :

(a) Capture d'un extrait de texte.

Sur les circonstances de la mort de Raspoutine, sir George n'apporte pas de détails inédits. Cet événement lui inspire la réflexion suivante :

(b) Retranscription d'un extrait de texte.

FIGURE 3.6 – Exemple de document sursegmenté contenu dans le corpus *NewsEye*.

de texte ne constitue pas nécessairement un article unique et complet. Il est évoqué précédemment en section 3.1 que la numérisation de documents historiques conduit à une sursegmentation du texte et à un accroissement du nombre de segments par rapport à la quantité réelle d'articles.

Répartition des documents dans le temps

Dans le même esprit qu'ont été réalisées les analyses documentées en section 2.3, la répartition temporelle des documents donne des indications sur les quantités de textes disponibles et la conservation des journaux sur le long terme. En premier lieu, la figure 3.7 projette le nombre de documents publiés en allemand, finnois, français et suédois sur la décennie 1910 à 1920. On y constate d'abord une différence entre ce qui était supposé précédemment, à savoir une présence de documents dans toutes les langues, suggéré par la figure 3.5 et la réalité. En 1910, aucun document n'existe en allemand, de même que manquent en 1919 les textes en français et en suédois.

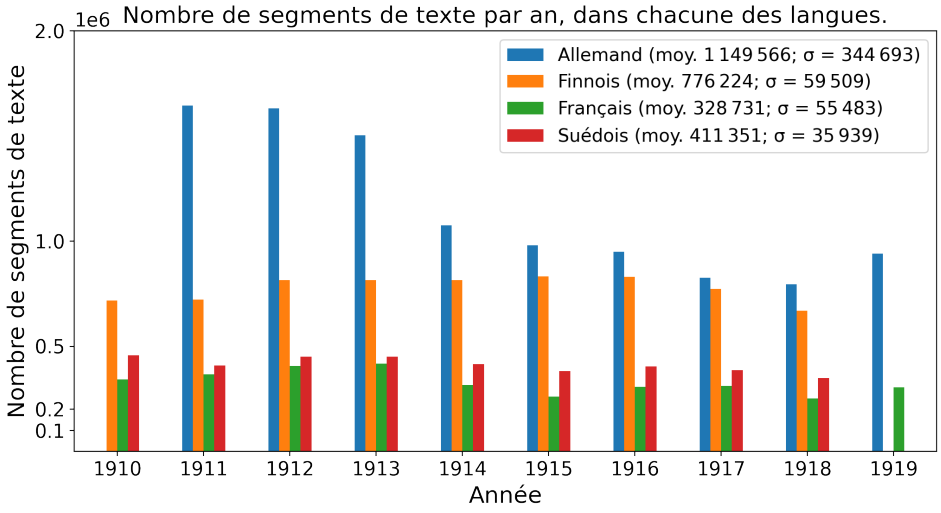
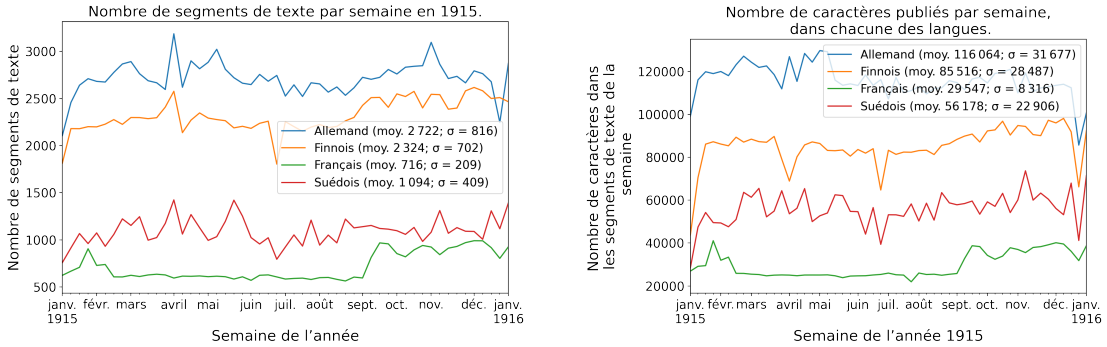


FIGURE 3.7 – Nombre de segments de texte pour chaque année, dans chaque langue de *NewsEye*.

La suite de cette analyse s'intéresse à l'année 1915 durant laquelle des articles sont

publiés dans toutes les langues disponibles et suivant une répartition de contenu qui semble homogène avec les années suivantes. Une autre différence domine : il est supposé y avoir sur cette décennie des articles provenant de cinq journaux en français, et c'est pourtant cette langue qui est la moins fournie en segments de texte. À ce stade, l'hypothèse est que les bibliothèques nationales qui fournissent les documents de *NewsEye* ne disposent pas des archives complètes sur ces périodes. La figure 3.8 donne deux types de statistiques sur l'année 1915 : le nombre de segments de texte par langue (figure 3.8a) ainsi que le nombre de termes (figure 3.8b) contenus dans tous les documents publiés. L'échelle de temps est la semaine. Le nombre de segments de texte est réparti relativement uniformément sur l'année. Pour le cas du français, une première explication à l'hypothèse suggérant sa moindre représentation est que le nombre de segments est faible (comparé aux autres langues) jusqu'au mois de septembre et double dans les mois qui suivent. Une nouvelle conjecture suppose le manque d'un ou plusieurs périodiques français entre février et septembre 1915. C'est la seule explication qui tienne compte du doublement du nombre de segments et de termes des textes entre septembre et décembre.



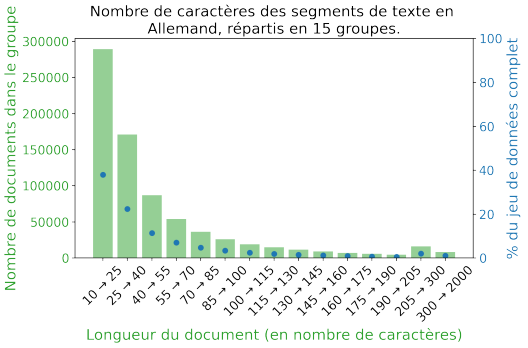
(a) Nombre de segments de texte. (b) Nombre de termes des segments de texte.

FIGURE 3.8 – Statistiques du contenu de NewsEye pour l'année 1915.

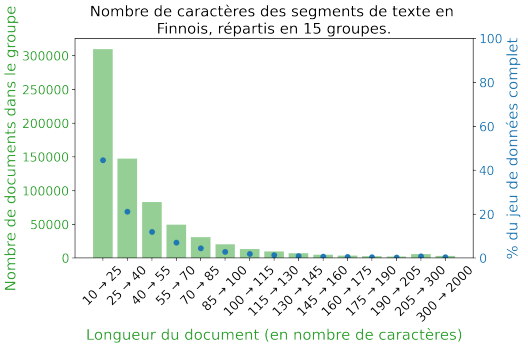
Longueur des documents

Nous l'avons déjà mentionné maintes fois, la longueur des documents est un facteur qu'il est nécessaire d'analyser et de connaître : le texte constitue la ressource de base des données. Nous comptons, de façon analogue à ce qui est réalisé pour les télégrammes, l'ensemble des termes de chaque segment de texte du corpus. Une projection de ces comptes, dans chaque langue, est présentée en figure 3.9. Les intervalles sont identiques à ceux de la figure 3.4a présentant les longueurs des télégrammes et les graphes comparables.

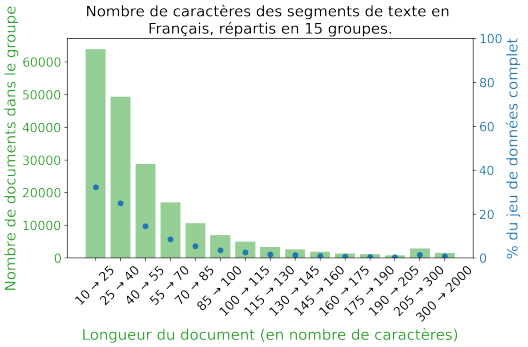
Le profil de toutes les courbes est identique : 40 % des segments de texte dans chacune des langues contiennent moins de vingt-cinq termes et les proportions de chaque intervalle sont toutes similaires, à quelques points près. Les documents de plus de 85 termes sont très rares et représentent légèrement plus de 5 % de la totalité des données. Ces



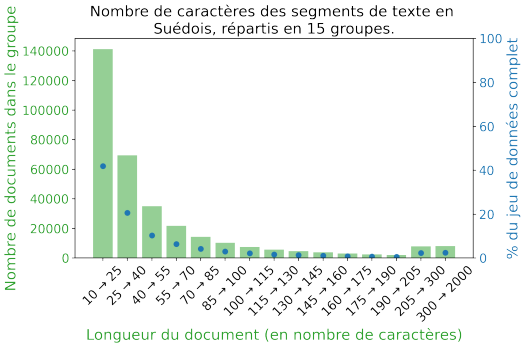
(a) En Allemand



(b) En Finnois



(c) En Français



(d) En Suédois

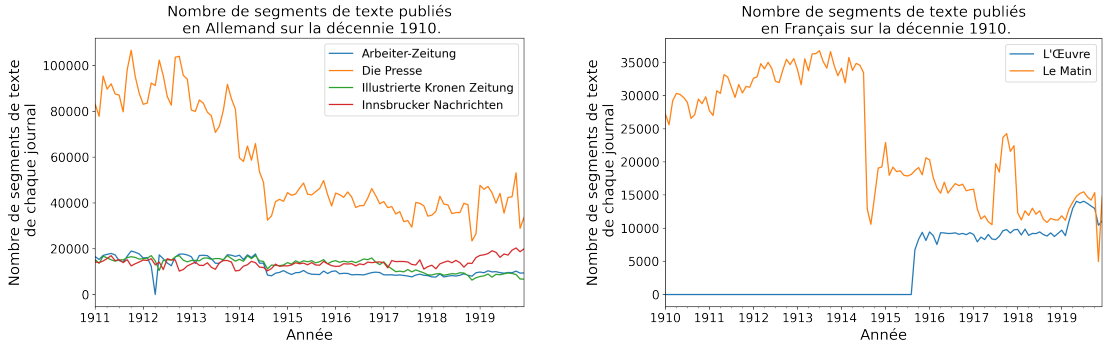
FIGURE 3.9 – Longueur des documents de *NewsEye* sur l'année 1915, en nombre de caractères

profils de longueurs accentuent l'hypothèse de sursegmentation du texte. Les documents historiques numérisés sont très fortement segmentés et contiennent de fait peu de termes, autant que des dépêches télégraphiques. Pourtant, là où les dépêches ont une unité de contenu, et répondent à des questions de base (quoi, quand, qui, etc.), ces segments de texte ne sont qu'une portion d'un article à part entière. Ils n'apportent pas toujours suffisamment d'informations pour appréhender le contexte des faits rapportés.

Disponibilité des périodiques

L'archivage et la numérisation des périodiques dépendent de l'état de conservation des journaux et de leur disponibilité. D'autres facteurs contextuels influent largement sur les publications et leur archivage. La Première Guerre mondiale a marqué la décennie 1910 – 1920, offrant l'opportunité d'analyser l'impact de la guerre sur les publications de presse. La figure 3.10 présente le nombre de segments de texte par journal en fonction du temps. Dans les deux langues, la quantité de contenu décroît significativement au milieu de l'année 1914, au déclenchement de la guerre entre ces deux pays. Les documents en allemand et en français n'ont pas la même origine : les premiers proviennent de la

Bibliothèque Nationale d'Autriche et les seconds de la Bibliothèque Nationale de France. L'origine des documents étant différente, cela ne peut constituer un phénomène isolé. Il est impossible, à partir de ces données, d'en déterminer la cause : cela peut-être dû à une diminution de la fréquence de publication de ces journaux, à la réduction de leur contenu où encore à un problème de conservation. Il ne s'agit pas d'une modification de la mise en page, qui aurait artificiellement réduit l'effet de la sursegmentation : le nombre de termes diminue dans les mêmes proportions que le nombre de segments. En allemand, la baisse est certes significative pour le journal Die Presse, mais elle l'est également et dans une moindre mesure pour les trois autres. En français, l'observation est identique. La quantité de segments de Le Matin diminue de moitié au milieu de l'année 1914. En outre, cette projection donne une explication à l'hypothèse formulée précédemment sur l'incohérence entre le nombre de publications supposées disponibles sur la période et la quantité de contenu effective. Pour la décennie, seulement deux périodiques sont disponibles au lieu de six, dont l'un à partir de l'année 1915, l'Œuvre.



(a) En Allemand (b) En Français

FIGURE 3.10 – Nombre de segments de texte par périodique et par mois, dans la décennie 1910.

3.2.3 Conclusion

Nous avons, dans cette section, analysé sous divers angles les documents du corpus *NewsEye*. Par la richesse de son contenu, son nombre de publications, la diversité des langues et parce qu'il intègre des données publiées sur un siècle, il est un candidat idéal à la problématique de détection et de suivi d'événements historiques rapportés par la presse. Comme explicité précédemment, les documents historiques numérisés se placent en opposition aux documents nativement numériques (sous-section 3.1.1) par leur forme et par leur contenu.

Chaque segment de texte est associé à quelques métadonnées, dont la date de publication, le nom du journal, etc. Aucune ne rattache les documents aux événements qu'ils mentionnent, c'est-à-dire, ne permet d'identifier de façon unique tous les documents rapportant le même événement historique. Cette limite majeure exclut le corpus *New-*

sEye du processus d'évaluation d'un algorithme recréant les histoires des événements à partir des documents. C'est le cas des jeux de données sélectionnés et présentés en section 2.3. Ceux-ci ne contiennent pourtant pas de données historiques numérisées et donc, n'intègrent pas les erreurs et défauts mentionnés précédemment. Pour faire du corpus *NewsEye* un jeu de données adapté à cet objectif, il est nécessaire d'annoter les documents avec des identifiants d'événements.

3.3 Association d'articles aux événements mentionnés

Les jeux de données *MUC* [CLH93], *TDT* [ALJ00], *ACE* [Dod+04] ou *Event Registry* [Rup+16] sont annotés d'événements. L'approche est différente pour chacun des projets, de même que les types d'annotations. Ce sont néanmoins ces travaux qui doivent inspirer une annotation d'événements pour tout projet dont l'objectif est la création d'un corpus adapté à la détection et au suivi d'événements mentionnés dans la presse.

Une annotation d'événement est, dans le cadre de ce travail, un identifiant qui associe différents documents de presse entre eux. Le point commun entre tous les documents annotés du même identifiant est qu'ils rapportent tous le même événement. Ils répondent collectivement aux mêmes questions « quoi », décrivant l'action, « qui » y prend part et « quand » a-t-elle eu lieu. C'est ce type d'annotation qui est fourni dans le corpus *Event Registry* [Rup+16]. Dans le tableau 3.1, deux documents sont rattachés au même événement, le n° 1129. Le troisième document rapporte un autre événement. D'après leurs contenus, ils semblent effectivement ne pas évoquer ni le même sujet ni le même événement.

#	Date	Titre	Corps de texte	Événement
4850	18/12/2013	STATE : Cal Fire arrests Clover Fire arson suspect	NORTHERN CALIFORNIA - Cal Fire law enforcement officers have arrested a Happy Valley man on suspicion of intentionally setting numerous fires throughout..	1129
48225	18/12/2013	Fire official : Man arrested on suspicion of arson in deadly California wildfire	REDDING, Calif. – Fire officials say a 29-year-old man has been arrested on suspicion of arson and murder in a blaze...	1129
16400	18/12/2013	Report : $4.5 million seized from bank chief's home	ANKARA, Turkey (AP) – Istanbul police leading a major corruption and bribery investigation targeting allies of Prime Minister Recep Tayyip Erdogan...	321

TABLEAU 3.1 – Documents d'*Event Registry* annotés d'identifiants d'événements identiques ou différents.

Des annotations de ce type ont également été réalisées au sein du projet *TDT* [Cie+02]. Des catégories de sujet, onze au total pour *TDT-2* et *TDT-3* couvrent des thématiques assez larges : les élections, les catastrophes naturelles, les accidents ou le sport, etc. Chaque catégorie est associée à un ensemble de règles d'interprétation per-

mettant d'uniformiser les annotations et aidant à déterminer si un document est bel et bien associé à un événement. Les annotateurs et annotatrices doivent connaître les onze différents types de sujets ainsi que les règles d'interprétation associées. Le domaine des sujets possibles est fini. Les annotations débutent avec un nombre de sujets limité, 100 pour *TDT-2*, 160 pour *TDT-3* que les annotateurs et annotatrices doivent connaître. L'annotation se fait en parcourant les documents et assigne une étiquette indiquant si le document est en relation avec un événement identifié. Les événements se construisent itérativement. Chaque document enrichit la description de l'événement annoté. Certains événements comme ceux du type d'élection ou catastrophe naturelle contiennent près de cinq cents documents différents.

Au sein d'*Event Registry*, les annotations sont similaires et manuelles [Rup+16], mais, à notre connaissance, le processus n'est pas détaillé. Il est probable, compte tenu du type d'annotation, que le procédé suivi soit semblable ou identique à celui défini dans *TDT*. Les auteurs ont développé une solution éponyme de détection et de suivi de mentions d'événements dans la presse numérique [Eve20]. On peut émettre l'hypothèse que leur processus d'annotation a été assisté par ordinateur. Ce serait une explication aux erreurs décrites dans la section 2.3.2.

Annoter des événements au sein du corpus *NewsEye* doit tenir compte de plusieurs contraintes. Le volume de données est très important ; il n'est pas possible d'annoter manuellement chaque document et le nombre d'événements mentionnés est critique : ce sont près de cent années d'événements qui sont rapportés dans les documents. Il est impossible pour un annotateur ou une annotatrice de connaître l'ensemble des événements référencés. Une stratégie possible mêle les deux procédés utilisés pour *TDT* et *Event Registry* : un nombre fini d'événements est sélectionné et les documents sont présélectionnés par un moyen numérique. La réduction artificielle du nombre d'événements et du nombre de documents est les conditions nécessaires pour envisager une annotation manuelle.

Le processus proposé est donc le suivant :

1. Le jeu de données est réduit pour ne contenir que des documents publiés sur une période courte, dans toutes les langues disponibles. L'analyse réalisée en sous-section 3.2.2 mène à considérer les années 1915 à 1920 comme de bonnes candidates dans le corpus *NewsEye*.
2. Les événements historiques sur la période sélectionnée sont collectés au moyen de sources ouvertes. Des bases de connaissances répertorient un grand nombre d'événements comme Wikidata [VK14] ou EventKG [GD19 ; AGD20]. Il est possible de les interroger pour obtenir l'ensemble des événements ayant eu lieu sur une plage donnée.
3. Le nombre d'événements est limité, de façon analogue à *TDT* : seul un sous-ensemble de ces événements est sélectionné. Pour rappel, *Event Registry* contient plus de 1 400 événements pour plus de 30 000 documents.
4. Des documents candidats sont associés à chaque événement sélectionné par des algorithmes ou des procédés numériques quelconques. Ils groupent les documents

qui traitent les mêmes thématiques. Ce sont les documents au sein de ces groupes qui sont annotés. Les chapitres 4 et 5 de ce document proposent des processus d'analyse simplifiant cette tâche.

5. L'annotateur ou annotatrice consulte un descriptif d'événement, comme c'est le cas dans *TDT* et indique si oui ou non le document consulté est lié à l'événement annoté.

Ce processus peut-être itératif et répété : à la première itération, les outils d'assistance mentionnés au point numéro 4 donnent des résultats moyens. Ils sont améliorés par apprentissage des premières annotations, générant de nouveaux groupes de documents pouvant ensuite être annotés, et ainsi de suite. Un tel processus itératif réduit le besoin d'expertise des annotateurs et annotatrices, automatise en partie le processus et limite les problèmes liés au volume de données utilisées. Les outils numériques d'assistance mentionnés ici sont par exemple ceux décrits dans les chapitres 4 et 5. L'annotation du corpus *NewsEye* pour le suivi de mentions d'événements est une perspective permise par ce travail de thèse.

En l'absence d'annotations d'identifiants d'événements sur les documents du corpus d'articles de *NewsEye*, nous proposons de réutiliser les jeux de données nativement numériques déjà annotés et mentionnés précédemment. La reproduction artificielle de certaines dégradations, comme celles mentionnées à la section 3.1, fournit une base pour analyser le comportement des algorithmes de détection et de suivi d'événements au sein de corpus historiques numérisés.

3.4 Simulation des caractéristiques de documents anciens

Nous l'avons étudié précédemment, les documents historiques présentent des spécificités de forme et de fond qui les distinguent des documents de presse récents et nativement numériques. Il est souhaitable que les travaux traitant des documents historiques manipulent effectivement des documents historiques. Le corpus *NewsEye* ne peut néanmoins pas être utilisé dans le contexte de cette thèse, car les annotations d'événements qu'il contient ne conviennent pas, comme soulevé plus haut. La seule solution permettant de pallier cette problématique est de reproduire artificiellement des dégradations sur des jeux de données annotés d'événements pour du suivi. Alors que les différences liées à la nature et au contenu même du texte sont difficilement reproductibles, il est possible d'altérer artificiellement du texte dans le but de reproduire les erreurs courantes d'un processus de numérisation. Les altérations du contenu proviennent du scanneur, de l'outil de segmentation ou de reconnaissance de caractères. Des documents récents sont utilisés pour reproduire en partie ces dégradations. Ce traitement est appliqué aux jeux de données annotés et choisis en section 2.3.

3.4.1 Dégradations introduites par la dégradation des images

La simulation de dégradations de texte par *OCR* a été réalisée de nombreuses fois par le passé, avec des études d'impact sur la détection d'entités nommées [Ham+19 ; Lin+19]

ou l'extraction d'événements [Bor+20b]. L'objectif de la simulation des dégradations est de reproduire au mieux les altérations qui peuvent apparaître dans un document ancien numérisé. En réalisant des dégradations spécifiques et contrôlées, nous serons en mesure d'obtenir des indications sur le type de correction ou d'amélioration à proposer aux algorithmes de détection et de suivi des mentions d'événements. Le processus est décrit schématiquement dans la figure 3.11 et se compose de quatre éléments fonctionnels :

— **La conversion de texte en images**. ImageMagick [LLC99] transforme un texte en image de 750 pixels de large, avec un font blanc, une police noire de type Times New Roman, police de presse utilisée par le journal britannique *The Times* dès l'année 1932 [C F46].
— **La dégradation des images pour simuler des documents anciens**. DocCreator [Jou+17 ; Na17] dégrade les images avec des effets de flou, de transparence, l'ajout de caractères fantômes ou de trous, par exemple. Ce sont en partie les dégradations liées à l'image que nous avons identifiées à la section 3.1.
— **La reconnaissance optique de caractères**. Tesseract-OCR [Rc75] extrait le texte des images. C'est ce composant du processus qui rend visibles les dégradations de l'image en introduisant des erreurs dans le texte qui n'existent pas dans celui d'origine.
— **Le chargement des données**. Les jeux de données sont reconstruits dans le format d'origine, pour être manipulés dans un contexte identique et être ainsi comparables.

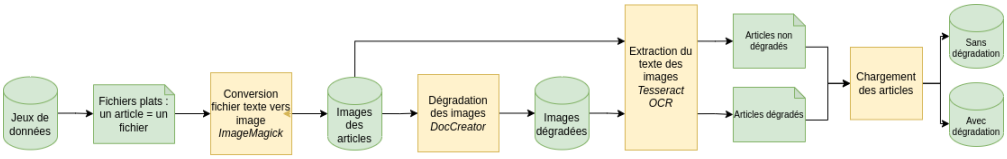


FIGURE 3.11 – Schéma du processus de dégradation de documents par OCR.

DocCreator reproduit des dégradations couramment observées dans les documents historiques numérisés. Elles sont additives dans le sens où les documents historiques présentent en général l'ensemble de ces altérations [Jou+17 ; Lin+19] et pas seulement un sous-ensemble distinct et isolé.

Aux trois jeux de données mentionnés, *Event Registry*, *CoAID* et *Fibvid*, un quatrième est ajouté, issu d'*Event Registry* duquel seuls les titres sont conservés. Les titres sont des textes courts à l'instar des dépêches télégraphiques, bien que leur contenu diffère énormément : ils sont plus courts et sont rédigés selon un style différent. Ils répondent pourtant aux mêmes questions pour cerner les événements : quoi, qui, et quand. Pour la suite, ce jeu de données est nommé *Event Registry Titles*.

Nous générons deux nouveaux corpus pour chacun des jeux de données sélectionnés à la section 2.3. Le premier est créé en transformant le texte en image puis en appliquant l'*OCR* sans dégrader les images. L'autre suit le même processus de création, mais les images sont abîmées. Nous altérons les images en dégradant les caractères, nous introduisons des caractères fantômes, ajoutons un effet de transparence et nous les floutons.

Ce sont ces dégradations que nous avons présentées précédemment à la sous-section 3.1.1 (page 71). La problématique de distorsion présentée dans l'exemple de la figure 3.1b n'est pas considérée.

Pour analyser l'effet des différentes dégradations, nous calculons les taux d'erreur de caractères (*CER*) et de mots (*WER*) après la dégradation des jeux de données *Event Registry*, *CoAID* et *FibVid*. Les résultats sont rapportés dans le tableau 3.2 pour *Event Registry*, et le tableau 3.3 pour *Event Registry Titles*, *CoAID* et *FibVid*. Les jeux de données traités dans la suite de ce document sont respectivement ceux sans dégradations et ceux les intégrant toutes. La dégradation des caractères est la seule à entraîner un taux d'erreur élevé. Les autres dégradations ont un impact mineur, par exemple l'effet de flou, ou négligeable comme la présence de caractères fantômes ou l'effet de transparence. Pour ces deux derniers, les taux d'erreurs sont comparables aux taux en l'absence d'altération sur les documents. Les ordres de grandeur de nos résultats sont les mêmes que ceux de l'étude dont nous nous inspirons [Lin+19].

Type d'erreur	**Dégradation**					
	Sans	Caractère	Fantôme	Transparence	Flou	**Toutes**
CER	0,282	4,154	0,274	0,275	0,582	4,577
WER	0,552	16,364	0,551	0,548	1,159	16,974

Tableau 3.2 – Taux d'erreurs de caractères et de mots après dégradation du corpus *Event Registry*.

Les conclusions sont identiques pour les jeux de données courts. Les taux d'erreurs sont légèrement plus élevés à cause de la taille très réduite des documents analysés. La dégradation de caractères est à nouveau l'altération qui affecte le plus les capacités de reconnaissance de l'*OCR*, suivie de loin par l'effet de flou. Les effets de caractères fantômes et l'effet de transparence ont ici aussi un impact négligeable.

Jeu de données	**Type d'erreur**	**Dégradations**					
		Sans	Caractère	Fantôme	Transparence	Flou	**Toutes**
Event Registry Titles	*CER*	2,421	6,940	2,414	2,422	2,874	7,178
	WER	1,127	19,785	1,124	1,131	2,035	19,894
CoAID	*CER*	2,105	6,358	2,105	2,122	2,616	7,898
	WER	2,494	20,230	2,496	2,580	3,726	20,230
FibVid	*CER*	1,463	6,089	1,461	1,467	1,935	6,359
	WER	2,065	20,797	2,041	2,052	2,868	21,396
Moyenne	CER	1,996	6,462	1,993	2,004	2,475	7,145
Écart type		0,448	0,435	0,486	0,488	0,485	0,770
Moyenne	WER	1,895	20,271	1,887	1,921	2,876	21,451
Écart type		0,699	0,507	0,699	0,734	0,846	1,585

Tableau 3.3 – Taux d'erreurs de caractères et de mots après dégradation des jeux de données courts : *Event Registry Titles*, *CoAID* et *FibVid*.

Pour chacun des quatre jeux de données sélectionnés, deux nouveaux sont synthétisés. Ils présentent de façon équivalente des défauts liés à l'application d'un *OCR* sur des

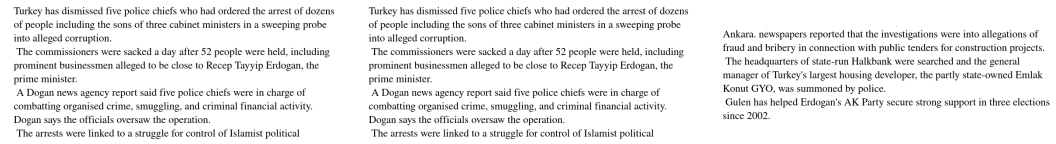
documents numérisés altérés. Pour le premier, l'image utilisée n'est pas dégradée, pour l'autre, l'ensemble des images le sont, entraînant une augmentation significative du taux d'erreur. L'objectif de ces multiples dégradations et de reproduire du mieux possible les défauts de numérisation qui existeraient dans un corpus de documents historiques. Au-delà des problématiques de reconnaissances de caractères, une autre dégradation est évoquée tout au long de ce document : la sursegmentation du texte. Cette caractéristique peut aussi être synthétisée et ne doit pas être négligée.

3.4.2 Dégradations introduites par la segmentation du texte

Au-delà des erreurs de reconnaissance du texte par les techniques d'*OCR*, les logiciels de segmentation de texte sont parfois dans l'incapacité d'identifier correctement les articles de presse complets. Le phénomène généralement constaté est celui de sursegmentation : les articles de presse sont divisés en un ensemble de paragraphes, dits segments, qui ne sont pas connectés entre eux pour former un article. Il est par conséquent impossible de supposer qu'un segment de texte constitue un unique article et que celui-ci est complet. L'analyse du contenu ou de la position des segments [WWC82 ; Pal+12 ; Mic+21] dans l'image numérisée permettent l'identification d'une continuité entre les segments et de reconstituer de la linéarité dans la lecture des segments de texte.

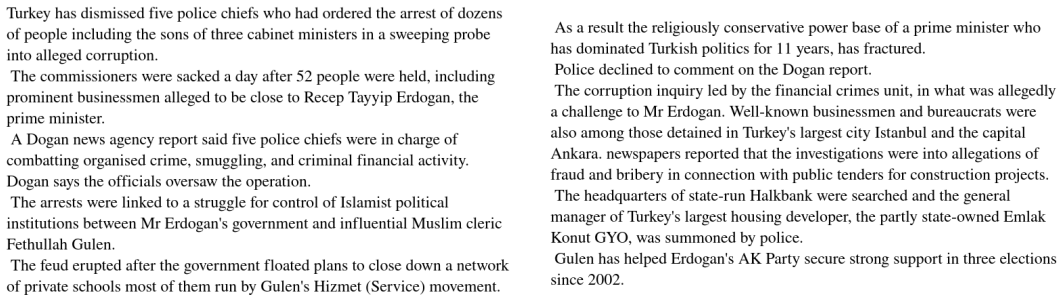
Les sauts de colonne, de paragraphes ainsi que la séparation des textes par des éléments visuels sont les origines principales de sursegmentation du texte. La tradition typographique en presse écrite est de répartir le texte sur une multitude de colonnes. La page elle-même peut-être subdivisée en plusieurs sections horizontales au sein desquelles les articles sont répartis sur plusieurs colonnes de texte. La séparation en segments d'un article est aussi causée par l'introduction d'images, de publicités ou d'éléments séparateurs variés au sein même d'une colonne. Le logiciel de segmentation réalise de fait parfaitement sa tâche : il segmente les paragraphes là où il semble y avoir une séparation nette. Les sauts de paragraphe ou de colonne sont des motifs légitimes de segmentation. C'est l'application de ces outils à la presse numérisée, historique ou non, qui ouvre la voie vers une nouvelle problématique : il devient nécessaire de reconstruire des articles à partir de données sursegmentées.

En sus des problèmes de reconnaissance de caractères, la sursegmentation est un fait rencontré lors du traitement de journaux numérisés. Parmi les corpus retenus, seul *Event Registry* contient des articles de presse et c'est le seul sur lequel il est possible de segmenter artificiellement le texte. Elle intervient après application des dégradations de la sous-section 3.4.1. Pour la simuler, les textes extraits par *OCR* sont divisés uniformément en deux ou en trois sections. Ce découpage correspond à une simulation de sauts de colonnes, mais qui traite le problème général d'un texte fractionné en plusieurs sous-ensembles de longueurs variables. Le nombre de divisions est arbitraire et vise à évaluer l'impact des segmentations par rapport à un même jeu de données n'en contenant aucune. Cette approche présente toutefois un biais. Contrairement aux données générées par un outil de segmentation de texte, les segments artificiels que nous créons contiennent un même nombre de lignes. Les figures 3.12 et 3.13 présentent des segmentations en deux ou en trois portions de l'article n° 21 200 du corpus *Event Registry*.



(a) Premier segment (b) Second segment (c) Troisième segment

FIGURE 3.12 – Article de presse n° 21 200 segmenté en trois portions.



(a) Premier segment (b) Second segment

FIGURE 3.13 – Article de presse n° 21 200 segmenté en deux portions.

Pour l'ensemble des documents segmentés, le titre de l'article n'est associé qu'au premier segment de texte. Dans ce cas particulier, nous émettons l'hypothèse que l'outil de segmentation associe correctement le titre de l'article au premier segment. La réalité est évidemment plus complexe et le titre peut être isolé des autres segments de l'article par le processus de segmentation.

Les articles des données *Event Registry*, dégradés ou non, sont segmentés en deux, ou en trois. Le nombre de jeux de données de presse double à nouveau et permet en plus de l'évaluation des erreurs liées à l'*OCR*, d'évaluer les effets négatifs ou positifs de la sursegmentation du contenu.

3.4.3 Jeux de données dégradées disponibles

Des trois jeux de données sélectionnés initialement, neuf sont générés artificiellement. Ces nouveaux corpus contiennent quelques dégradations courantes d'articles de presse historique numérisés. Celles retenues sont typiquement liées à l'altération du document d'origine, à sa numérisation ou aux outils de reconnaissance optique de caractères. Dans un second temps, des problématiques de segmentation sont reproduites en divisant les articles en portions de longueurs identiques. Pour synthétiser, les nouveaux jeux de données générés sont répertoriés dans le tableau 3.4.

Le jeu de données d'articles *Event Registry* est exploitable suivant six configurations différentes, selon la présence ou l'absence des dégradations et des segmentations. Pour *Event Registry Titles*, *CoAID* et *FibVid* qui ne contiennent que des brèves ou titres de presse, courts, ils ne sont pas divisibles et contiennent seulement des erreurs liées à

Jeu de données	Langues	Dégradation *OCR*	Segmentation	Nombre de documents	
				Entraînement	Test
Event Registry	anglais, espagnol, allemand	Non	Non	20 803	13 004
		Non	Division en 2	41 567	25 979
		Non	Division en 3	61 450	38 576
		Oui	Non	20 803	13 004
		Oui	Division en 2	41 567	25 979
		Oui	Division en 3	61 450	38 576
Event Registry Titles	anglais, espagnol, allemand	Non	Non	20 803	13 004
		Oui	Non		
CoAID	anglais	Non	Non	72 045	31 200
		Oui	Non		
FibVid	anglais	Non	Non	988	402
		Oui	Non		

TABLEAU 3.4 – Descriptif des jeux de données synthétisés, intégrant des erreurs de reconnaissance ou de segmentation.

l'*OCR*. Tous les jeux de données dégradés sont publiés dans le cadre de cette thèse, de même que les images issues des dégradations décrites dans cette section : *Event Registry* [Ber22i](images [Ber22g]), *Event Registry Titles* [Ber22l] (images [Ber22k]), *CoAID* [Ber22f] (images [Ber22d]) et enfin *FibVid* [Ber22p] (images [Ber22n]). De la même façon, nous mettons à dispositions tous les outils utilisés pour dégrader ces documents afin qu'ils puissent être réutilisés dans un contexte similaire à celui décrit dans ce chapitre [Ber22s].

3.5 Format de données pour des expériences reproductibles

La question de la reproductibilité de la recherche a atteint un *« point critique »* à la fin des années 2010 [Des+19]. La non-reproductibilité des expériences scientifiques est un élément concourant à la défiance du public et des institutions envers la communauté scientifique. Spécifiquement en informatique, cette notion de reproductibilité recouvre de nombreuses réalités : gestion des données, des codes sources, des modèles d'apprentissage, des nombres aléatoires, etc. Les programmes logiciels liés à cette thèse sont archivés sur la plate-forme *Software Heritage* [DZ17], qu'ils soient dédiés à l'analyse de données ou qu'ils implémentent les algorithmes que nous verrons dans les chapitres à venir.

Pour pallier ce problème, la solution semble la conservation complète des données traitées. En effet, certains outils d'extraction d'entités, pour ne citer qu'eux, reposent sur des modèles qui peuvent évoluer ou disparaître. Il n'est pas raisonnable de considérer qu'il sera éternellement possible de les obtenir pour reproduire des expériences. En 2022, le coût du giga-octet de stockage est faible et autorise ce mode de raisonnement. Pour des documents historiques de presse rapportant des événements, différents types de traitements peuvent être réalisés et des informations extraites :

— **Nettoyage du texte**. Des post-traitements sont appliqués au texte qui peut

contenir des marques laissées par l'*OCR* (espaces surnuméraires, sauts de ligne, etc.) ou dans lequel des corrections sont appliquées. Enfin, une post-correction des sorties d'*OCR* réduit les erreurs présentes dans les textes [Ngu20].

— **Identification des termes**. Les termes (mots, jetons) sont identifiés dans le texte suivant différentes techniques, qui peuvent ou non tenir compte des mots-composés, des noms propres, et sont adaptés à la langue analysée.
— **Détection des entités nommées** [HHD20]. Les noms de lieux, de personnes, les dates ou d'autres entités sont détectées dans le texte et un type leur est affecté. Comme expliqué précédemment, ce sont les entités nommées qui portent l'essentiel de l'information dans des documents textuels [WL11 ; YB18].
— **Identification de l'opinion**. Un sentiment est assigné à chaque entité nommée détectée dans le texte, qu'elle soit négative (-1), neutre (0) ou positive (1).
— **Identification des entités nommées**. Les entités nommées sont connectées à des bases de connaissances par des techniques d'*Entity Linking* [Lin+19 ; Lin+20]. Il s'agit d'identifier précisément le concept représenté par l'entité.
— **Détection des événements**. Les événements rapportés dans les différentes phrases du texte sont identifiés et typés selon les taxonomies proposées par ACE 2005 [Dod+04].

La seule diffusion des textes dégradés ou non des jeux de données est problématique, au regard de la seule reproductibilité. Pour certaines techniques de vectorisation de document, le texte doit être nettoyé et les procédés diffèrent, la détection d'entités nommées avec un modèle plutôt qu'un autre peut donner des résultats dissemblables. C'est pour contrer cette limite qu'il semble prudent de proposer un format de document adapté à la conservation des données de presse desquelles des caractéristiques sont extraites (jetons, événements, mentions, etc.). Ce sont ces documents, enrichis qui devraient être diffusés en lieu des textes seuls. Ils concourent à faciliter la reproductibilité des expériences scientifiques et à limiter les tâches, parfois complexes, de préparation des documents.

Le code 3.1 est un exemple de document, tronqué, intégrant toutes les extractions mentionnées ci-dessus. Le texte est disponible en deux versions, brute ou nettoyée, les jetons ou termes identifiés au sein des phrases, les entités nommées détectées et identifiées, de même que les événements. On y découvre également l'annotation d'événement telle que proposée en section 3.3. Ce document fait partie des nombreux articles du corpus *NewsEye* décrivant l'attentat de Sarajevo, survenu en 1914. Une collection de documents de ce type forme un corpus, qu'il est plus aisé d'identifier et de publier sur des archives numériques telles Zenodo [EO13]. Les différents jeux de données disposent alors d'un identifiant objet (*DOI*) unique et les corpus utilisés pour une expérience peuvent être cités.

Code 3.1 – Exemple de document formaté pour la conservation des différentes caractéristiques extraites à partir du texte.

```
1 {
2   "_id": "arbeiter_zeitung_aze19140703_article_40",
3   "language": "de",
4   "date": "1914-07-03",
```

```
  "newspaper": {
    "title": "Arbeiter-Zeitung",
    "languages": ["de"],
    "uri": ["https://www.wikidata.org/wiki/Q627083"],
    "inception": "1889-07-12",
    "end": "1991-10-31",
    }
  "event": {
    "label": "Assassination␣of␣Archduke␣Franz␣Ferdinand"
    "uri": "https://www.wikidata.org/wiki/Q192050"
  },
  "text": {
    "raw": "[...]",
    "cleaned": "[...]␣Vor␣einigen␣Wochen␣habe␣er␣in␣Belgrad␣in␣
       ↪ einem␣Kaffeehause␣gesessen␣und␣in␣einem␣Blatt␣gelesen,␣
       ↪ da␣Erzherzog␣Franz␣Ferdinand␣Ende␣Juni␣in␣Sarajevo␣
       ↪ eintreffe.␣[...]␣das␣Vaterland␣sterben␣wollen.␣[...]",
    "sentences": [
      [
        ...
        {
          "id": 14, "token": "Belgrad",
          "start_index": 93, "end_index": 100
        },
        ...
        {
          "id": 116, "token": "sterben",
          "start_index": 633, "end_index": 640
        },
      ],
    ]
  },
  "mentions": [
    {
      "type": "LOC",
      "text": "Belgrad",
      "article_start_index": 93,
      "article_end_index": 100,
      "stance": 0,
      "id": 1,
      "linked_entities": [
        {"uri": "https://wikidata.org/wiki/Q3711"}
      ],
      "tokens": [14]
    }
    ...
  ],
  "events": [
    ...
```

```
    {
      "id": 4, "trigger": {"type": "Die", "tokens": [116]}
    }
  ]
}
```

Ce document suit les principes du format *JSON*, largement utilisé dans des bases de données dites « *NoSQL* » et des systèmes d'indexation de documents. Des outils comme Apache *SOLR* [Fou22b] ou *ElasticSearch* [BV22] sont en mesure de conserver et d'indexer ces documents. Ils fournissent un système d'interrogation permettant de rechercher et d'interroger les textes en utilisant le programme Apache LUCENE [Fou22a] adapté à la fouille de texte.

3.6 Conclusion

Nous l'avons vu tout au long de ce chapitre, la presse historique numérisée est un type de document bien particulier. Une fois dépassé ce qui rapproche presse récente et presse ancienne, la forme journalistique et l'objectif d'information du public, les différences se font jour. Tout d'abord, le processus de numérisation et l'application successive des outils d'extraction de contenu sont générateurs d'erreurs. La source de l'information est le document physique, qui peut être dégradé ou mal numérisé. Les outils de segmentation et d'*OCR* y sont sensibles : le texte est très largement sursegmenté, comme nous l'avons démontré dans l'analyse des données de *NewsEye*. Des défauts très couramment observés dans les copies numériques de pages de journal mènent à des erreurs de reconnaissance du texte. Le texte historique numérisé est grêlé d'erreurs diverses : les articles détectés ne sont pas intègres et des substitutions, ajouts ou suppressions de texte apparaissent çà et là. Le temps est également un facteur primordial lorsque l'on étudie la diffusion d'informations dans la presse historique. Ces dernières transitent par des télégrammes, les dépêches circulent lentement dans un premier temps, au gré des réseaux de communication disponibles à l'époque, et parfois du contexte politique ou géographique.

Ensuite, le projet *NewsEye*, par sa dimension européenne et la quantité de documents qu'il permet d'analyser est un excellent candidat pour détecter et suivre les mentions d'événements dans des documents historiques. Contenant des articles rédigés pendant un siècle, dans plusieurs langues, il offre l'opportunité d'étudier les événements majeurs de l'histoire et comment ils ont été rapportés par la presse de l'époque. L'analyse de ces articles met en avant une disparité de contenu au niveau des langues et des journaux disponibles. Sa richesse souffre d'un défaut dans notre contexte d'étude : les annotations d'événements sont au format *ACE* et nous ne pouvons pas les utiliser, elles ne correspondent pas au format d'annotation que nous avons présenté au chapitre 2. Nous proposons cependant une stratégie d'annotation se basant sur les travaux passés réalisés au sein des projets *TDT* et *Event Registry*. Les travaux développés durant cette thèse permettront de mener une campagne d'annotation suivant les principes et règles présentés.

Pour suppléer ces données manquantes, un processus de dégradation reprenant les défauts courants de documents historiques (erreurs d'*OCR* et sursegmentation des textes) est proposé. Il se base sur des données récentes, nativement numériques, auxquels est appliquée une série de dégradations. L'objectif est d'étudier les comportements d'algorithmes de suivi de mentions d'événements face à des documents présentant ces typologies d'erreurs. Ce sont au total neuf jeux de données différents qui sont générés et peuvent être utilisés à des fins expérimentales.

Enfin, la question de la reproductibilité des expériences de recherche scientifique est évoquée. En traitement du langage naturel, de nombreux outils existent pour réaliser un ensemble de tâches variées : extraction d'entités nommées, détection d'événements, etc. Ces travaux utilisent parfois des modèles et des ressources spécifiques ou non publiées, ce qui limite voire empêche la reproduction de certaines expériences. Pour pallier ce phénomène, nous proposons un format de document qui organise et stocke l'ensemble des traitements applicables à des données de presse : détection d'événements, d'entités, nettoyage du texte, extraction des termes, etc.

Après avoir défini les contraintes des documents de presse ancienne numérisée, nous allons nous intéresser aux procédés algorithmiques et aux systèmes capables d'identifier et de reconstruire les déroulés des événements rapportés dans la presse historique.

Chapitre 4

Des documents de presse aux événements

Sommaire

Nous venons de le voir au chapitre précédent, manipuler des numérisations de presse historique implique de manier des documents pour partie dégradés. Ces dégradations ont nécessairement un impact sur les différents algorithmes que l'on peut exploiter dans le cadre du suivi d'événements. En état de l'art, à la section 2.2, nous présentons des techniques basées sur du *clustering* pour identifier des groupes d'articles décrivant des événements identiques. À partir d'un ensemble d'articles de presse, les algorithmes identifient ceux qui décrivent les mêmes sujets et les groupent au sein de *clusters*. Les informations qui décrivent les événements sont extraites de chacun de ces groupes : quel est l'événement, quand a-t-il eu lieu, ou/et qui y a participé. C'est en ce sens que cette approche est qualifiée « de documents aux événements » : les événements sont caractérisés après avoir groupé les articles qui les décrivent.

Tous les algorithmes que nous y avons cités partagent un point commun : les documents qu'ils manipulent sont encodés numériquement, condition préalable à leur traitement par les algorithmes que nous présentons. Nous allons dans un premier temps, à la section 4.1, évoquer la problématique de vectorisation du texte. Les encodages influent sur les algorithmes et, pour des données identiques, nous supposons qu'ils expliquent des variations dans les résultats obtenus. Nous introduirons les différentes possibilités de vectorisation des documents, les méthodes statistiques de pondération *TF-IDF* puis les encodages générés par des modèles d'apprentissage profond.

De nombreux algorithmes ont été proposés au fil des ans pour suivre des événements mentionnés dans la presse [All02b ; Pou+04 ; AY06 ; Rup+16 ; Mir+18 ; MBC19 ; LH20 ; SMM22]. Certains sont supervisés et d'autres non. Ces deux catégories répondent à deux besoins distincts. En section 4.2, nous présenterons deux algorithmes que nous avons implémentés pour suivre les mentions d'événements dans la presse.

Enfin, à la section 4.3, nous présenterons nos résultats issus de l'application des différents algorithmes sur les données synthétisées au chapitre 3. Enfin, nous évaluerons les résultats selon deux types de configurations, représentatives du contexte des mentions d'événements dans la presse ancienne : la première concerne les articles de presse longs, dégradés et segmentés, ceux issus du corpus *Event Registry*. L'autre configuration de notre analyse se base sur l'étude des brèves et des textes courts en utilisant les jeux de données *CoAID*, *FibVid* et *Event Registry Titles*.

4.1 Vectorisation des documents

Pour être traités par des algorithmes, les documents ne peuvent être manipulés tels quels, sous forme textuelle ou structurée comme présentée dans la section 3.5. Les ordinateurs et les algorithmes traitent des informations numériques et les documents doivent être transformés, le plus généralement, en vecteurs de nombres : c'est la vectorisation de document.

L'objectif est toujours d'obtenir, à partir d'un texte source, un vecteur de nombres encapsulant les données par une application bijective. Chaque texte est associé à un vecteur toujours identique, décrivant son contenu d'une manière unique. Deux moyens sont couramment utilisés pour obtenir ces vectorisations : l'application de la pondé-

ration statistique *TF-IDF*, qui génère des vecteurs creux et l'encodage par un modèle d'apprentissage profond qui génère des vecteurs denses. Le premier est un vecteur de taille variable, dépendant du vocabulaire d'entrée, dont la plupart des composantes sont nulles. Le second est un vecteur de taille fixe dont les composantes sont toutes valuées.

Par ces multiples vectorisations, il est possible d'évaluer le comportement des algorithmes en fonction du type d'encodage réalisé. Considérant les dégradations liées à la nature des documents historiques, nous cherchons à identifier un cadre expérimental approprié au suivi des mentions d'événements. Nous cherchons à confirmer ou infirmer l'hypothèse selon laquelle les vecteurs denses ne sont pas adaptés à la problématique de suivi de mentions d'événements dans la presse [Sta+19]. Les travaux menant à ces conclusions se basent sur des techniques publiées durant la première moitié des années 2010, notamment la vectorisation `doc2vec` [LM14]. Nous proposons d'affiner ces conclusions en tenant compte de l'évolution récente de l'état de l'art et de la publication des architectures d'apprentissage profond comme *ELMo* [Pet+18], *BERT* [Dev+19] ou *XL-NET* [Yan+20].

4.1.1 Par pondération des constituants du texte

La première méthode de vectorisation utilisée est appelée *TF-IDF* pour *Term-Frequency (TF) Inverse Document Frequency (IDF)*. Cette méthode mesure le poids, l'importance d'un terme dans un document, relativement à sa présence dans un corpus. Ce poids augmente ou diminue en fonction de la fréquence d'apparition du terme dans un document rapporté au reste du corpus de textes dont il provient. Cette méthode est largement utilisée dans la littérature pour la pondération de constituants d'un texte [ALJ00 ; Pou+04 ; AY06 ; AY10 ; RM12 ; Leb+14 ; Mir+18 ; MBC19 ; SMM22]. Le poids de chaque terme d'un document se calcule à l'aide des équations présentées ci-après, et une fois le processus d'extraction de termes (*tokenization*) achevé.

Le calcul de fréquence (équation 4.1) de chaque terme est le rapport entre le nombre d'occurrences du terme t dans le document d, noté $f_{t,d}$ et le nombre total de termes du corpus $\sum_{t' \in d} f_{t',d}$.

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{4.1}$$

La fréquence d'apparition de chaque terme est pondérée à la présence globale du terme dans le corpus. L'objectif du calcul *TF-IDF* est de donner un poids plus fort aux termes les moins fréquents et un poids plus faible à ceux plus courants. Le calcul de pondération (équation 4.2) est l'application d'une fonction logarithme sur le rapport entre le nombre de documents du corpus $|D|$ et le nombre de documents dans lesquels le terme t est présent. L'application d'une fonction logarithme, en base de deux ou en base dix accentue les effets du rapport entre le numérateur et le dénominateur.

$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \tag{4.2}$$

La pondération d'un terme t dans un document d relativement à un corpus de documents D s'obtient par multiplication de la fréquence de ce terme dans le document multiplié par sa fréquence relative au reste du corpus, décrite par l'équation 4.3

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \tag{4.3}$$

Pour chaque terme d'un texte donné, la méthode de pondération *TF-IDF* entraîne une importante quantité de calculs, d'autant plus que la quantité $|D|$ de documents est importante.

Une pratique courante de vectorisation *TF-IDF* est de calculer la pondération de chaque terme d'un corpus D relativement à sa présence au sein de ce même corpus D. Ce procédé implique de posséder à l'initialisation tous les documents constituant le corpus D pour calculer l'$idf(t, D)$. Dans le cas contraire, un corpus de documents de référence est nécessaire. S'il est de taille suffisamment importante, il offre un aperçu statistique de l'utilisation des termes dans une langue, pour une pratique donnée (littérature, journalisme, etc.). C'est-à-dire que si l'on cherche à pondérer des termes issus de documents de presse, ce corpus de référence doit être composé d'articles de presse rédigés dans la même langue que les documents à encoder.

Il est fait mention dans l'état de l'art d'algorithmes [PSD08 ; Rup+16 ; Mir+18 ; LH20] qui traitent des flux de documents itérativement. Dans un tel contexte, les articles ne sont pas connus par avance puisqu'ils ne sont pas encore publiés par les organes de presse. Pour surmonter cette difficulté, il est nécessaire de créer des corpus de pondération spécifiques en collectant en amont des articles de presse pour chacune des langues. Les corpus sélectionnés pour l'évaluation étant rédigés en anglais, en allemand et en espagnol, ce sont autant d'articles dans ces langues qui sont à collecter.

Nous l'avons dit, les corpus utilisés pour la vectorisation *TF-IDF* doivent être les plus proches possibles des données originales. Puisque nous traitons des corpus de presse (*Event Registry*) et des corpus de brèves et informations publiées sur Twitter (*CoAID* et *FibVid*), ce sont des données du même type que nous allons collecter. Nous construisons d'abord un corpus d'articles de presse puis un corpus de tweets. Le site Internet de presse *Deutsche Welle*[1] publie des articles dans plusieurs langues européennes, dont l'allemand, l'anglais et l'espagnol. Des articles dans ces trois langues sont collectés pour créer des corpus de pondération. Les textes du titre et du corps de chaque article sont conservés, de même que quelques métadonnées. Pour pondérer les tweets, ce sont des publications d'organes de presse sur Twitter qui sont collectées. Il s'agit par exemple du compte de la BBC (@BBCNews[2]) ou de l'AFP (@afp[3]). Alors qu'il est possible de télécharger sans restriction depuis le site de *Deutsche Welle*, l'API Twitter pose des limites à son utilisation. Il est impossible de télécharger plus de 3 200 publications par compte[4]. Pour outrepasser cette restriction, nous avons multiplié les sources et téléchargé

1. `https://www.dw.com`
2. `https://twitter.com/bbcnews`
3. `https://twitter.com/afp`
4. API Twitter v1 : `get-statuses/user-timeline` (archive sur `https://web.archive.org`)

les publications de 19 agences et journaux pour chaque langue. Nous obtenons ainsi des milliers de tweets dans chacune des langues, comme le montre le tableau 4.1.

Les statistiques du contenu de chacun des corpus, dans chaque langue, sont présentées dans ce même tableau. Le nombre de termes, de lemmes et d'entités est décompté pour donner un aperçu de la diversité linguistique des données.

Jeu de données	**Langue**	**Documents**	**Nombre de composants total**		
			Termes	Lemmes	Entités
Presse	Anglais	66 215	13 135 162	12 205 181	881 298
	Espagnol	69 785	16 837 878	15 517 875	1 667 356
	Allemand	72 641	12 838 604	12 254 457	989 835
Tweets	Anglais	53 907	546 625	544 538	48 708
	Espagnol	56 760	484 906	483 608	50 169
	Allemand	51 648	448 368	447 711	49 338

TABLEAU 4.1 – Contenu des deux jeux de données utilisés pour la pondération *TF-IDF* des textes.

Les deux algorithmes de suivi d'événements présentés ci-après en section 4.2 exploitent diverses composantes du texte : les termes, les lemmes et les entités. Chacune est constituée de chaînes de caractères pondérables par *TF-IDF* : le texte d'un terme, d'un lemme ou d'une entité. À chaque composante, dans chaque langue correspond un modèle de pondération *TF-IDF*. C'est-à-dire que les entités nommées de chaque document en anglais sont pondérées en fonction des entités nommées de tous les documents en anglais et seulement par rapport à celles-ci. Au lieu d'utiliser un unique vecteur de caractéristiques de document, la proposition faite par Miranda et coll. [Mir+18] est de définir de multiples vecteurs de poids pour chacun des articles. Le tableau 4.2 synthétise les vecteurs de caractéristiques à calculer pour chaque type de document. Sont donc calculées, pour chacun des articles de presse, trois pondérations sur le titre de l'article, une sur les termes, une sur les lemmes et une dernière sur les entités. Il en est de même pour le corps du texte. Une dernière pondération est calculée sur la concaténation du titre et du corps de texte. Les tweets ne contiennent pas de titre, par conséquent il n'est possible de calculer de vecteurs que d'après leur texte seul.

Jeu de données	**Langue**	**Titre**			**Corps**			**Titre + Corps**		
		Termes	Lemmes	Entités	Termes	Lemmes	Entités	Termes	Lemmes	Entités
Presse	Anglais	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Espagnol	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Allemand	✓	✓	✓	✓	✓	✓	✓	✓	✓
Tweets	Anglais				✓	✓	✓			
	Espagnol				✓	✓	✓			
	Allemand				✓	✓	✓			

TABLEAU 4.2 – Vecteurs de caractéristiques *TF-IDF* décrivant les documents au sein des corpus de presse.

Après pondération des composants de tous les documents, les articles de presse sont

décrits par neuf vecteurs *TF-IDF* et les tweets par trois (tableau 4.2). Un logiciel [Ber21b] est publié sur *Software Heritage* pour calculer les pondérations des composants de documents, en prenant en compte le fait qu'ils contiennent de la presse ou des tweets. Les corpus collectés depuis *Deutsche Welle* et Twitter ne sont pas publiables en l'état, car ce serait contrevenir aux droits des auteurs. En lieu et place sont partagés les identifiants des tweets utilisés [Ber21d], les URLs des articles téléchargés [Ber21c] ainsi que les fichiers de caractéristiques extraites [Ber22r] et présentées dans le tableau 4.1.

4.1.2 Par calcul de vecteurs agnostiques aux langues

Dans les années 2010, de nouvelles méthodes de vectorisation de texte apparaissent, basées sur des approches de plongement lexical. Il a d'abord été question de plongement de mots, de représentations distribuées ou de *word embeddings* [LM14 ; PSM14]. Ces trois appellations sont utilisées indistinctement dans la littérature. Le premier modèle proposé, `word2vec` [LM14] encode des termes d'un texte en tenant compte de leur proximité sémantique. Par exemple, « roi » et « reine » sont encodés par des vecteurs x_{roi} et x_{reine} proches dans leur espace de représentation vectoriel. Le fondement logique est que la distance euclidienne qui sépare deux vecteurs est d'autant plus faible que les termes sont sémantiquement similaires. Ce type de vecteur résout le problème de dimensionnalité propre aux représentations pondérées présentées ci-devant. La dimension des vecteurs creux est liée à la diversité du vocabulaire d'entrée. *A contrario*, les vectorisations denses encodent les textes dans des vecteurs de taille fixe, en général entre cent et mille points.

C'est dans cet esprit que plus tard, les représentations denses sont utilisées pour être comparées aux vectorisations pondérées par *TF-IDF* [Sta+19]. Il en résulte que les représentations par *TF-IDF* sont plus efficaces que des vecteurs générés par `doc2vec`, une variante du plongement lexical adapté à l'encodage de documents.

Quelques années plus tard sont conçus des modèles plus complexes basés sur des mécanismes d'attention [Vas+17] dont les transformateurs ou *Transformers* [Dev+19 ; TDP19] proposés par Google en 2018. C'est parce que ces modèles d'apprentissage profond ont outrepassé de nombreux états de l'art dans divers domaines comme la traduction ou la réponse aux questions qu'il est pertinent de considérer leur utilisation dans un processus d'encodage de documents. Les modèles d'apprentissage profond utilisés en traitement automatique des langues se basent sur des modèles linguistiques (*language models*). Ces représentations statistiques des langues sont une distribution de probabilités sur des séquences de termes de texte. Chaque modèle apprend à calculer la probabilité que le terme t succède à une séquence de termes $(t_1, \ldots, t_{m-1})$. Le réseau de neurones apprend donc la probabilité conditionnelle $P(t_m|t_1, \ldots, t_{m-1})$ d'après un corpus de texte qui lui sert de référence.

Le modèle *BERT* proposé initialement fonctionne tel un encodeur-décodeur pour respectivement vectoriser du texte ou calculer une représentation textuelle d'un vecteur. Ce dernier projette la représentation sémantique du texte dans un espace vectoriel. Cette capacité est exploitée dans des travaux récents [RG19 ; LH20 ; SMM22] dont l'objectif est d'encoder des articles de presse en vecteurs denses obtenus après affinage (*fine-tuning*) du modèle. Il génère des vecteurs alignés dans différentes langues : c'est-à-dire qu'un

vecteur représentant x_{roi} en français sera colinéaire ou presque du vecteur encodant le même concept en anglais, x_{king}. Ces vecteurs sont dits agnostiques aux langues.

Nous proposons d'utiliser ces nouveaux modèles d'encodage de documents afin d'obtenir des représentations vectorielles des articles de presse, comme cela avait pu être fait avec le mécanisme `word2vec` ou `doc2vec` [LM14]. Avec un encodage par *BERT*, le contexte est pris en compte. Une première hypothèse que l'on peut émettre est que ce type de vecteur serait moins sensible aux erreurs de reconnaissance de caractères évoquées en chapitre 3, justement car il y a prise en compte du contexte. Également, l'obtention de vecteurs alignés dans différentes langues offre une première possibilité de suivi d'événements à travers les langues. Enfin le temps de calcul nécessaire aux algorithmes présentés en section 4.2 est affecté par la dimension des vecteurs. Réduire leur taille devrait logiquement mener à une diminution des temps de traitement.

Plus précisément, c'est la tâche de similarité sémantique entre documents qui est utile pour regrouper des articles de presse rapportant des événements identiques. Elle est exprimée par des vecteurs proches dans l'espace et c'est un calcul de similarité cosinus entre les vecteurs qui détermine cette proximité. Reimers et Gurevych [RG19 ; RG20] proposent le modèle *Sentence-BERT* qui améliore les performances du modèle *BERT* pour l'encodage de phrases. Les auteurs proposent également une implémentation [RG] ainsi que des modèles préentraînés [5] et adaptés à des tâches spécifiques telles la similarité sémantique ou la réponse aux questions. Nous proposons d'utiliser ces modèles préentraînés, parce que disponibles, au lieu de réaliser les opérations d'affinage manuellement, comme cela a pu être fait précédemment [LH20]. Parmi ces modèles, quatre sont adaptés à la recherche sémantique. Ils sont également multilingues et produisent des vecteurs alignés pour des concepts sémantiques proches représentés dans des langues différentes. À l'écriture de ces lignes, les deux modèles [RG20] offrant les meilleures performances sont `distiluse-base-multilingual-cased-v1` [6] (ci-après nommé *USE*) et `paraphrase-multilingual-mpnet-base-v2` [7] (ci-après nommé *MPNet*). Ils sont entraînés avec des données collectées dans plus de quinze langues différentes. Pour de plus amples détails concernant le modèle d'apprentissage, les données utilisées ainsi que le mécanisme de transfert entre modèles sous la forme d'un enseignant et d'un étudiant (*knowledge distilation* [Abb+20]) nous renvoyons le lecteur ou la lectrice vers l'article original publié par les auteurs du concept *S-BERT* [RG20 ; Tha+21] ainsi qu'à la documentation de leur outil [RG].

Contrairement à la pondération statistique *TF-IDF*, la vectorisation dense par analyse sémantique produite par les modèles *S-BERT* traite le texte dans son ensemble. Il n'est pas question ici d'extraction de termes, de lemmes ou d'entités. Les phrases des documents sont simplement encodées séquentiellement. À terme, un vecteur de taille fixe représente chaque document. Aussi, la quantité de données traitées par le modèle est réduite, dans le cas de *BERT* à 512 termes, pour des raisons de performances et

5. `https://www.sbert.net/docs/pretrained_models.html`
(archive sur `https://web.archive.org`)

6. `https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1`

7. `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`

pour limiter le nombre d'inférences requises [LLY21]. *S-BERT* manipule des phrases, mais présente également une limitation similaire. Le tableau 4.3 donne un aperçu de ces caractéristiques.

Nom du modèle	**Données d'entraînement**	**Taille du vecteur**	**Maximum de phrases**
USE	*Multi-Lingual model of Universal Sentence Encoder for 15 languages*	512	128
MPNet	*Multi-Lingual model of Universal Sentence Encoder for 50 languages*	768	128

Tableau 4.3 – Caractéristiques des modèles denses utilisés avec *S-BERT*.

Chaque modèle est utilisé pour encoder tous les documents des jeux de données. Les documents d'*Event Registry* contenant un titre et un corps, un vecteur est calculé pour le titre, le corps de l'article et la concaténation du corps et du titre, tout comme pour la pondération *TF-IDF*. Les jeux de données au contenu court, *Event Registry Titles*, *CoAID* et *FibVid* ne sont décrits quant à eux que par un unique vecteur, issu du corps du texte. Un logiciel [Ber21a] est publié sur *Software Heritage* pour encoder des documents de presse en vecteurs denses. Les différents modèles sont également distribués à travers la plate-forme `https://huggingface.co` par les auteurs du concept *S-BERT*.

4.1.3 Synthèse

Chacun des documents issus des jeux de données que nous avons présentés en section 2.3 puis dégradés en section 3.4 est encodé en un ou plusieurs vecteurs pour être utilisé par les algorithmes de suivi de mentions d'événements. Toutes ces représentations n'ont qu'un seul but : identifier les types de vecteurs qui fournissent les meilleurs résultats et découvrir des tendances qui mèneront à recommander un type de vectorisation plutôt qu'un autre en fonction du contexte d'expérimentation et des données.

Le tableau 4.4 rappelle les vectorisations réalisées sur chacun des jeux de données. Logiquement, seul *Event Registry* n'est pas pondéré avec des modèles *TF-IDF* issus de textes de tweets. Ce ne serait pas conforme à l'affirmation selon laquelle le corpus utilisé pour la pondération doit être similaire aux données vectorisées.

Jeux de données	**Pondération TF-IDF**		**Vectorisation dense**	
	Presse	Tweets	MPNet	USE
Event Registry [Ber22i]	✓		✓	✓
Event Registry Titles [Ber22l]	✓	✓	✓	✓
CoAID [Ber22f]	✓	✓	✓	✓
FibVid [Ber22p]	✓	✓	✓	✓

Tableau 4.4 – Types de vecteurs calculés pour chacun des jeux de données évalués.

À la section 3.4 (page 87), nous avions généré douze jeux de données différents,

après application ou non des différentes dégradations, par *OCR* ou sursegmentation. Pour chacun, ce sont désormais quatre vectorisations qui sont disponibles, dans chaque langue que nous avons répertoriée au tableau 2.10 (page 66). Après avoir vectorisé les données des textes, c'est-à-dire rendu possible leur utilisation par les algorithmes de suivi d'événements mentionnés précédemment, nous allons nous intéresser en détail à deux d'entre eux. Ils répondent à ces objectifs différents, mais tous deux exigent des données encodées pour fonctionner.

4.2 Algorithmes pour la construction d'histoire de presse

Divers types d'algorithmes permettent le suivi des mentions d'événements rapportées dans la presse. Parmi les différentes méthodes présentées en état de l'art (section 2.2), nous en avons retenu deux, adaptées à des cas d'usage courants. La première est un algorithme supervisé, que l'on nommera ici « Miranda et coll. » en référence à ses concepteurs [Mir+18]. L'autre est une adaptation de l'algorithme des K-Moyennes (*K-Means* [Mac67]) adapté au suivi de mentions d'événements. Les deux permettent un suivi des événements rapportés dans la presse, selon les critères adoptés dans la littérature : il s'agit de construire une chronologie de documents rapportant des événements identiques.

Leur fonctionnement est similaire : ils groupent des articles au sein de *clusters* de documents. Les articles de chaque groupe sont supposés ne traiter qu'un unique sujet. L'extraction des caractéristiques de chaque groupe (date, lieu, titre, entités participantes) offre la possibilité de reconstruire l'événement en répondant aux questions de base : *quoi*, *quand*, *où* l'action se déroule-telle, et quels acteurs (*qui*) met-elle en jeu.

4.2.1 Suivi supervisé d'événements

Dans cette section, nous produisons une explication détaillée du fonctionnement de l'algorithme proposé par Miranda et coll. [Mir+18] afin d'en comprendre les rouages et les défauts qui pourraient affecter la compréhension des résultats sur des documents historiques. Il s'agit d'un algorithme supervisé qui nécessite une phase d'entraînement préalable à partir de données annotées. L'entraînement de l'algorithme est évoqué ci-après.

Il traite les documents sous la forme d'un flux continu de données. Cette notion de flux lui permet de s'intégrer aisément avec des outils d'agrégations d'articles tels *NewsFeed* [TN12] qui collecte et unifie des flux d'articles provenant de sources variées. Les articles sont traités itérativement, l'un après l'autre et incorporés au sein de groupes de documents (*clusters*). Chaque nouveau document est comparé à chacun des groupes d'articles parmi l'ensemble de ceux connus. Chaque groupe est un candidat potentiel au sein duquel le document peut être intégré. La condition à remplir est que la similarité du document avec ce groupe soit supérieure à un seuil T_1. Si de multiples candidats existent, celui avec la similarité la plus grande remporte la sélection. Au contraire, en l'absence de candidat positif (où la similarité est supérieure à T_1), un nouveau groupe de documents est créé et initialisé par le document en cours de traitement.

Le calcul de similarité entre un document d_i et un groupe (*cluster*) C_j est défini par l'opération $sim(d_i, C_j)$ détaillée par l'équation équation 4.6.

L'algorithme traite des données hétérogènes : du texte et des horodatages de documents. La question de la temporalité est importante pour le traitement de documents de presse. Un événement est en effet ancré dans une unité temporelle donnée. Dans les temps qui suivent l'événement, des articles qui le décrivent de façon détaillée sont publiés, puis dans un second temps émergent des éditoriaux, des documentaires ou des rétrospectives [Lej+15]. Bien qu'ils traitent et mentionnent des événements identiques, ce sont deux types de contenu bien distincts, séparés entre autres par le moment de leur publication.

Nous avons vu en état de l'art (chapitre 2, page 40) que le temps joue un rôle majeur dans la définition des événements (définition n° 10). Comparer les événements implique de comparer les moments dans lesquels ils sont survenus. Pour cela, nous disposons de l'horodatage des articles de presse. Les groupes de documents gardent la trace de deux dates : une basse correspondant à la date de publication du plus ancien document du groupe et une haute correspondant à la date de publication la plus récente. Pour calculer les similarités, une fonction gaussienne (ϕ) est utilisée avec les paramètres $\mu = 0$ et $\sigma = 3$. Ce dernier paramètre doit être interprété comme le nombre de jours à partir duquel la similarité décroît significativement. On considère ainsi que les documents publiés plus avant ou après sont trop éloignés dans le temps pour être similaires. La valeur du paramètre σ doit par conséquent être adaptée à chaque contexte d'étude. Dans le cas où les documents diffusent rapidement, elle peut être diminuée, pour des mentions dans des articles circulant lentement, cette valeur doit être augmentée. Est conservée ici la valeur par défaut proposée par Miranda et coll. La date de publication de chaque article est comparée aux bornes basses et hautes de chacun des groupes. En résultent deux valeurs de similarité temporelle. Ce calcul est représenté par l'équation 4.4. La fonction gaussienne initialisée avec ces paramètres est représentée en figure 4.1. On y remarque que les similarités les plus élevées sont centrées autour de zéro, là où la distance qui sépare les deux dates est la plus faible.

$$f(d_{date}, C_{date}) = \phi_{\mu,\sigma^2}(|d_{date} - C_{date}|) \tag{4.4}$$

Le calcul des similarités entre textes intervient dans un second temps. Les documents sont encodés sous forme de vecteurs et la mesure de similarité cosinus ($\theta(d^k, C^k)$) calcule la distance qui les sépare. Ce calcul est décrit dans l'équation 4.5. Le vecteur de nombres réels encodant un groupe de documents est formé par moyennage de l'ensemble des vecteurs des documents qu'il contient. Si de multiples vecteurs encodant le texte (un pour les termes, un pour les lemmes, etc.) sont disponibles, ils sont comparés paire à paire entre le document et le *cluster*. Le nombre de ces vecteurs est noté K. Pour les représentations par pondération *TF-IDF* et les représentations denses, ces valeurs diffèrent. Pour les documents de presse encodés par pondération *TF-IDF*, $K = 9$ et $K = 3$ pour les données provenant de Twitter, comme expliqué dans le tableau 4.2. Dans l'équation 4.5, les coefficients β_k pondèrent l'importance des différents vecteurs dans le calcul de similarité. Ils donnent par exemple plus de poids aux vecteurs représentant

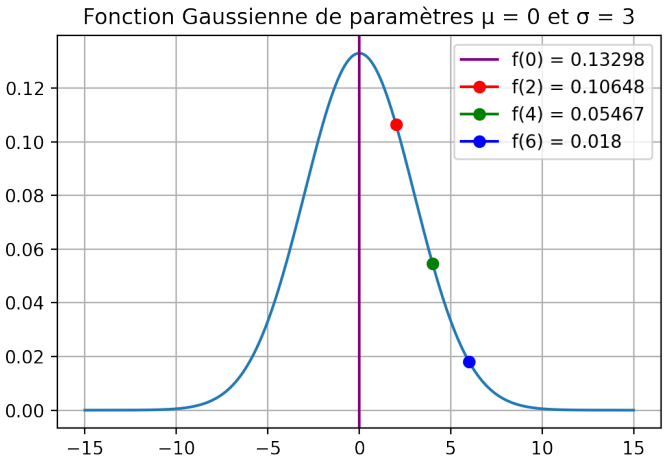


FIGURE 4.1 – Fonction gaussienne ϕ utilisée pour la comparaison des dates entre un document et un groupe. Ici, $\mu = 0$ et $\sigma = 3$.

les entités nommées dans le corps du texte qu'à ceux encodant les termes du titre. Les valeurs β_k et α sont obtenues par entraînement d'un modèle comme *SVM* [EMN92] ou la régression logistique, utilisés tous deux pour des tâches de classification binaire.

$$g(d_i, C_j) = \sum_{k=0}^{K} \beta_k \times \theta(d_i^k, C_j^k) + \sum_{k=0}^{K=2} \beta_k \times f(d_i^k, C_j^k) + \alpha \tag{4.5}$$

Pour contraindre les scores de similarités dans l'intervalle $[0; 1]$, le résultat de la fonction $g(d_i, C_j)$ est injecté dans une fonction sigmoïde, de la même façon que pour une régression logistique. Le calcul final de similarité est donné par l'équation 4.6.

$$sim(d_i, C_j) = \frac{1}{1 + \mathrm{e}^{-g(d_i, C_j)}} \tag{4.6}$$

Les coefficients du modèle β_k ainsi que le seuil T_1 sont déterminés par apprentissage. Dans leur publication originale [Mir+18], les auteurs utilisent l'algorithme *SVM-Rank* [Joa02]. Une implémentation [Joa09] est publique, mais face aux difficultés posées par son utilisation (intégration dans un processus complet, reproductibilité des expérimentations), le choix est fait d'utiliser de façon analogue a ce qui est proposé par Linger et coll. [LH20] une régression logistique. Cet algorithme fournit des coefficients β_k pour chaque caractéristique (ici, chacun des vecteurs encodant des parties du document) et une ordonnée à l'origine α. Il répond aux mêmes problématiques et s'inscrit dans le même contexte que l'algorithme *SVM-Rank*.

Entraînement du modèle

Le modèle à entraîner est un modèle de classification binaire, chargé de fournir une réponse à la question « *le document d_i doit-il, d'après sa similarité avec le groupe C_j, y être inclus ?* ». Si la réponse est oui, cela signifie que la similarité entre les deux est supérieure à T_1. Le modèle de régression logistique est un modèle mathématique statistique répondant à ce problème. À partir d'un ensemble de similarités, il est entraîné et les coefficients du modèle ajustés pour séparer au mieux les réponses positives des négatives. Pour déterminer le seuil d'inclusion T_1, les scores sont calculés avec le modèle le plus adapté aux données.

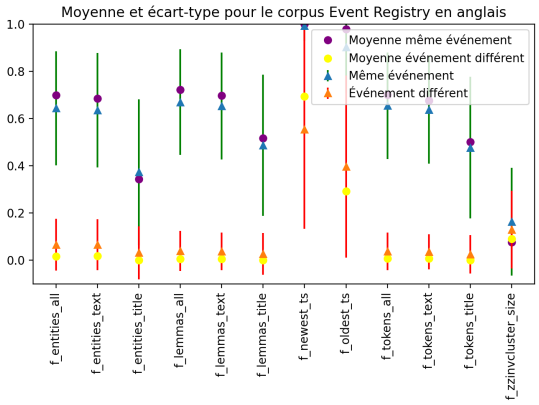
Dans un premier temps, à partir de données d'entraînement, les groupes de documents sont construits itérativement d'après les étiquettes connues dans la vérité terrain. Chaque itération entraîne une série de calculs de similarités entre le document d_i et chaque groupe C_j candidat. Les similarités entre les paires de vecteurs de document et de *clusters* sont toutes conservées. Pour une représentation vectorielle par *TF-IDF* à neuf vecteurs, $K = 9$ dans l'équation 4.6 (comme présenté en sous-section 4.1.1), ce sont neuf mesures de similarité qui sont calculées. Par exemple, les moyennes de ces calculs de similarité pour chacune des caractéristiques du corpus *Event Registry* sont présentées en figure 4.2a. En rouge, les similarités calculées pour des paires $d_i - C_j$ qui ne mentionnent pas le même événement, et en vert les similarités pour les paires qui décrivent les mêmes événements.

Le nombre d'exemples négatifs est très fortement déséquilibré ($> 99{,}5$ %) par rapport aux positifs. Sur l'ensemble des calculs réalisés, à un unique document d_i correspond un unique *cluster* C_j. Ce déséquilibre est réduit par la sélection des vingt exemples négatifs de plus hauts scores pour chaque comparaison. Ce sont ceux les plus près de la frontière de décision du modèle. Il en résulte une base d'exemples à environ 95 % d'éléments négatifs et 5 % de positifs.

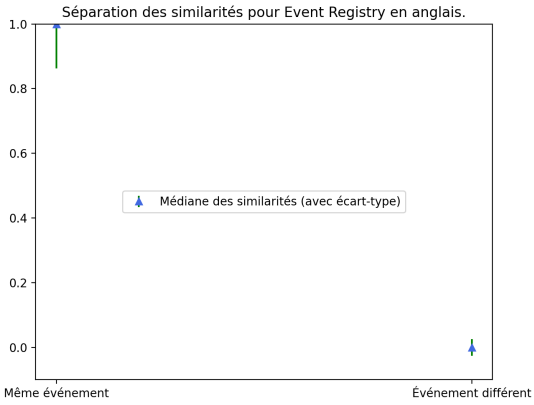
Un modèle de régression logistique nourri avec ces données est recherché sur une grille de paramètres. Le meilleur modèle est celui dont la moyenne harmonique par étiquette (négative ou positive) est la plus élevée. Une moyenne harmonique de type *macro* élimine le défaut d'équilibre dans les données. Du modèle trouvé sont extraits les coefficients β_k et l'ordonnée à l'origine α.

Les scores finaux de similarités entre les documents d'entraînement d_i et les *clusters* C_j sont calculés par l'équation 4.6. En résulte un unique nombre réel, la similarité entre le document est le groupe. Dans la figure 4.2b, on constate que les scores pour des paires de documents – groupe différents sont proches de zéro, proches d'un sinon. Enfin, le seuil T_1 est déterminé depuis ces scores, en cherchant à maximiser la moyenne harmonique sur les données de développement. Un exemple est présenté en figure 4.3. Pour chaque valeur de seuil T_1 possible, la moyenne harmonique est calculée. Le seuil T_1 correspond au point d'abscisse qui maximise cette moyenne.

Nous proposons notre propre implémentation de cet algorithme et de sa procédure d'entraînement [Ber21f]. Pour valider cette implémentation, nous comparons les résultats qu'elle fournit avec ceux partagés par les auteurs originaux [Mir+18]. Ils utilisent le jeu de données *Event Registry* pour lequel ils fournissent des vectorisations par pondération



(a) Similarités cosinus de chaque vecteur de couple d_i - C_i rapportant des événements identiques ou non.



(b) Scores de similarités pour les couples d_i - C_i rapportant des événements identiques ou non.

FIGURE 4.2 – Scores de similarités durant le processus d'entraînement de l'algorithme, avant et après le calcul de l'équation 4.5 et application de la fonction sigmoïde

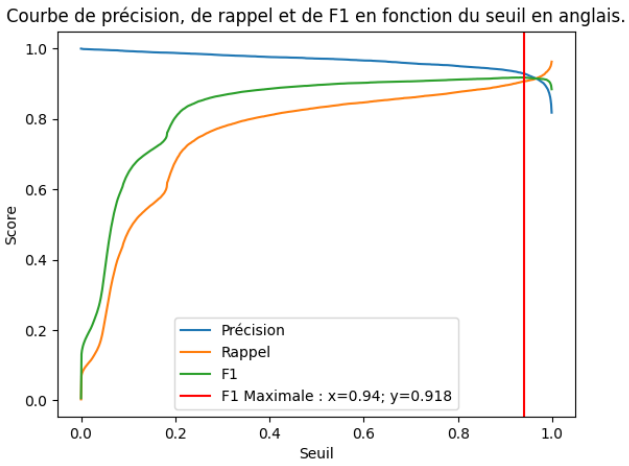


FIGURE 4.3 – Courbe de précision, de rappel et de moyenne harmonique des scores pour sélectionner le seuil T_1.

TF-IDF et créent des modèles pour les trois langues. Nous réutilisons ces vectorisations dans les mêmes conditions pour valider nos outils. Leurs résultats et les nôtres sont répertoriés dans le tableau 4.5 en testant sur les mêmes données, mais avec nos propres modèles. Les résultats sont similaires et *a minima*, la différence entre leurs résultats et ceux que nous obtenons est faible. Par conséquent, nous considérons que notre implémentation est de qualité et c'est ce processus d'apprentissage qui est utilisé pour les expérimentations réalisées dans ce chapitre. La différence sur les résultats en anglais

	Langue	Résultats standards			**Résultats *BCubed***			**Nombre de groupes**	
		F1	Précision	Rappel	F1	Précision	Rappel	Prédits	Réels
Miranda et coll.	Anglais	94,03	98,14	90,25	92,36	94,57	90,25	326	222
	Allemand	97,19	99,86	94,67	93,64	98,92	88,90	229	118
	Espagnol	96,83	97,01	96,65	91,61	96,44	87,25	281	149
Notre modèle	Anglais	90,70	97,60	84,70	92,50	95,60	89,70	392	222
	Allemand	95,20	99,90	91,00	93,7	99,00	88,90	203	118
	Espagnol	96,70	97,20	96,20	91,20	96,40	86,50	262	149
Différence	Anglais	3,33	0,54	5,55	- 0,14	- 1,03	0,55	-66	222
	Allemand	1,99	- 0,04	3,67	- 0,06	- 0,08	0	26	118
	Espagnol	0,13	- 0,19	0,45	0,41	0,04	0,75	19	149

TABLEAU 4.5 – Comparatif des résultats obtenus sur le jeu de données *Event Registry* en utilisant les vecteurs *TF-IDF* fournis par Miranda et coll. [Mir+18].

s'explique par les avancées publiées par les mêmes auteurs quelques années plus tard [SMM22]. Nous utilisons une mesure pour nos résultats nommée *BCubed* [Ami+07]. Nous y reviendrons plus tard lorsque nous évoquerons l'évaluation de nos résultats, à la page 112.

4.2.2 Suivi non supervisé d'événements

L'autre mécanisme envisagé pour le suivi des mentions d'événements dans du texte est un algorithme non supervisé. Nous l'avons déjà démontré, les jeux de données annotés et adaptés à notre problématique sont rares, ce qui limite le recours possible à des algorithmes exigeant des données d'entraînement.

Dans la continuité des suggestions de plusieurs auteurs précédents [RM12 ; MBC19 ; Sta+19], l'algorithme non supervisé qui peut être envisagé est celui des K-Moyennes [Mac67] en utilisant la similarité cosinus comme mesure de distance entre les vecteurs de documents. Cet algorithme répond parfaitement à la problématique du manque de données d'entraînement spécifique à ce domaine de recherche. Tout comme pour l'algorithme supervisé présenté précédemment, deux types de données sont pris en compte pour identifier des groupes de documents : le temps qui s'écoule ainsi que le texte.

La prise en compte du temps se fait au sein de fenêtres temporelles. Les documents sont ordonnés chronologiquement et divisés en groupes de n jours consécutifs que l'on nomme fenêtres. Tous les documents publiés dans une fenêtre temporelle sont traités ensemble, l'une après l'autre, limitant le volume de données manipulées à chaque itération. Cette stratégie repose sur l'hypothèse que des articles publiés au sein de la même unité de temps relatent potentiellement les mêmes événements [SSS02]. Dans le cadre du développement de l'algorithme de suivi de mentions d'événements *newsLens*, il est déterminé qu'une fenêtre glissante d'une semaine avec 50 % de recouvrement est la plus adaptée. C'est-à-dire que les fenêtres sont par exemple j à $j+7$, $j+3$ à $j+10$, $j+6$ à $j+13$, etc.

Pour le texte, les articles sont encodés en une multitude de vecteurs qui représentent le titre ou le texte ainsi que dans le cas de pondérations *TF-IDF*, les termes, lemmes et entités de ces titres ou corps d'articles. L'algorithme de K-Moyennes n'accepte dans

sa forme originale qu'un seul vecteur pour encoder les données associées à un article. Considérons un document présenté comme dans le tableau 4.6 : les deux vecteurs encodant le titre et le corps du texte côtoient l'horodatage numérique du document. Ces données sont concaténées pour former un unique vecteur de caractéristiques et encoder le document pour l'algorithme de K-Moyennes.

Horodatage	**Vecteur du titre**	**Vecteur du corps**
1448454804	[2 ; 5,6 ; 8,4 : 23,9 ; . . .]	[7 ; 4,6 ; 8,6 ; 12,6 ; . . .]

TABLEAU 4.6 – Exemple de vecteurs et de données décrivant un article de presse, dont les parties sont encodées numériquement.

La problématique principale posée par l'algorithme de K-Moyennes est qu'il doit être initialisé avec un nombre de groupes de documents (*clusters*) k qu'il utilise pour partitionner les données. Trois métriques peuvent être utilisées pour déterminer un nombre de k groupes qui maximise leur cohérence, c'est-à-dire le fait que les documents des groupes sont similaires entre eux, et que les groupes quant à eux sont dissemblables. Chacune nécessite le calcul de tous les partitionnements possibles tels que $k \in [a + s, a + 2s, a + 3s, \ldots, b]$, a étant le nombre minimal de groupes possibles, b le nombre maximal. Dans le pire des cas, $a = 1$, $b = N$, N étant le nombre total de documents, le pas s valant 1. Pour la première méthode, dite « du coude », la somme totale des carrés intragroupe (*total within-cluster sum of squares, WSS*) pour chaque k est projetée sur un graphe à la recherche d'un « coude ». La valeur k où se situe le coude indique le nombre de groupes potentiellement proche de la réalité. Une autre méthode repose sur le coefficient de silhouette [Rou87] qui mesure la compacité d'un *cluster*, la plus forte étant la meilleure, au maximum d'un. Une troisième, la statistique *gap* [TWH06] compare la modification de dispersion intra*cluster* avec une distribution de référence calculée au préalable. Les deux premières sont implémentées au sein du paquet logiciel `scikit-learn` [Cou] que nous utilisons. Comme le montre la figure 4.4, le calcul du coefficient de silhouette donne une approximation toujours meilleure du nombre correct de *clusters* par rapport à la réalité. La projection est ici faite avec le jeu de données *Event Registry*, mais le comportement est similaire avec les deux autres corpus.

Par conséquent, c'est la méthode basée sur le calcul du coefficient de silhouette qui est utilisée pour déterminer le nombre de groupes k. La méthode du coude donne des résultats trop éloignés de la réalité des données pour être utilisable. Cette recherche de la valeur k implique des calculs de *clustering* pour chaque valeur possible de l'intervalle. Cette méthode sera par conséquent plus lente comparé à celle, supervisée, introduite précédemment.

4.2.3 Synthèse

Chacun des deux algorithmes, Miranda et coll. puis K-Moyennes répondent à leur propre problématique : un algorithme supervisé est utilisé en présence de données d'entraînement pour construire un modèle de décision. En l'absence de ces données, seul un

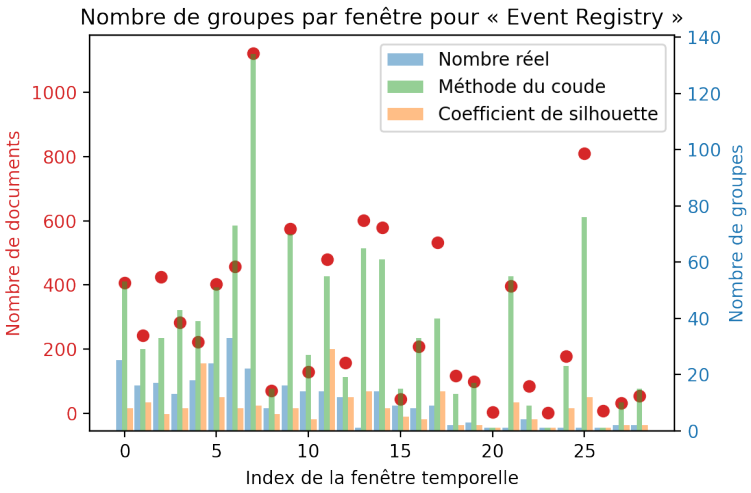


FIGURE 4.4 – Nombre de *clusters* par fenêtre temporelle découvert par la méthode du « coude » ou du coefficient de silhouette par rapport à la réalité.

algorithme non supervisé permet de suivre les mentions d'événements dans les articles de presse. Dans ce chapitre, les deux algorithmes sont utilisés dans le même contexte, avec les mêmes données afin de déterminer dans quelles situations il est possible d'envisager l'un plutôt que l'autre et de comparer leurs performances respectives sur les jeux de données synthétisés. Ils prennent en compte la question du temps qui s'écoule de différentes manières, par un calcul de similarité entre dates ou en opérant une sélection d'après des fenêtres temporelles. Tous deux utilisent des représentations vectorielles de documents, encodées par les mécanismes présentés en section 4.1.

Une implémentation de ces deux algorithmes est distribuée [Ber21f] dans le cadre de cette thèse, avec une interface en ligne de commande [Ber21h] pour lancer les entraînements des modèles et les tests sur des corpus de données. Les expérimentations réalisées dans ce chapitre se basent sur ces implémentations.

4.3 Reconstruction de la chronologie des événements

Les expériences décrites dans ce chapitre se basent sur les jeux de données présentés en section 2.3, artificiellement dégradés en section 3.4. Tous les articles des huit jeux de données sont encodés à l'aide de deux méthodes différentes, la pondération *TF-IDF* et l'encodage dense par des modèles d'apprentissage profond. Enfin, pour chacun de ces encodages, les deux algorithmes présentés ici sont utilisés pour grouper les articles qui mentionnent des événements identiques. L'objectif de toutes ces expérimentations est d'identifier les encodages et les algorithmes de suivi les plus performants pour notre problématique. Il s'agit également de quantifier l'impact des dégradations issues de l'*OCR* et de la sursegmentation du contenu. Le processus expérimental complet est schématisé

en figure 4.5.

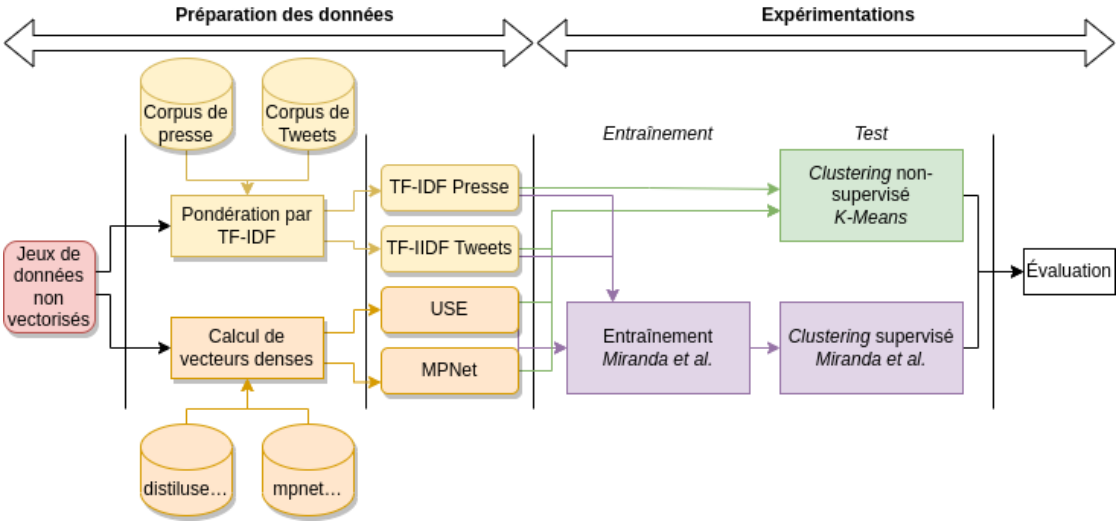


FIGURE 4.5 – Description du processus de traitement global, depuis les documents jusqu'à l'évaluation de la qualité du suivi des mentions d'événements.

Les résultats de toutes les expériences, de même que l'ensemble des modèles entraînés pour l'implémentation de l'algorithme supervisé sont publiés conjointement à cette thèse [Ber22q]. Ce sont ces résultats qui vont être présentés et analysés dans les sections suivantes. Dans ce chapitre, nous montrons les différences qui existent entre les algorithmes qui manipulent des articles de presse encodés eux aussi par des procédés divers. Ce sont ces différences qui sont mises en avant et analysées ci-après.

Deux types d'expérimentations sont proposés, un premier sur les articles de presse complets, c'est-à-dire ceux du corpus *Event Registry* et un second sur les brèves de presse et les tweets en utilisant les jeux de données *Event Registry Titles*, *CoAID* et *FibVid*.

La qualité des regroupements de documents fournis par les algorithmes est évaluée selon deux types de métriques. La première est la mesure standard de précision, de rappel et leur moyenne harmonique (*F1*) basée sur des matrices de confusion à paires qui comparent le *clustering* avec le résultat attendu. La seconde méthode repose sur les mesures *BCubed* [Ami+07] qui prennent en compte des contraintes formelles telles l'homogénéité des *clusters*, leur complétude, leur taille, etc. Dans les diagrammes présentés et analysés dans cette section, seule la *F1* de la métrique *BCubed* est présentée pour des raisons de clarté. Les analyses sont identiques, quelle que soit celle utilisée.

Pour chaque évaluation, un temps indicatif est annoncé. Les expériences ont été exécutées sur un ordinateur doté d'un processeur *Intel® Xeon® CPU E5-2630 v2 @ 2.60 GHz* avec 64 Go de mémoire vive. La durée indiquée est le temps processeur consommé par le processus plutôt que le temps physique écoulé entre le début et la fin des expériences. Cette information peut rentrer en ligne de compte pour évaluer le choix d'un processus expérimental par rapport à un cas d'utilisation donné. Certains

paramètres des algorithmes sont initialisés par des générateurs de nombres aléatoires. Pour satisfaire des conditions de reproductibilité, ces nombres sont initialisés par une valeur arbitraire, assurant ainsi des résultats déterministes.

4.3.1 Propagation des mentions d'événements dans la presse

Dans un premier temps, nous analysons les jeux de données *Event Registry* comportant ou non des dégradations d'*OCR* et différents niveaux de segmentation : d'abord sans puis avec les articles segmentés en deux et enfin en trois. L'effet des segmentations est évalué séparément.

Event Registry

Les résultats bruts des expérimentations sont présentés dans les tableaux 4.7 pour l'algorithme supervisé et dans le 4.8 pour le non supervisé. L'analyse des dégradations et de l'impact des vectorisations est présentée en figure 4.6.

Données	Vecteurs	Langue	Résultats standards			Résultats *BCubed*			Nombre de groupes		Durée
			F1	P	R	F1	P	R	Prédits	Réels	
Non dégradé	Presse	Anglais	89,80	97,30	83,40	91,20	95,40	87,30	453	222	00 :56 :33
	USE		66,20	96,50	50,30	81,50	93,10	72,50	428		00 :26 :27
	MPNet		66,10	97,30	50,00	82,40	94,90	72,90	451		00 :24 :46
	Presse	Espagnol	96,40	96,80	96,00	91,20	97,40	85,70	283	149	00 :04 :22
	USE		80,00	96,60	68,30	86,70	93,90	80,50	234		00 :02 :37
	MPNet		69,10	93,40	54,90	83,40	93,60	75,20	253		00 :04 :23
	Presse	Allemand	91,90	99,40	85,40	90,50	98,00	84,10	207	118	00 :04 :12
	USE		75,90	99,10	61,50	85,30	96,20	76,50	182		00 :04 :18
	MPNet		75,80	99,40	61,20	85,50	96,40	76,80	178		00 :03 :40
Dégradé	Presse	Anglais	92,20	97,70	87,20	91,70	96,10	87,80	487	222	01 :04 :35
	USE		62,80	95,60	46,80	80,20	91,80	71,20	437		00 :26 :11
	MPNet		65,20	96,70	49,10	81,40	93,90	71,90	446		00 :25 :49
	Presse	Espagnol	96,00	96,70	95,40	90,50	97,10	84,70	291	149	00 :04 :36
	USE		78,70	97,20	66,10	85,90	94,70	78,60	247		00 :04 :25
	MPNet		69,80	91,90	56,20	82,00	92,00	74,00	256		00 :04 :28
	Presse	Allemand	91,30	99,40	84,50	91,50	97,80	85,90	213	118	00 :06 :06
	USE		81,20	99,20	68,80	85,70	95,40	77,80	182		00 :04 :10
	MPNet		83,30	99,00	72,00	84,70	93,90	77,20	173		00 :03 :43

TABLEAU 4.7 – Résultats des expérimentations sur le jeu de données *Event Registry* en appliquant l'algorithme supervisé.

Les figures 4.6a et 4.6c montrent tout d'abord que les dégradations du corpus par *OCR* entraînent une diminution de la qualité des regroupements par les algorithmes. La baisse est plus faible avec l'algorithme supervisé, seuls deux résultats sont diminués de plus d'un point, en anglais avec les vecteurs *USE* et en espagnol avec les vecteurs *MPNet*. Avec l'algorithme non supervisé, la baisse de qualité des *clusters* est plus importante en général, mais aussi plus constante : seule une classification sur les neuf n'est pas diminuée, mais au contraire augmentée. Cela signifie que le modèle a mieux réussi à regrouper les documents dégradés que les documents non dégradés. Ce phénomène se retrouve sur les résultats de l'algorithme supervisé, avec les vecteurs *TF-IDF* pondérés sur le corpus de

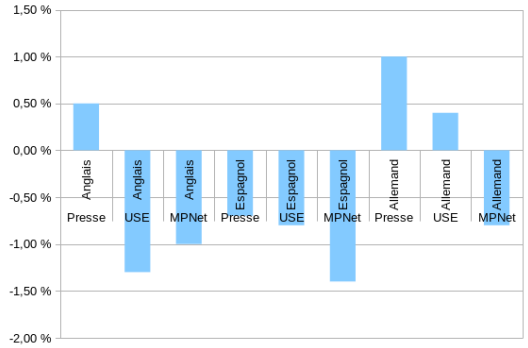
Données	Vecteurs	Langue	Résultats standards			Résultats *BCubed*			Nombre de groupes		Durée
			F1	P	R	F1	P	R	Prédits	Réels	
Non dégradé	Presse	Anglais	63,80	79,10	53,40	77,20	81,00	73,70	212	222	15 :49 :18
	USE		73,50	86,70	63,90	80,90	84,30	77,80	186		16 :49 :45
	MPNet		73,70	85,70	64,60	80,60	83,50	77,90	181		03 :01 :50
	Presse	Espagnol	62,40	70,60	55,80	75,20	77,10	73,40	141	149	02 :04 :07
	USE		75,80	86,40	67,50	79,80	83,70	76,20	173		08 :42 :15
	MPNet		74,20	85,00	65,90	78,80	81,30	76,40	153		09 :42 :08
	Presse	Allemand	65,10	94,20	49,70	75,10	87,70	65,60	141	118	03 :53 :16
	USE		68,70	98,50	52,70	81,50	92,60	72,70	141		08 :57 :58
	MPNet		69,20	97,60	53,60	80,10	91,10	71,50	145		09 :48 :15
Dégradé	Presse	Anglais	61,60	73,70	52,90	76,20	79,10	73,50	196	222	20 :09 :06
	USE		73,30	86,00	63,90	80,50	83,10	78,10	179		15 :16 :00
	MPNet		70,40	84,20	60,40	79,40	82,80	76,20	190		02 :51 :26
	Presse	Espagnol	69,60	74,90	65,10	74,70	73,70	75,80	121	149	02 :04 :43
	USE		77,00	89,30	67,70	80,20	83,60	77,10	162		07 :38 :51
	MPNet		74,20	85,20	65,70	77,60	79,50	75,80	147		09 :51 :11
	Presse	Allemand	72,40	92,90	59,30	79,50	84,90	74,70	114	118	03 :54 :56
	USE		69,30	97,90	53,60	81,10	91,10	73,10	133		11 :07 :53
	MPNet		68,00	83,10	57,60	77,40	80,30	74,80	105		09 :58 :18

TABLEAU 4.8 – Résultats des expérimentations sur le jeu de données *Event Registry* en appliquant l'algorithme non supervisé.

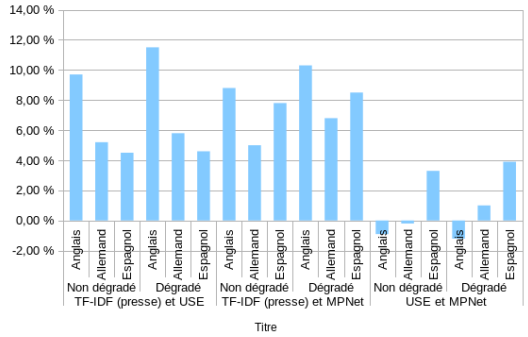
presse en anglais et en allemand. La tendance est en général à la baisse dans tous les cas, la dégradation par *OCR* diminue de peu l'efficacité des algorithmes (-0,46 point en moyenne pour l'algorithme supervisé, -0,39 point pour le non supervisé, moyenne affectée par l'amélioration significative en langue allemande avec les vecteurs *TF-IDF*).

Les figures 4.6b et 4.6d quant à elles comparent les différents encodages utilisés deux à deux. Premièrement, avec l'algorithme supervisé, les pondérations *TF-IDF* sont systématiquement meilleures que les vectorisations denses, que les données soient dégradées ou non. Comparés aux vecteurs denses, les vecteurs *TF-IDF* offrent même de meilleures performances sur les données dégradées. Pour les vecteurs denses, qu'ils soient encodés par *MPNet* ou *USE*, ils donnent des résultats assez similaires entre eux, qu'importe la langue.

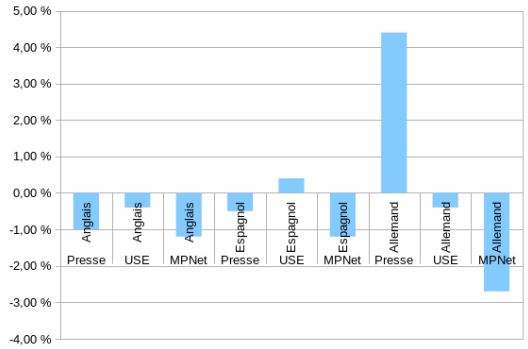
Que l'algorithme soit supervisé ou non, une sortie d'*OCR* qui contient des erreurs entraîne une diminution, même faible, d'efficacité des algorithmes. Les vecteurs par pondération *TF-IDF* sont meilleurs avec l'algorithme supervisé et ce sont les vecteurs denses qui fonctionnent mieux, de peu, avec l'algorithme de K-Moyennes. À ce stade, seules les dégradations introduites par l'application de l'*OCR* peuvent expliquer des diminutions dans les résultats, par rapport à une version non dégradée des textes. Une récente étude a mis en évidence l'incapacité pour les *OCR* de reconnaître correctement les entités nommées [HHD20] pour réaliser de la post-correction de texte. Les algorithmes présentés ici bénéficient de la détection des entités nommées, comme nous l'avons expliqué en section 4.1 où les documents sont représentés après avoir extrait leurs entités. Un faible taux de reconnaissance des entités entraîne nécessairement une diminution de la qualité de représentation du document. Les deux algorithmes testés utilisent ces vectorisations : c'est une voie possible d'explication de nos résultats. L'utilisation de vecteurs denses



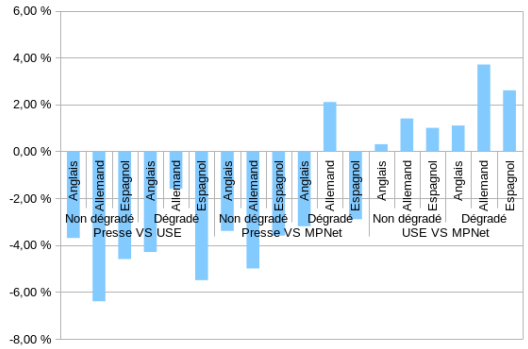
(a) Exécution sur articles dégradés contre articles non dégradés (algo. supervisé).



(b) Qualité de regroupement en fonction de la dégradation et de l'encodage (algo. supervisé).



(c) Exécution sur articles dégradés contre articles non dégradés (algo. non supervisé).



(d) Qualité de regroupement en fonction de la dégradation et de l'encodage (algo. non supervisé).

FIGURE 4.6 – Qualité des résultats de regroupement selon les dégradations ou le type de vecteur par langue du corpus *Event Registry* dégradé ou non.

(*USE* et *MPNet*) ne semble pas apporter de nouveauté par rapport aux pondérations *TF-IDF*. En effet, les niveaux de dégradations semblent identiques. Puisque ces vectorisations encapsulent la sémantique de la phrase, l'introduction d'erreurs par l'*OCR* a pu affecter également le sens des termes et des phrases, empêchant les algorithmes d'identifier correctement deux documents évoquant les mêmes événements.

Event Registry, segmenté en deux

Les résultats bruts des expérimentations sont présentés dans les tableaux 4.9 pour l'algorithme supervisé et dans le 4.10 pour le non supervisé. L'analyse des dégradations et de l'impact des vectorisations est présentée en figure 4.7.

Le premier niveau de segmentation est une division des articles en deux. Ici, sur les figures 4.7a et 4.7c, la diminution des résultats est accentuée par la segmentation, comparée aux seules erreurs introduites par l'*OCR*. Avec les mêmes dégradations d'*OCR*,

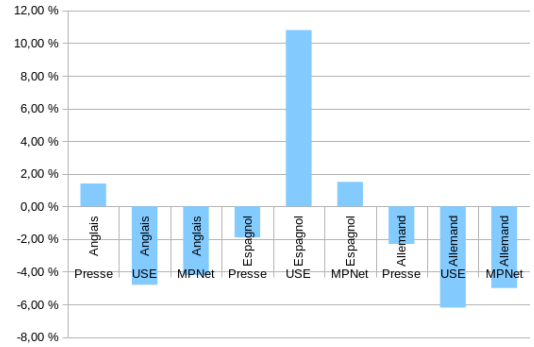
Données	Vecteurs	Langue	Résultats standards			Résultats *BCubed*			Nombre de groupes		Durée
			F1	P	R	F1	P	R	Prédits	Réels	
Non dégradé	Presse	Anglais	67,20	96,50	51,50	82,00	95,50	71,90	735	222	01 :28 :50
	USE		46,70	94,60	31,00	68,90	91,20	55,40	667		00 :35 :46
	MPNet		47,20	95,50	31,30	70,20	92,50	56,50	565		00 :31 :32
	Presse	Espagnol	92,00	96,70	87,80	87,40	96,10	80,20	322	149	00 :07 :46
	USE		49,90	95,50	33,70	67,70	95,80	52,40	420		00 :06 :27
	MPNet		69,60	92,10	56,00	76,10	91,00	65,40	285		00 :05 :11
	Presse	Allemand	69,50	99,70	53,40	79,80	97,60	67,40	255	118	00 :07 :46
	USE		56,60	98,40	39,70	73,90	94,70	60,50	227		00 :04 :53
	MPNet		49,00	97,60	32,70	72,80	94,80	59,00	198		00 :04 :32
Dégradé	Presse	Anglais	71,60	96,80	56,80	83,40	95,80	73,90	732	222	01 :39 :28
	USE		40,90	93,00	26,20	64,10	89,30	49,90	576		00 :31 :23
	MPNet		40,10	93,10	25,60	66,00	90,80	51,80	573		00 :31 :20
	Presse	Espagnol	92,70	96,10	89,50	85,50	95,10	77,60	330	149	00 :08 :16
	USE		73,90	90,60	62,40	78,50	88,20	70,70	270		00 :04 :56
	MPNet		71,30	89,30	59,40	77,60	86,90	70,10	241		00 :04 :36
	Presse	Allemand	67,90	99,60	51,50	77,50	96,70	64,70	263	118	00 :07 :46
	USE		41,40	97,50	26,30	67,70	93,70	52,90	264		00 :05 :08
	MPNet		42,70	95,90	27,40	67,80	91,50	53,90	201		00 :04 :27

TABLEAU 4.9 – Résultats des expérimentations sur le jeu de données *Event Registry* avec documents segmentés en deux, en appliquant l'algorithme supervisé.

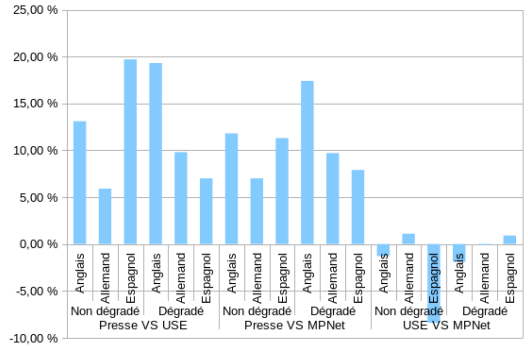
Données	Vecteurs	Langue	Résultats standards			Résultats *BCubed*			Nombre de groupes		Durée
			F1	P	R	F1	P	R	Prédits	Réels	
Non dégradé	Presse	Anglais	67,50	73,40	62,50	71,70	77,80	66,50	206	222	09 :05 :51
	USE		67,50	79,10	58,80	67,70	79,20	58,10	232		19 :45 :47
	MPNet		56,40	63,30	50,90	66,30	74,50	59,70	201		05 :59 :22
	Presse	Espagnol	46,30	60,90	37,40	66,80	75,00	60,30	190	149	06 :04 :45
	USE		59,70	68,60	52,90	62,30	69,50	56,50	163		08 :09 :23
	MPNet		59,30	69,00	51,90	60,70	70,10	53,50	183		09 :26 :47
	Presse	Allemand	64,30	94,20	48,80	62,90	86,90	49,30	211	118	08 :20 :25
	USE		64,90	84,50	52,60	67,90	80,10	58,80	165		08 :59 :19
	MPNet		64,60	83,90	52,50	66,90	79,00	58,10	165		12 :36 :00
Dégradé	Presse	Anglais	52,90	65,30	44,40	67,50	77,80	59,70	281	222	21 :59 :15
	USE		55,10	63,10	49,00	58,50	70,90	49,80	208		20 :14 :28
	MPNet		56,20	60,30	52,70	65,60	69,40	62,20	157		03 :39 :38
	Presse	Espagnol	45,00	62,80	35,20	64,80	75,30	56,90	191	149	06 :43 :26
	USE		60,50	69,90	53,30	61,50	68,80	55,70	160		08 :14 :18
	MPNet		50,60	48,20	53,20	54,50	54,80	54,20	111		09 :39 :21
	Presse	Allemand	57,80	93,80	41,70	58,90	86,70	44,60	224	118	08 :41 :30
	USE		64,80	84,80	52,40	66,60	80,90	56,60	181		09 :09 :18
	MPNet		64,20	83,40	52,20	66,00	78,60	56,80	169		15 :07 :19

TABLEAU 4.10 – Résultats des expérimentations sur le jeu de données *Event Registry* avec documents segmentés en deux, en appliquant l'algorithme non supervisé.

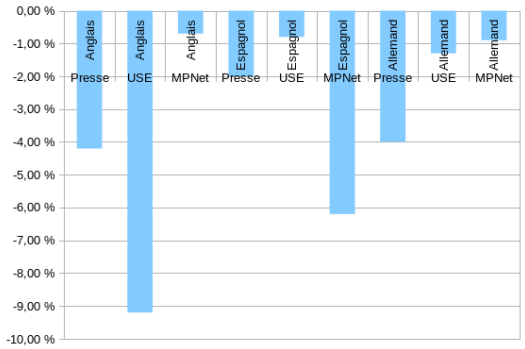
la segmentation entraîne une diminution plus importante de la qualité des *clusterings*, quel que soit l'algorithme (en moyenne -1,19 point pour l'algorithme supervisé, -2,37 dans l'autre cas). Pour ce dernier, l'algorithme non supervisé, on constate avec certaines vectorisations denses des chutes importantes : -9,20 en anglais avec des vecteurs *USE*, -6,20 en espagnol avec des encodages *MPNet*. Cette chute de la *F1* s'explique en partie par une forte diminution du rappel, associée à une augmentation du nombre d'événements



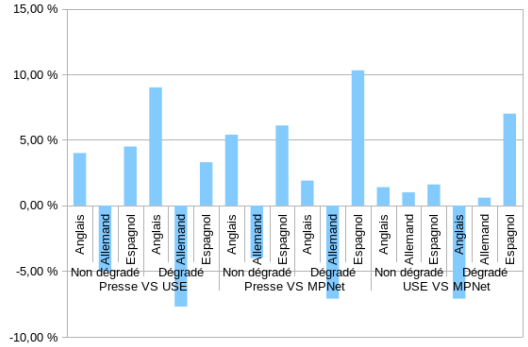
(a) Exécution sur articles dégradés contre articles non dégradés (algo. supervisé).



(b) Qualité de regroupement en fonction de la dégradation et de l'encodage (algo. supervisé).



(c) Exécution sur articles dégradés contre articles non dégradés (algo. non supervisé).



(d) Qualité de regroupement en fonction de la dégradation et de l'encodage (algo. non supervisé).

FIGURE 4.7 – Qualité des résultats de regroupement selon les dégradations ou le type de vecteur par langue du corpus *Event Registry* dégradé ou non dont les articles sont segmentés en deux.

trouvés par les algorithmes, et donc de *clusters*.

La comparaison des vectorisations entre elles, dans les figures 4.7b et 4.7d, montre qu'aucune des vectorisations *USE* ou *MPNet* ne se détache, elles produisent des résultats relativement similaires. Les vecteurs creux, obtenus par *TF-IDF* surpassent les vectorisations denses pour l'algorithme supervisé et conservent un léger avantage sur le non supervisé, notamment en termes de précision. Là où, avec l'algorithme supervisé, il est clair que les vecteurs *TF-IDF* sont plus performants, les vecteurs denses en allemand uniquement fonctionnent mieux avec l'algorithme K-Moyennes pour opérer un suivi de mentions d'événements.

Ici, nous avons, en plus des dégradations liées à l'*OCR*, le phénomène de sursegmentation qui altère les résultats. En plus des raisons que nous avons évoquées précédemment liées à la représentation des documents et des entités nommées, la segmentation tronque les phrases et les documents. Pour les vectorisations denses *USE* et *MPNet* qui vecto-

risent des phrases, une césure change intégralement le contenu des vecteurs. On constate par exemple qu'en anglais, les pondérations *TF-IDF* basées sur des corpus de presse produisent les mêmes résultats sur données dégradées ou non. Par contre, la diminution est plus faible, toujours dans cette langue avec les vectorisations denses. Cette capture du contexte sémantique, grande différence par rapport aux pondérations *TF-IDF* peut expliquer dans certains cas pourquoi ces vectorisations sont meilleures, par exemple en espagnol dans la figure 4.7a.

Event Registry, segmenté en trois

Les résultats bruts des expérimentations sont présentés dans les tableaux 4.11 pour l'algorithme supervisé et dans le 4.12 pour le non supervisé. L'analyse des dégradations et de l'impact des vectorisations est présentée en figure 4.8.

Données	Vecteurs	Langue	Résultats standards			Résultats *BCubed*			Nombre de groupes		Durée
			F1	P	R	F1	P	R	Prédits	Réels	
Non dégradé	Presse	Anglais	61,10	97,00	44,60	76,10	95,00	63,50	887	222	02 :18 :03
	USE		33,10	91,00	20,20	58,60	89,50	43,50	799		01 :03 :26
	MPNet		37,40	93,80	23,40	62,10	91,40	47,00	645		00 :51 :13
	Presse	Espagnol	86,60	97,10	78,20	81,60	95,70	71,20	382	149	00 :07 :46
	USE		56,50	90,10	41,20	70,70	90,30	58,00	324		00 :04 :58
	MPNet		57,20	89,40	42,00	70,90	88,10	59,30	297		00 :04 :39
	Presse	Allemand	58,30	99,60	41,20	71,30	96,70	56,40	326	118	00 :08 :16
	USE		34,60	96,10	21,10	64,90	92,50	49,90	251		00 :04 :41
	MPNet		33,40	98,00	20,10	63,80	94,80	48,00	250		00 :04 :47
Dégradé	Presse	Anglais	63,70	96,90	47,40	76,00	94,60	63,50	864	222	02 :26 :10
	USE		31,50	92,40	19,00	55,80	89,30	40,50	817		01 :05 :11
	MPNet		36,00	90,70	22,50	60,20	87,90	45,80	541		00 :43 :28
	Presse	Espagnol	81,80	94,90	71,90	77,50	92,40	66,80	375	149	00 :08 :33
	USE		58,70	90,80	43,40	68,60	86,00	57,10	281		00 :05 :03
	MPNet		54,70	87,50	39,80	68,60	86,00	57,00	309		00 :05 :25
	Presse	Allemand	53,10	99,50	36,20	69,40	95,80	54,40	307	118	00 :07 :58
	USE		38,90	95,80	24,40	63,80	90,80	49,20	248		00 :04 :41
	MPNet		36,10	95,90	22,20	63,50	89,80	49,10	196		00 :04 :10

TABLEAU 4.11 – Résultats des expérimentations sur le jeu de données *Event Registry* avec documents segmentés en trois, en appliquant l'algorithme supervisé.

L'autre niveau de segmentation testé est une division des articles en trois portions. À ce niveau, le constat est toujours le même dans les figures 4.8a et 4.8c : les vecteurs creux sont plus efficaces pour cette tâche que les vecteurs denses obtenus par des modèles d'apprentissage profond. À nouveau, la segmentation accentue la diminution des résultats par rapport à des dégradations sans segmentation avec une diminution de 1,84 point pour l'algorithme supervisé, et de 3,52 points pour l'algorithme non supervisé. La forte diminution de la *F1* est toujours liée à une chute encore plus importante du rappel et à l'augmentation du nombre de *clusters*. Les algorithmes faisant face à une réduction du volume de données disponibles par document (lié à leur segmentation) n'associent pas certains articles aux groupes auxquels ils sont rattachés, ce qui augmente le nombre de groupes total. Le rappel est dans le pire des cas inférieur de trente points par rapport au rappel du même algorithme, mais sur des données non segmentées.

Données	Vecteurs	Langue	Résultats standards			Résultats *BCubed*			Nombre de groupes		Durée
			F1	P	R	F1	P	R	Prédits	Réels	
Non dégradé	Presse	Anglais	60,60	67,40	55,00	64,70	76,40	56,00	439	222	22 :07 :18
	USE		69,20	74,40	64,60	69,00	75,50	63,50	236		06 :53 :45
	MPNet		61,90	62,00	61,80	70,20	72,60	67,90	209		06 :32 :00
	Presse	Espagnol	42,70	58,00	33,70	59,50	74,00	49,70	245	149	10 :00 :23
	USE		61,80	65,10	58,90	65,50	68,00	63,20	134		13 :14 :45
	MPNet		52,50	50,30	54,90	57,90	61,20	55,00	134		15 :19 :12
	Presse	Allemand	57,50	94,40	41,40	52,60	87,20	37,60	319	118	15 :09 :29
	USE		67,20	81,20	57,30	67,50	74,80	61,50	139		13 :53 :23
	MPNet		66,80	84,20	55,40	67,00	76,80	59,40	150		15 :34 :00
Dégradé	Presse	Anglais	61,10	66,40	56,60	64,70	74,10	57,50	293	222	03 :08 :59
	USE		56,10	55,70	56,60	59,00	59,90	58,10	157		07 :40 :23
	MPNet		61,30	61,10	61,60	65,20	64,60	65,80	143		05 :32 :12
	Presse	Espagnol	56,40	69,00	47,70	57,30	73,40	46,90	264	149	10 :53 :46
	USE		50,50	44,30	58,60	57,50	55,20	60,00	104		14 :33 :49
	MPNet		39,50	30,90	54,90	44,60	38,30	53,40	71		16 :26 :40
	Presse	Allemand	56,20	95,30	39,80	50,20	86,30	35,40	325	118	19 :35 :53
	USE		66,90	85,90	54,90	64,50	74,50	56,90	135		14 :15 :28
	MPNet		66,70	83,70	55,40	67,20	74,10	61,50	117		16 :24 :15

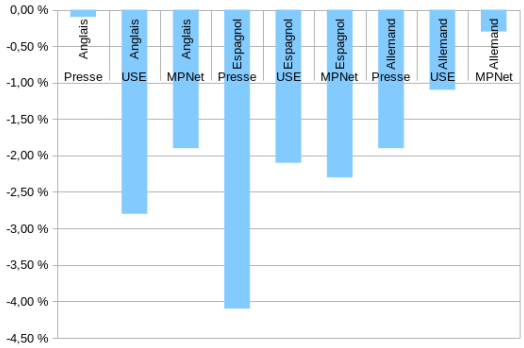
TABLEAU 4.12 – Résultats des expérimentations sur le jeu de données *Event Registry* avec documents segmentés en trois, en appliquant l'algorithme non supervisé.

Ce qui divise les deux algorithmes dans cette configuration, ce sont les types de vectorisation, décris aux figures 4.8b et 4.8d. Pour l'algorithme supervisé et un contexte non dégradé, les vecteurs *TF-IDF* sont toujours meilleurs que des vectorisations denses, de parfois 15 points, comme en espagnol comparé à *USE*. Pour l'algorithme non supervisé, le constat diffère selon les langues. Les vectorisations creuses sont meilleures, que les vectorisations denses en anglais et en espagnol. En allemand par contre, ce sont les vectorisations denses, obtenues par *MPNet* ou *USE* qui offrent les meilleurs résultats, de près de 15 points au-dessus de ceux obtenus par une vectorisation par pondération *TF-IDF*.

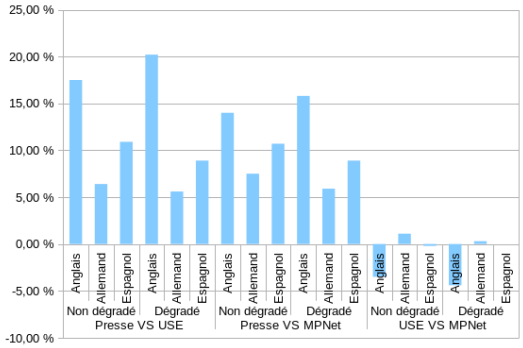
Nous reprenons les explications formulées à la section précédente pour une sursegmentation en trois portions. La baisse de rappel peut s'expliquer, dans tous les cas, par l'incapacité à discriminer les documents correctement. Les césures excessives qui séparent les sections décrivant les événements affectent les représentations par *TF-IDF* et les vectorisations denses. Également, nous avons vu en introduction de ce document que le style journalistique consiste en partie à répondre à des questions de base sur les événements. En divisant les réponses à ces questions, certaines sections se décontextualisent, rendant les algorithmes inefficaces.

Comparaison des niveaux de segmentation

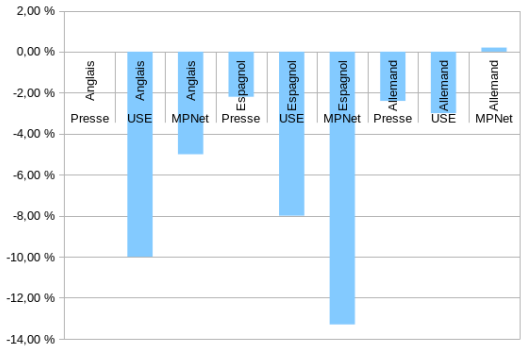
La comparaison des niveaux de segmentation, en deux ou en trois, par rapport aux versions non segmentées d'*Event Registry* montre tout d'abord que l'impact le plus important est lié à la segmentation. Les dégâts causés par l'*OCR* sont minimes (diminution d'un point en moyenne), mais ils se surajoutent aux effets produits par la segmentation. Ces tendances sont claires pour les deux algorithmes dans la figure 4.9 pour une seg-



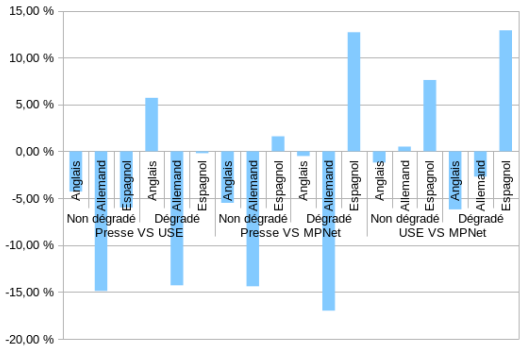
(a) Exécution sur articles dégradés contre articles non dégradés (algo. supervisé).



(b) Qualité de regroupement en fonction de la dégradation et de l'encodage (algo. supervisé).



(c) Exécution sur articles dégradés contre articles non dégradés (algo. non supervisé).



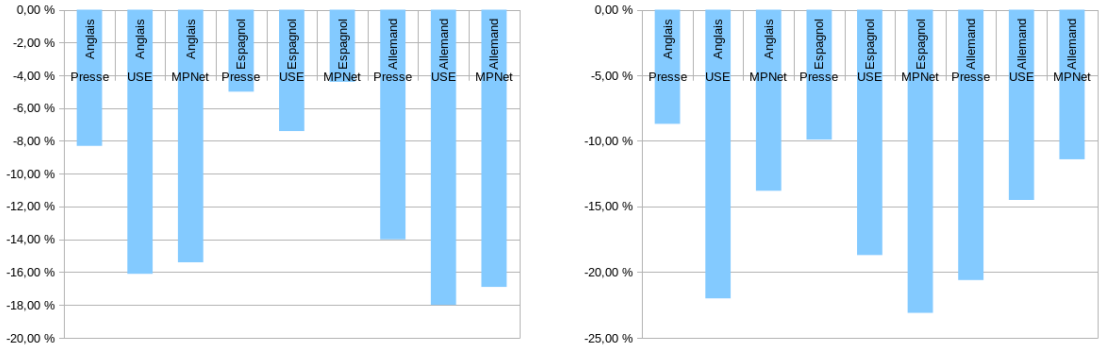
(d) Qualité de regroupement en fonction de la dégradation et de l'encodage (algo. non supervisé).

FIGURE 4.8 – Qualité des résultats de regroupement selon les dégradations ou le type de vecteur par langue du corpus *Event Registry* dégradé ou non dont les articles sont segmentés en trois.

mentation en deux et dans la figure 4.10 pour une segmentation en trois.

Tout d'abord, le rappel, quel que soit l'algorithme, décroît proportionnellement au niveau de segmentation, et diminue par voie de conséquence, la *F1*. Cette diminution du rappel est liée à la création dynamique des nombreux nouveaux *clusters*, les algorithmes peinant à identifier, avec le peu d'informations contenues dans les documents segmentés, les groupes auxquels chaque document pourrait être rattaché.

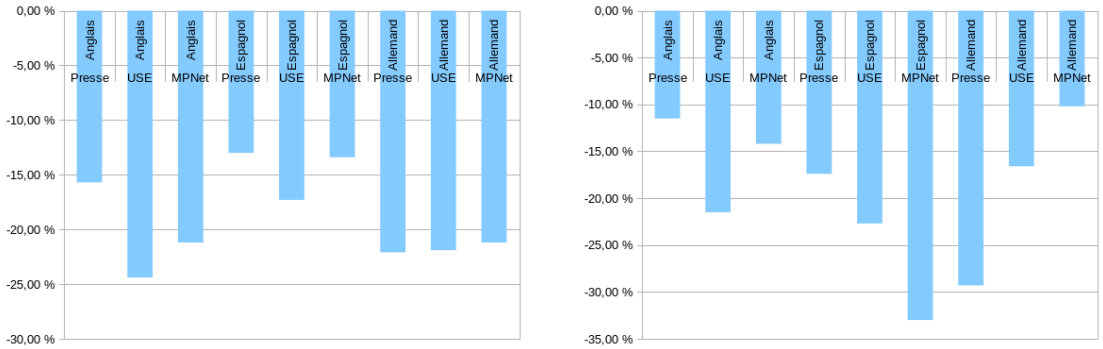
Avec une segmentation en deux portions ou en trois, aucune tendance ne semble se dégager quant aux vecteurs à préférer. Les *TF-IDF* semblent moins souffrir des dégradations des documents, comparé aux exécutions basées sur des vecteurs denses : la diminution est plus faible. Certaines langues ou plutôt les représentations vectorielles dans ces langues « résistent » mieux aux effets de segmentation que les autres. En anglais par exemple, les vectorisations creuses ou denses subissent davantage, de quelques points, les effets de segmentation. Pour l'allemand, le rappel diminue de près de 30 %



(a) Avec algorithme supervisé. (b) Avec algorithme non supervisé.

FIGURE 4.9 – Comparaison des deux algorithmes sur le jeu de données *Event Registry* segmenté en deux comparé à *Event Registry* non segmenté.

pour certaines vectorisations, entraînant dans sa chute la *F1*. L'algorithme supervisé subit moins ces effets de segmentation selon certaines dégradations, mais pas de façon systématique. En anglais avec les vecteurs *TF-IDF* par exemple, la diminution est moins importante, mais la qualité des résultats était déjà faible, comparée à l'exécution de l'algorithme supervisé.



(a) Avec algorithme supervisé. (b) Avec algorithme non supervisé.

FIGURE 4.10 – Comparaison des deux algorithmes sur le jeu de données *Event Registry* segmenté en trois comparé à *Event Registry* non segmenté.

4.3.2 Brèves et télégrammes : diffusion de messages courts

Les brèves et autres messages courts sont particuliers : la presse numérique et historique en regorge, comme nous l'avons indiqué en section 3.1. L'étude de la propagation des mentions d'événements dans des textes brefs vise à identifier les moyens et potentiels problèmes liés au traitement de messages courts comme les télégrammes. Les brèves de presse peuvent être semblables aux articles sursegmentés en termes de longueur, mais pas

en termes de contenu. Là où les brèves répondent en un texte condensé à des questions de base sur les événements : quoi, quand, qui et où, les articles le font en développant certains aspects plus en longueur, avec une narration spécifique. La segmentation divise alors potentiellement les réponses à ces questions en plusieurs portions, indépendantes.

Cette section s'intéresse à l'analyse de la diffusion des mentions d'événements dans des textes courts, telles les brèves de presse contenues dans les corpus *CoAID* et *FibVid*. Une expérience est également conduite sur le jeu de données *Event Registry* duquel seuls les titres sont extraits, à des fins de comparaison : *Event Registry Titles*.

CoAID

Les résultats bruts des expérimentations sont présentés dans les tableaux 4.13 pour l'algorithme supervisé et dans le 4.14 pour le non supervisé. L'analyse des dégradations et de l'impact des vectorisations est présentée en figure 4.11.

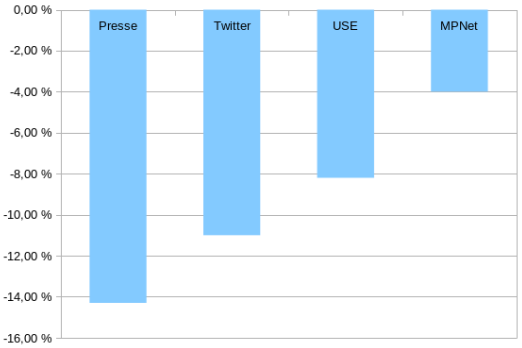
Données	Vecteurs	Résultats standards			Résultats *BCubed*			Nombre de groupes		Durée
		F1	P	R	F1	P	R	Prédits	Réels	
Non dégradé	Presse	53,40	67,60	44,10	59,80	79,40	48,00	6439	125	02 :17 :39
	Twitter	56,00	68,50	47,30	62,50	81,60	50,60	6077		01 :58 :15
	USE	11,70	12,50	11,00	20,70	36,70	14,40	6578		03 :44 :05
	MPNet	13,00	35,70	8,00	20,60	74,70	12,00	16745		18 :35 :19
Dégradé	Presse	38,60	55,30	29,60	45,50	69,90	33,70	6930		02 :35 :13
	Twitter	44,40	60,50	35,10	51,50	75,50	39,00	6963		02 :22 :44
	USE	8,30	8,40	8,30	12,50	14,90	10,70	1492		00 :56 :27
	MPNet	10,30	36,20	6,00	16,60	53,40	9,90	6990		03 :54 :13

TABLEAU 4.13 – Résultats des expérimentations sur le jeu de données *CoAID* en appliquant l'algorithme supervisé.

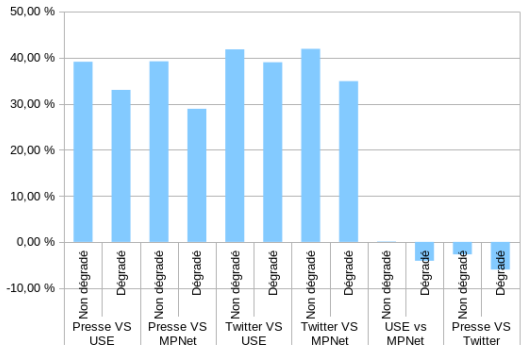
Données	Vecteurs	Résultats standards			Résultats *BCubed*			Nombre de groupes		Durée
		F1	P	R	F1	P	R	Prédits	Réels	
Non dégradé	Presse	31,30	40,20	25,60	34,30	62,50	23,70	669	125	18 :29 :30
	Twitter	26,90	53,50	18,00	27,50	66,20	17,30	835		07 :56 :54
	USE	18,40	47,90	11,40	19,90	54,60	12,10	669		20 :42 :56
	MPNet	13,10	38,10	7,90	15,30	45,90	9,20	791		07 :35 :14
Dégradé	Presse	20,60	23,50	18,30	26,10	50,60	17,60	570		03 :38 :32
	Twitter	25,70	33,00	21,10	29,30	55,70	19,90	590		18 :00 :40
	USE	12,90	42,10	7,60	15,10	50,40	8,90	836		20 :32 :16
	MPNet	7,50	4,10	41,00	29,00	24,00	36,70	330		21 :40 :06

TABLEAU 4.14 – Résultats des expérimentations sur le jeu de données *CoAID* en appliquant l'algorithme non supervisé.

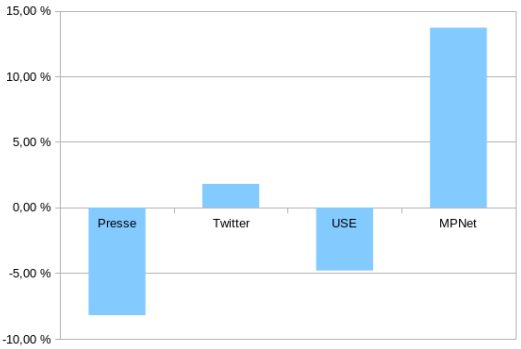
Pour les brèves du jeu de données *CoAID*, les vectorisations *TF-IDF* pondérées par des tweets et de la presse donnent des résultats plutôt bons avec l'algorithme supervisé (environ 60 % de *F1 BCubed* sans dégradation, et au moins 45 % dans un contexte dégradé). Les résultats obtenus par vectorisation dense, que ce soit par *USE* ou *MPNet* sont très faibles, inférieurs à 20 % sans dégradation et à 16 % avec des articles dégradés.



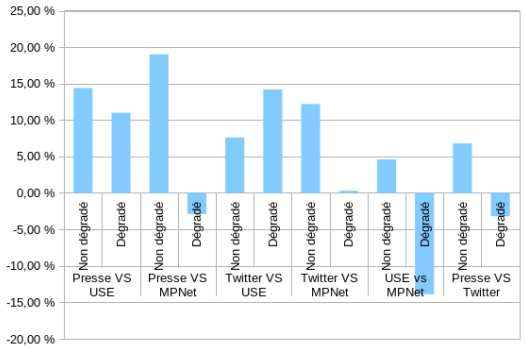
(a) Exécution sur articles dégradés contre articles non dégradés (algo. supervisé).



(b) Qualité de regroupement en fonction de la dégradation et de l'encodage (algo. supervisé).



(c) Exécution sur articles dégradés contre articles non dégradés (algo. non supervisé).



(d) Qualité de regroupement en fonction de la dégradation et de l'encodage (algo. non supervisé).

FIGURE 4.11 – Qualité des résultats de regroupement selon les dégradations ou le type de vecteur du corpus *CoAID* dégradé ou non.

La dégradation des documents semble avoir un impact moindre sur les résultats présentés en figure 4.11a. Cela doit donc être nuancé : certes, les résultats obtenus par vectorisation dense sont moins sensibles aux erreurs d'*OCR*, mais ils fournissent systématiquement des résultats inexploitables. Ceux obtenus par pondération *TF-IDF* tendent à valider l'hypothèse formulée en sous-section 4.1.1 : les corpus de vectorisation *TF-IDF* doivent être les plus proches des données à vectoriser. Ici, la pondération *TF-IDF* basée sur un corpus de tweets journalistiques est moins affectée que celle basée sur des articles de presse et donne des résultats bruts supérieurs de quelques points par rapport aux autres.

Ce constat est plus partagé concernant l'algorithme non supervisé dont les résultats sont visibles à la figure 4.11c. Étonnamment, la vectorisation par modèle *MPNet* donne de meilleurs résultats sur les données dégradées que sur les données non dégradées. La tendance générale est néanmoins toujours baissière : les dégradations diminuent la capacité de l'algorithme à regrouper les articles ensemble. Même si l'impact est faible, les résultats bruts sont toujours largement inférieurs (de plus de 20 points) à ce que peut

faire l'algorithme supervisé.

Pour les vecteurs, dans un contexte supervisé (figure 4.11b), les pondérations *TF-IDF* sont meilleures que les vectorisations denses (de parfois près de 40 points), qui elles-mêmes sont assez similaires entre elles. Avec les K-Moyennes en figure 4.11d, les différences sont plus modestes entre vectorisations creuses et denses (maximum 20 points, pour la comparaison corpus *TF-IDF* de presse contre vectorisation par *MPNet*). Ce dernier encodage semble avoir un intérêt, avec cet algorithme, dans un contexte dégradé : il donne parfois de meilleurs résultats que les autres modèles utilisés.

Il est important de noter que, conformément à ce qui a été présenté en section 2.3, la répartition temporelle des données de *CoAID* n'est pas homogène dans le temps. Beaucoup de documents sont publiés sur une courte période correspondant à une seule fenêtre pour l'algorithme des K-Moyennes. Dans ce cas, cela peut entraîner un biais : les documents publiés sur cet intervalle de temps sont tous groupés ensemble davantage parce qu'ils sont publiés à des dates proches que parce que leur contenu est similaire.

Pour expliquer ces résultats, nous devons retenir le contenu même des brèves de presse. Nous avons évoqué, à la section 3.1, que les télégrammes condensent, en peu de texte, les informations sur l'événement. Bien que *CoAID* ne contienne pas de télégrammes, nous assimilons son contenu à des brèves de presse. Les réponses aux questions (qui, où, quand) étant en général des entités nommées, nous conservons la même analyse que pour le corpus *Event Registry*. L'incapacité à identifier correctement ces entités influe sur les vectorisations et donc les résultats des algorithmes. Les pondérations *TF-IDF* résistent mieux que les vectorisations denses. En ayant peu de texte à encapsuler, les modèles *USE* et *MPNet* génèrent peut-être des vectorisations trop générales qui rendent impossible la discrimination des documents entre eux, comme nous pouvons le constater dans les tableaux 4.13 et 4.14.

FibVid

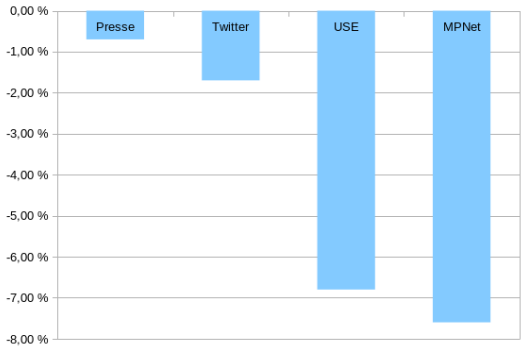
Les résultats bruts des expérimentations sont présentés dans les tableaux 4.15 pour l'algorithme supervisé et dans le 4.16 pour le non supervisé. L'analyse des dégradations et de l'impact des vectorisations est présentée en figure 4.12.

Données	Vecteurs	Résultats standards			Résultats *BCubed*			Nombre de groupes		Durée
		F1	P	R	F1	P	R	Prédits	Réels	
Non dégradé	Presse	19,40	26,80	15,10	37,60	70,10	25,70	203	52	00 :00 :02
	Twitter	30,70	34,00	28,00	41,10	48,10	35,90	99		00 :00 :01
	USE	10,30	6,00	37,70	16,30	9,90	45,10	12		00 :00 :01
	MPNet	27,80	27,90	27,70	38,10	41,50	35,20	85		
Dégradé	Presse	20,80	30,70	15,70	36,90	63,20	26,10	182		00 :00 :02
	Twitter	25,10	42,40	17,80	39,40	66,60	27,90	183		00 :00 :01
	USE	3,60	1,80	91,60	9,50	5,00	92,50	6		00 :00 :00
	MPNet	21,70	16,90	30,40	30,50	25,40	38,30	38		00 :00 :02

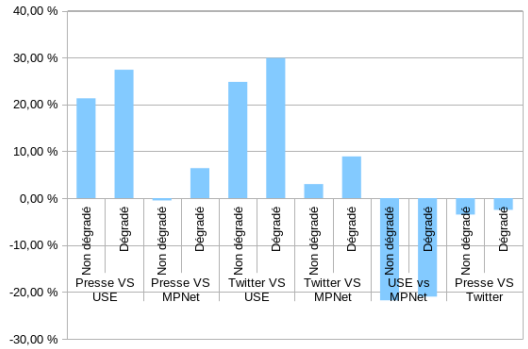
TABLEAU 4.15 – Résultats des expérimentations sur le jeu de données *FibVid* en appliquant l'algorithme supervisé.

Données	Vecteurs	Résultats standards			Résultats *BCubed*			Nombre de groupes		Durée
		F1	P	R	F1	P	R	Prédits	Réels	
Non dégradé	Presse	19,30	20,70	18,10	32,30	38,50	27,90	74	52	00 :09 :19
	Twitter	19,80	22,40	17,70	33,10	41,00	27,80	79		00 :09 :32
	USE	25,50	26,40	24,60	36,20	39,60	34,40	70		00 :10 :02
	MPNet	24,90	26,10	23,80	35,70	39,60	32,40	68		00 :10 :48
Dégradé	Presse	18,40	20,10	16,90	32,20	39,80	27,10	82		00 :09 :14
	Twitter	18,40	18,30	18,50	31,70	35,70	28,50	67		00 :09 :12
	USE	24,20	24,70	23,70	34,60	37,20	32,20	69		00 :10 :02
	MPNet	24,00	25,20	22,90	35,30	39,00	32,10	68		00 :10 :53

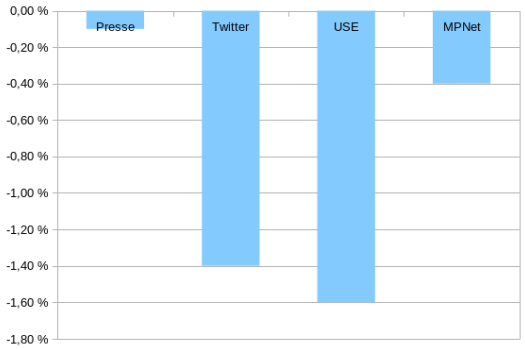
TABLEAU 4.16 – Résultats des expérimentations sur le jeu de données *FibVid* en appliquant l'algorithme non supervisé.



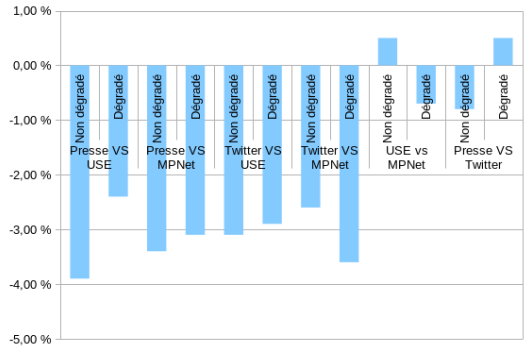
(a) Exécution sur articles dégradés contre articles non dégradés (algo. supervisé).



(b) Qualité de regroupement en fonction de la dégradation et de l'encodage (algo. supervisé).



(c) Exécution sur articles dégradés contre articles non dégradés (algo. non supervisé).



(d) Qualité de regroupement en fonction de la dégradation et de l'encodage (algo. non supervisé).

FIGURE 4.12 – Qualité des résultats de regroupement selon les dégradations ou le type de vecteur du corpus *FibVid* dégradé ou non.

Le cas des données *FibVid*, rapporté en figure 4.12, est intéressant, car les résultats sont assez mauvais quel que soit l'algorithme ou l'encodage utilisé, à au maximum 40 %

de *F1 BCubed*. L'effet des dégradations est davantage marqué sur les résultats de l'algorithme supervisé, où la diminution pour l'encodage *MPNet* est de plus de 7 points. Le meilleur résultat pour l'algorithme supervisé est obtenu avec la pondération *TF-IDF* basée sur des tweets (tableau 4.15). Vient ensuite l'encodage dense *MPNet*. Cela tend à nouveau à confirmer l'hypothèse sur le besoin d'utiliser des corpus proches en termes de contenu avec les données à pondérer si l'on utilise *TF-IDF*. Les vectorisations *TF-IDF* basées sur des corpus de presse ou de tweets limitent fortement la diminution des résultats causée par la dégradation des documents.

Pour l'algorithme supervisé, les encodages *TF-IDF* sont plus adaptés, comparativement aux vectorisations denses, qui, elles tendent à mieux fonctionner pour l'algorithme non supervisé. La différence de résultats entre les deux algorithmes est d'ailleurs étonnamment faible, de seulement quelques points, là où elle était jusqu'à 30 points pour le jeu de données *CoAID*. Enfin, la vectorisation *USE* n'est pas du tout adaptée pour l'algorithme supervisé, donnant les pires résultats du groupe. Elle donne pourtant les meilleurs résultats avec l'algorithme des K-Moyennes.

Event Registry Titles

Les résultats bruts des expérimentations sont présentés dans les tableaux 4.17 pour l'algorithme supervisé et dans le 4.18 pour le non supervisé. L'analyse des dégradations et de l'impact des vectorisations est présentée en figure 4.13.

Données	Vecteurs	Langue	Résultats standards			Résultats *BCubed*			Nombre de groupes		Durée
			F1	P	R	F1	P	R	Prédits	Réels	
Non dégradé	Presse	Anglais	54,00	97,10	37,40	67,60	93,30	53,00	1056	222	00 :03 :50
	Twitter		53,60	96,20	37,10	64,10	92,50	49,10	1196		00 :04 :11
	USE		50,00	91,30	34,40	70,30	88,90	58,10	606		00 :12 :34
	MPNet		49,40	93,40	33,60	71,60	90,90	59,00	507		00 :12 :25
	Presse	Espagnol	59,10	94,30	43,00	67,40	92,80	52,90	542	149	00 :00 :20
	Twitter		51,20	97,30	34,70	59,20	96,90	42,60	700		00 :00 :32
	USE		71,20	74,00	68,60	70,50	65,50	76,30	121		00 :00 :43
	MPNet		73,30	78,30	68,80	76,00	73,40	78,70	159		00 :00 :56
	Presse	Allemand	26,00	89,10	15,20	53,20	85,60	38,60	320	118	00 :00 :14
	Twitter		26,90	85,00	16,00	53,70	79,20	40,60	221		00 :00 :10
	USE		53,10	95,70	36,70	70,20	88,40	58,20	210		00 :01 :17
	MPNet		30,70	98,90	18,20	64,80	96,40	48,80	284		00 :02 :04
Dégradé	Presse	Anglais	46,90	97,40	30,90	59,40	93,50	43,50	1406	222	00 :04 :59
	Twitter		49,40	97,00	33,10	59,50	92,00	44,00	1347		00 :04 :27
	USE		42,70	93,20	27,70	64,50	91,70	49,70	819		00 :17 :44
	MPNet		54,00	85,80	39,50	68,90	77,30	62,10	209		00 :04 :59
	Presse	Espagnol	52,40	84,10	38,00	56,70	76,80	44,90	354	149	00 :00 :25
	Twitter		51,50	89,80	36,10	57,20	88,30	42,50	619		00 :00 :21
	USE		57,20	87,80	42,40	72,10	86,30	61,90	297		00 :01 :49
	MPNet		56,70	89,70	41,50	72,70	86,10	62,90	243		00 :01 :13
	Presse	Allemand	20,60	88,80	11,70	48,30	85,40	33,70	376	118	00 :00 :14
	Twitter		21,80	87,00	12,40	49,50	82,50	35,30	306		00 :00 :23
	USE		78,00	77,40	78,70	66,20	56,00	81,00	61		00 :00 :32
	MPNet		29,70	92,50	17,70	60,60	87,80	46,30	218		00 :01 :11

TABLEAU 4.17 – Résultats des expérimentations sur le jeu de données *Event Registry Titles* en appliquant l'algorithme supervisé.

Données	Vecteurs	Langue	Résultats standards			Résultats *BCubed*			Nombre de groupes		Durée
			F1	P	R	F1	P	R	Prédits	Réels	
Non dégradé	Presse	Anglais	49,80	57,70	43,70	58,40	72,60	48,80	259	222	06 :22 :19
	Twitter		51,90	59,70	46,00	60,30	73,50	51,10	245		06 :20 :49
	USE		48,30	84,90	33,80	70,80	82,80	61,80	222		20 :27 :59
	MPNet		61,90	75,50	51,60	74,30	79,70	69,50	243		00 :00 :35
	Presse	Espagnol	40,50	43,20	38,10	54,50	61,60	48,90	167	149	00 :51 :58
	Twitter		46,90	47,60	46,20	56,70	63,40	51,30	175		00 :51 :45
	USE		74,00	83,20	66,70	74,10	76,90	71,50	153		02 :57 :02
	MPNet		71,30	82,40	62,90	72,80	76,70	69,30	157		03 :43 :24
	Presse	Allemand	26,10	46,60	18,20	52,60	69,70	42,30	169	118	01 :00 :17
	Twitter		33,70	64,20	22,90	52,90	75,10	40,90	192		00 :59 :44
	USE		62,70	80,20	51,50	74,40	81,10	68,60	128		03 :34 :42
	MPNet		61,30	84,70	48,10	72,10	81,50	64,60	138		03 :52 :16
Dégradé	Presse	Anglais	47,20	60,70	38,60	56,50	69,40	47,60	229	222	06 :37 :25
	Twitter		49,20	54,40	44,80	57,50	68,50	49,50	237		06 :27 :29
	USE		53,70	65,90	45,40	70,60	76,40	65,60	192		22 :09 :13
	MPNet		60,30	76,20	50,00	72,20	78,40	66,80	206		23 :33 :37
	Presse	Espagnol	42,50	41,00	44,10	53,20	59,70	48,00	185	149	00 :51 :59
	Twitter		35,00	33,70	36,40	52,20	55,50	49,40	150		00 :52 :09
	USE		53,70	69,60	43,80	67,30	71,20	63,80	142		02 :59 :34
	MPNet		55,50	65,80	47,90	67,00	68,70	65,30	138		03 :27 :49
	Presse	Allemand	38,50	58,40	28,70	56,10	67,80	47,80	127	118	01 :00 :49
	Twitter		49,30	68,00	38,70	55,80	70,30	46,20	163		00 :59 :58
	USE		49,20	72,60	37,20	69,30	76,90	63,10	125		03 :26 :40
	MPNet		62,60	85,20	49,50	69,60	76,80	63,60	112		03 :51 :13

TABLEAU 4.18 – Résultats des expérimentations sur le jeu de données *Event Registry Titles* en appliquant l'algorithme non supervisé.

Les titres d'articles du corpus *Event Registry* peuvent être analysés de la même façon que le sont *CoAID* et *FibVid*, étant entendu que leur nature, leur forme et le fond du texte diffèrent entre ces deux types de documents. L'algorithme supervisé donne de meilleurs résultats en général que l'algorithme non supervisé, de parfois seulement quelques points. Pour les deux algorithmes, les vectorisations denses obtenues par les modèles *MPNet* ou *USE* donnent de meilleurs résultats, de façon constante, que les vectorisations obtenues par pondération *TF-IDF*. La plus grande différence de résultats pour les deux algorithmes est en allemand, comme cela est visible dans les figure 4.13c et figure 4.13d. Dans les deux cas, la baisse de performance est d'environ 20 points par rapport au même jeu de données dans lequel les corps des articles sont conservés et utilisés. Par rapport à ce corpus, *Event Registry*, l'algorithme Miranda et coll., donne des résultats, en termes de précision, assez semblables à ceux issus du traitement des textes complets et pas seulement des titres. Le rappel est dans tous les cas assez bas, autour de 50 % et 60 %, que ce soit avec cet algorithme ou les K-Moyennes. Pour ce dernier, l'écart entre résultats obtenus à l'aide des vecteurs creux et des vecteurs denses s'accroît. K-Moyennes fonctionne à nouveau mieux avec des encodages denses qu'avec des vecteurs obtenus par pondération *TF-IDF*, quel que soit le corpus utilisé, de presse ou de tweets, pour pondérer les termes.

Concernant l'effet des dégradations, avec l'algorithme supervisé, la perte de *F1 BCubed* entre les documents non dégradés et les dégradés est de 4,42 points en moyenne, et de 2,32 avec l'algorithme de K-Moyennes. Ce dernier écart est certes plus faible, mais il

est à contrebalancer avec des résultats bruts qui sont de base bien inférieurs. En outre, sur les figures 4.13a et 4.13b, la différence par type de vectorisation est assez flagrante : l'algorithme K-Moyennes y semble bien moins sensible que celui supervisé, la majorité des baisses de résultats étant contenues à -3 points. Pour l'algorithme Miranda et coll., la vectorisation par pondération *TF-IDF* est celle qui entraîne une plus forte dégradation, comparée aux encodages denses. Aucun modèle ne se dégage des résultats obtenus par les K-Moyennes. Enfin, les figures 4.13c et 4.13d confortent les précédentes affirmations. Pour l'algorithme supervisé, aucune vectorisation creuse par *TF-IDF* ne donne de meilleurs résultats que les encodages denses, quelle que soit la langue ou l'état de dégradation des documents. Ce sont les encodages denses qui, pour *Event Registry Titles* donnent de meilleurs résultats de *clustering*. Pour l'algorithme K-Moyennes, les pondérations obtenues par *TF-IDF* se basant sur un corpus de tweets donnent des résultats proches de ceux obtenus par vectorisation dense.

En comparant avec les résultats obtenus sur les corpus *CoAID* à *FibVid*, on note une différence importante, notamment pour la métrique *F1 BCubed*. Les résultats sur ce jeu de données sont meilleurs d'environ dix points par rapport à *CoAID*. L'explication de cette différence est très certainement liée à l'observation que nous avons faite à la section 2.3.2 : des documents sont dupliqués. La duplication ou réutilisation partielle de texte, si elle est présente dans de longs documents, peut être noyée dans le texte. En nous focalisant seulement sur les titres de presse, le contenu est plus court et s'il est dupliqué, il est identique entre tous les documents. Les vectorisations produites sont identiques et par conséquent, les algorithmes les traitent en commettant moins d'erreurs.

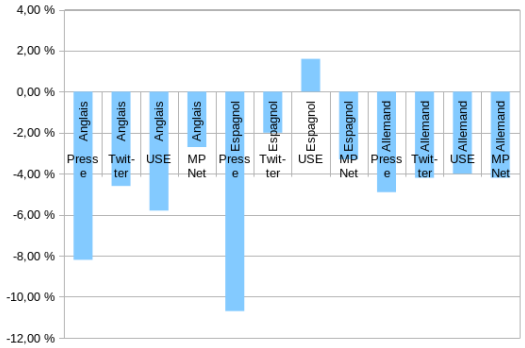
4.4 Conclusion

Dans ce chapitre, nous avons analysé et comparé le comportement de deux algorithmes, l'un supervisé et l'autre non face à la problématique de regroupement d'articles de presse dégradés décrivant les mêmes événements. Les jeux de données intègrent différents niveaux de dégradation, par *OCR* et par sursegmentation. Les documents sont encodés pour les algorithmes de quatre manières différentes : deux utilisent des pondérations *TF-IDF*, les autres des vectorisations par des modèles multilingues d'apprentissage profond.

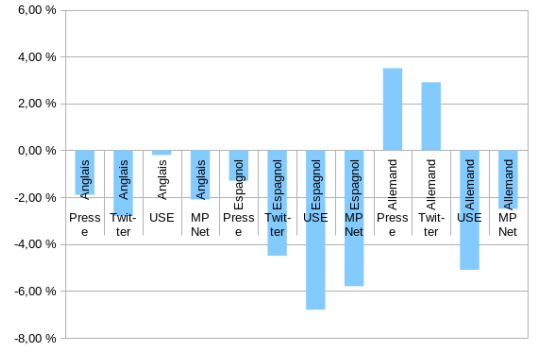
Plusieurs conclusions peuvent être tirées de ces travaux concernant l'impact des dégradations et de la segmentation, le contexte d'utilisation des algorithmes supervisés et non supervisés ainsi que les encodages creux ou denses.

— Les dégradations de contenu introduites par l'*OCR* ont un impact négligeable sur les résultats fournis par ces algorithmes. La baisse de performance dépend des encodages utilisés.

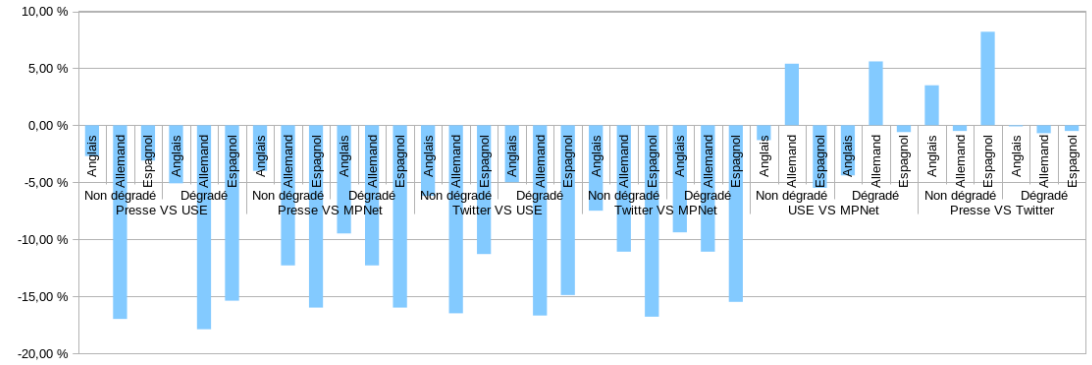
— L'effet de segmentation est le plus important : en divisant les articles de presse en plusieurs segments, les deux algorithmes ne sont pas en mesure d'identifier de grands groupes de documents traitant des mêmes événements. À la place, on constate une forte diminution du rappel, et donc l'augmentation du nombre de *clusters*.



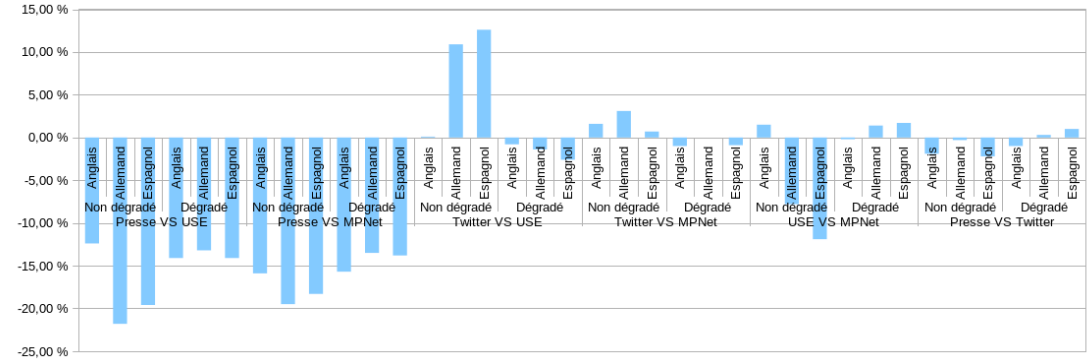
(a) Exécution sur articles dégradés contre articles non dégradés (algo. non-supervisé).



(b) Exécution sur articles dégradés contre articles non dégradés (algo. non-supervisé).



(c) Qualité de regroupement en fonction de la dégradation et de l'encodage (algo. supervisé).



(d) Qualité de regroupement en fonction de la dégradation et de l'encodage (algo. non supervisé).

FIGURE 4.13 – Qualité des résultats de regroupement selon les dégradations ou le type de vecteur du corpus *Event Registry Titles* dégradé ou non.

— L'effet de segmentation est accentué par les dégradations *OCR*. Alors qu'elles ont peu d'impact seules, l'ajout de dégradations à un document déjà segmenté réduit la capacité des outils à reconnaître correctement les groupes d'articles traitant des mêmes événements.

— Les vectorisations par *TF-IDF* sont les plus adaptées pour l'utilisation de l'algorithme supervisé présenté par Miranda et coll. À l'inverse, les encodages denses donnent de meilleurs résultats, en général, sur l'algorithme de K-Moyennes.
— Pour *Event Registry*, il est possible d'analyser les résultats par langue. Les vectorisations par pondération *TF-IDF* en allemand sont souvent battues par les vectorisations fournies par les modèles *USE* et *MPNet*, pour les mêmes données et mêmes algorithmes. Cela peut s'expliquer par la morphologie syntaxique de l'allemand, qui nécessiterait un corpus *TF-IDF* de termes et de lemmes plus important que celui que nous avons collecté en section 4.1 pour mieux discriminer les documents.
— Les expériences menées ici tendent à nuancer la conclusion présentée par Staykovski et coll. [Sta+19] selon laquelle les encodages denses ne sont pas adaptés au suivi d'événements. Avec l'algorithme de K-Moyennes, et pour certains types de données, les résultats avec ces vecteurs surpassent ceux obtenus avec des pondérations *TF-IDF*.

La qualité des données est un frein à l'analyse de ces résultats. L'une des limites majeures, parmi d'autres, et qui n'est visible qu'après une analyse exploratoire comme celle réalisée en section 2.3 est liée à la présence de doublons dans le contenu. Certains articles ou brèves sont entièrement rédigés par des agences et republiés par des sites d'information ou des journaux, entraînant une duplication du contenu. Nous avons déjà évoqué, à la section 2.2, la réutilisation de texte [Oiv+19] pour le suivi d'événements dans des textes historiques. La duplication est alors le témoin d'une circulation d'une information dans la presse. De nouvelles expériences doivent évaluer l'importance de cette problématique dans les données d'évaluation.

Depuis l'écriture de ces lignes, de nouveaux modèles pour *S-BERT* [RG20] sont régulièrement publiés. Le dernier en date pour l'anglais uniquement, `all-mpnet-base-v2`[8], basé sur un modèle fourni par Microsoft, est plus performant que *USE* ou *MPNet* sur les tâches de similarité sémantique. Il est entraîné avec un volume de données encore plus important : l'utiliser pourrait améliorer les résultats présentés dans ce chapitre. Entraînée seulement sur des données en anglais, l'utilisation de ce modèle ne devrait pas remettre en cause les résultats que nous avançons ici, notamment sur les deux autres langues.

Les processus présentés dans ce chapitre posent des limites à leur mise en œuvre. Les modèles supervisés nécessitent des données d'entraînement là où l'autre algorithme peut s'en affranchir. Les temps d'exécution des expériences avec K-Moyennes sont très longs, souvent trop longs pour des cas d'utilisation réels. Pour l'algorithme de Miranda et coll., les temps d'exécution sont quadratiques, en fonction du nombre de documents à traiter. Les résultats sont comparables à ceux de l'état de l'art pour les documents nativement numériques [Mir+18 ; Mir+18 ; LH20 ; SMM22]. Nous venons de montrer que la dégradation des documents entraîne une diminution de leur efficacité. La contrepartie de ces bons résultats est qu'ils exigent d'importantes ressources calculatoires pour obtenir des résultats et fournir une analyse. Nous proposons dans le prochain chapitre une stratégie

8. `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

alternative, basée sur un autre compromis entre la performance et la rapidité d'exécution. Comparée à la proposition présentée ici, cette autre stratégie donne des résultats de moins bonne qualité, mais instantanés.

Chapitre 5

Des événements aux documents de presse

Sommaire

Les dégradations du climat et de la biodiversité sont induites par les activités humaines polluantes et destructrices [Int13 ; Suá+22]. Les questions d'économies et de sobriété sont au cœur des enjeux de notre début de siècle. La décroissance est un concept politique, économique et social qui est lié à la préservation de notre environnement et qui peut s'appliquer à l'informatique. Les puissances de calcul augmentaient jusqu'à la dernière décennie comme la rapporte la loi empirique dite « Moore », théorisant le doublement annuel des capacités de calcul [Les22a]. Cette affirmation s'est avérée exacte pendant un temps, mais ne l'est plus. Pourtant, la vision qui en découle affecte les consommations et orientations du numérique. La puissance de calcul disponible permet de résoudre des problèmes qui se complexifient en utilisant des architectures et des modèles de plus en plus consommateurs de matériel et d'énergie [Lig21]. Nous souhaitons dans ce chapitre explorer une autre voie qui répond à la problématique de suivi de mentions d'événements historiques. Des contraintes à ce processus sont fixées en amont : il doit fonctionner sans données d'entraînement, fournir des réponses rapides, voire instantanées et réduire au maximum le besoin en ressources informatiques (temps processeur et mémoire vive de l'ordinateur).

La solution conçue et évaluée dans ce chapitre reprend les concepts des moteurs de recherche, mais adaptés aux documents de presse. Considérons un ou une scientifique en sciences sociales qui souhaite analyser des documents relatifs à un événement historique donné. Il ou elle le décrit grâce à son expertise en répondant aux questions fondamentales quoi, quand, où, qui et comment. De ces informations on peut générer une requête afin de sélectionner les documents qui décrivent cet événement.

Dans une première partie (section 5.1), nous introduirons un mécanisme de description automatique d'événements à partir de bases de connaissances publiques comme Wikipédia. La description des événements est fortement liée à la langue et aux lieux où ils se produisent. En section 5.2 nous analyserons l'impact des choix linguistiques pour décrire des événements à partir de données publiques. Ensuite, nous verrons que les annotations décrites au chapitre 2 pour grouper les événements ne sont pas suffisantes. Nous devons faire le lien entre les événements du corpus *Event Registry* et des concepts d'événements se trouvant dans des bases de connaissances. C'est ce que nous proposons en section 5.3. Enfin, aux sections 5.4 et 5.5 le moteur de recherche lui-même est décrit et évalué. De la même façon qu'au chapitre 4, l'objectif premier est d'examiner l'impact des dégradations d'*OCR* et de segmentation de texte sur les méthodes et outils développés. Un second objectif est de proposer une alternative économique en ressources, en opposition aux méthodes proposées au chapitre précédent.

5.1 Caractérisation des événements

Pour construire une requête efficace et rechercher des documents de presse rapportant des événements, la description et la représentation de ces derniers sont cruciales. Il est nécessaire de collecter des informations exhaustives sur eux, depuis par exemple des bases de données publiques. L'objectif est de décrire les événements le plus précisément possible [STH09 ; EN11]. Ce sont souvent les entités nommées qui fournissent les informa-

tions élémentaires sur les événements (quoi, quand, etc.) [YB18] et qui sont considérées comme des qualificateurs d'événements. En l'état actuel de nos connaissances, aucune méthode ou outil n'existe pour extraire une représentation aussi exhaustive que possible des événements historiques. Les projets de recherche proposent en général leurs propres définitions et représentations, adaptées à leurs besoins, et bien que des bases de références sont disponibles, rien n'existe pour les exploiter à travers un processus unifié.

La méthode proposée dans ce chapitre exploite les ressources présentes sur Wikidata [VK14] et Wikipédia. Wikipédia est une encyclopédie internationale et universelle à laquelle chacun peut contribuer. Wikidata est un graphe de connaissance reposant sur une riche ontologie décrivant l'ensemble des concepts de cette encyclopédie, dont des événements. Il est supposé que ces deux bases de données fournissent suffisamment d'informations pour caractériser les événements. Ce travail s'inscrit dans la continuité des recherches du projet *ACE* [Dod+04] et des ontologies d'événements [STH09 ; vHag+11]. Les qualificatifs d'événements sont utilisés dans le même contexte que peuvent l'être les arguments dans *ACE*. Nous l'avons vu et défini au chapitre 2, les arguments, généralement des entités nommées, contextualisent les événements. Nous avons mentionné ces qualificateurs dans notre définition d'événement (définition 10). La méthode exploite Wikidata pour extraire des informations de base (le type, le nom, la date, etc.) et Wikipédia pour agréger l'ensemble des entités qui y sont rattachées et qui y participent.

5.1.1 Collecte des informations élémentaires

Wikidata décrit les événements à l'aide d'une grande variété de types, séparés en deux catégories. Chacune a ses subtilités. La première s'applique à des événements avec prémisses (concerts, élection, cérémonie) alors que les autres n'en ont pas (catastrophe naturelle, attaque) : ils ne sont pas annoncés dans la presse. La définition proposée dans d'anciens travaux qualifie de *Wikidata Event Type* (*WET*) ces deux catégories d'événements [Rud+19]. La définition adoptée dans ce contexte est celle d'« actions du monde réel ancrées dans le temps, dans l'espace et au sein desquels s'impliquent des participants », comme nous l'avons vu au chapitre 2. De chaque concept représentant un événement, ne sont collectés que le type, la date, les lieux éventuellement mentionnés ainsi que les différents noms qui lui sont associés. Ces propriétés discriminent les événements entre eux. Il est hautement improbable que deux événements soient décrits par un même tuple contenant ces quatre propriétés, c'est-à-dire que deux événements distincts du même type se produisent au même endroit, dans un même lieu et fassent intervenir les mêmes personnes.

En tant que projet communautaire, Wikidata n'est pas une source de données exhaustive. S'attendre à une extraction complète de qualificateurs d'événements en ne comptant que sur Wikidata n'est pas possible comme le montre le tableau 5.1. Par exemple, presque aucun événement ne possède d'attribut de type participant et seulement la moitié sont ancrés temporellement et géographiquement.

Type d'entité nommée [Sun95]	Propriété Wikidata	Nombre d'événements	Proportion
PER[SON]	Participant (*P710*)	58,885	6.18%
DATE	Temps (*P585*, *P580*, *P582*)	511,312	53.69%
LOC[ATION]	Lieu (*P7*, *P276*)	524.532	55.08%

TABLEAU 5.1 – Proportion de *WET* décrits par des attributs de lieux, de date ou de participants. Il y a un total de 952 351 événements.

5.1.2 Les entités impliquées dans les événements

Face au constat que les informations manquent si Wikidata est utilisée seule, et pour aller plus loin, nous proposons d'analyser Wikipédia à la recherche d'autres entités participantes. Chaque article débute par un en-tête qui officie comme résumé et comporte les informations essentielles pour le comprendre [The21b]. Dans Wikipédia, des liens internes connectent les articles entre eux et à Wikidata. Nous utilisons ces liens, extraits des en-têtes d'articles Wikipédia pour détecter les entités impliquées dans l'événement, quelles qu'elles soient. Par ce biais, il est possible d'enrichir la représentation obtenue de Wikidata avec des informations temporelles, géographiques ainsi que des participants supplémentaires lorsque ces données sont absentes du graphe de connaissances.

Il y avait en avril 2021 310 éditions linguistiques actives de Wikipédia. Pour des questions de performance et d'efficacité, il n'est pas possible d'analyser l'ensemble de ces éditions pour extraire les entités liées à chaque événement. N'en retenir qu'un maximum de N, chiffre choisi arbitrairement est une possible solution. Le concept de langue pivot (*hub* [RM12]) suppose que les éditions linguistiques les plus volumineuses sont plus fortement interconnectées aux autres éditions et répertorient davantage de concepts. Se concentrer sur les dix plus grandes éditions de Wikipédia paraît alors un choix raisonnable. Des statistiques sur ces différentes éditions sont présentées en tableau 5.2. En sus du nombre d'articles, elles couvrent le nombre de pages modifiées, de contributeurs (distingués des contributeurs actifs) et la profondeur des articles [The20]. Cette dernière information est un indicateur de qualité des articles basé sur l'édition des contenus.

Rang	Langue	Articles	Pages modifiées	Contributeurs	Contributeurs actifs	Profondeur d'article
		en millions			*en milliers*	
1	**Anglais**	6,151	1200	386	32	1026,81
4	**Allemand**	2,475	241	50	5,5	93,6
5	**Français**	2,246	280	54	5,1	237,67
7	*Russe*	1,657	173	40	3,4	135,94
8	**Italien**	1,631	183	44	2,5	169,03
9	**Espagnol**	1,622	270	87	4,2	208,81
10	*Polonais*	1,425	113	14	1,3	30,99

TABLEAU 5.2 – Propriétés des différentes éditions linguistiques de Wikipédia. Éditions triées par nombre décroissant d'articles (données collectées en octobre 2020).

Parmi les dix plus grandes éditions de Wikipédia, celles en cebuano (2e), suédois (3e) et danois (6e) sont principalement rédigées par un robot [Les22b] et par conséquent exclues de cette sélection. En cas de conflit, plus grande est la profondeur d'article, meilleure sera classée l'édition de Wikipédia. Ne sont ainsi retenues que cinq langues : l'anglais, l'allemand, le français, l'italien et l'espagnol. Ces langues représentent 30 % des locuteurs natifs dans le monde (2,361 [EGC21] sur 7,874 milliards de terriens en 2021) et 25,14 % de tous les articles de Wikipédia (14,125 [The21a] sur 56,615 millions d'articles). Il est supposé que ces éditions linguistiques de Wikipédia sont suffisantes pour collecter des qualificateurs d'événements précis. Néanmoins, cette sélection est biaisée. Elle exclut la plupart des langues asiatiques et africaines comprenant beaucoup de locuteurs et locutrices comme le mandarin ou l'arabe. Les éditions de Wikipédia dans ces langues sont plus modestes. Elles contiennent respectivement 1,292 million et 1,176 million d'articles [The22].

Depuis les résumés d'articles sont collectées les entités correspondant à des personnes, à des lieux et à des organisations ou des entités géopolitiques. Le nombre d'occurrences de chaque entité est calculé pour toutes les éditions linguistiques analysées. Cet index dénote l'importance de l'entité au regard de l'événement. Il est supposé que celles mentionnées au sein de multiples en-têtes d'articles sont importantes dans la description de l'événement.

Prenons un exemple comme l'assassinat de Raspoutine [1], identifié par le concept `Q2882749` sur Wikidata. Depuis le graphe de connaissance, la date, les lieux, participants et noms de l'événement sont extraits. Les propriétés associées au type de l'événement sont inconnues à ce stade, comme l'auteur du crime pour un événement de type « assassinat ». Les participants et les lieux sont liés à des concepts Wikidata et identifiés par des *URI*s dans le graphe. Cette représentation est enrichie par l'analyse des articles Wikipédia décrivant l'assassinat de Raspoutine. Les seuls articles qui existent à ce sujet sont rédigés en français et en espagnol. Après application du processus décrit précédemment, on obtient parmi d'autres ces triplets : `(PER, Q312997, [Felix Yusupov, auteur], 3)`, `(PER, Q43989, [Grigori Rasputin], cible], 2)`, `(GPE, Q34266, [Russian Empire], 1)`. Les poids, respectivement 3, 2 et 1 tendent à montrer, selon l'hypothèse formulée précédemment, que connaître qui a tué la victime est plus important, pour la rédaction de Wikipédia dans ces langues, que de savoir où l'action s'est déroulée. Ces entités et ces dénombrements synthétisent la connaissance historique et fournissent une information objective de l'implication des entités dans les événements.

5.1.3 Transcrire les événements en langage naturel

La description d'un événement revient à associer des propriétés absolues comme des dates ou des noms avec des éléments identifiés par des *URI*s et liés à des graphes de connaissance. Cette description, en tant que telle, c'est-à-dire sous forme numérique, est agnostique à la langue. Il n'est pas affirmé ici que ce sont les événements qui le sont : la représentation et la connaissance d'un événement sont liées à la langue et aux pratiques

1. `https://fr.wikipedia.org/wiki/Assassinat_de_Raspoutine`

culturelles. C'est-à-dire que certains points de vue locaux, exprimés par des pratiques linguistiques distinctes, peuvent modifier les interprétations liées aux événements. Néanmoins, notre approche se base sur la collecte d'entités nommées, et non la sémantique qui y est associée. Cette approche tend à limiter ce biais.

Dans la plupart des cas, les entités Wikidata sont décrites par un ou plusieurs noms alternatifs qui sauvegardent les différentes orthographes utilisées pour un même concept. Pour poursuivre avec l'exemple précédent, en français, l'entité `Q312997` [2] sur Wikidata est indifféremment écrite *Félix Youssoupoff* ou *Felix Youssoupov*. Ces nuances orthographiques se constatent également dans les corpus historiques qui intègrent les usages de leurs époques. Dans les textes, nous ne pouvons pas utiliser `Q312997`, car cet identifiant n'a de sens que pour Wikidata. Par conséquent, connaître ces différentes alternatives textuelles est nécessaire pour rechercher ces entités dans les textes, ces dernières pouvant avoir été représentées par différentes graphies.

La dernière étape de ce processus de description d'événement transforme cette représentation abstraite de liens Wikidata en une description dans une langue particulière. Elle prend en compte l'ensemble des noms alternatifs de chacune des entités qui décrivent l'événement et les enregistre dans le langage cible. Cette approche permet d'obtenir une représentation d'un événement dans une langue même si elle n'a pas été analysée sur Wikipédia. Pour terminer avec l'exemple de l'assassinat de Raspoutine, cela signifie obtenir une description en italien alors que cet événement n'est décrit sur Wikipédia qu'en français et en espagnol.

5.1.4 Synthèse

Dans cette première section, nous avons proposé une méthode pour caractériser les événements du monde réel tels que définis dans ce domaine de recherche. La méthode se base sur l'ontologie de Wikidata et sur l'analyse des articles de Wikipédia pour extraire toutes les entités qui participent à l'événement. L'approche est multilingue, les entités sont identifiées par des *URI*s desquels on obtient une représentation dans n'importe quelle langue, sous réserve de données. Nous supposons qu'en sélectionnant un sous-ensemble de toutes les éditions linguistiques de Wikipédia, il est possible de collecter efficacement la plupart des entités de l'événement. Ce processus est implémenté au sein d'un paquet logiciel [Ber20]. Il fournit une *API* pour extraire une représentation d'événement et toutes ses entités participantes à partir d'un simple identifiant Wikidata. Il est possible d'étendre les capacités de l'outil en ajoutant des connexions vers des graphes de connaissance comme YAGO2 [Hof+13] ou *EventKG* [GD19 ; AGD20].

5.2 Description des événements et influence de la langue

La méthode présentée à la section précédente permet de collecter des qualificateurs d'événements disponibles dans les bases encyclopédiques. Néanmoins, la sélection arbitraire de certaines langues pivots est biaisée. Plusieurs langues largement employées sont

2. `https://www.wikidata.org/wiki/Q312997`

exclues, car elles sont moins visibles sur Wikipédia. Les éditions linguistiques en arabe, mandarin, hindi, bengali, portugais ou russe sont concernées, pour ne citer qu'elles. Un compromis doit être trouvé entre un nombre faible de langues analysées, pour la performance, et la complétude des descriptions. Une solution peut consister à exploiter les langues vernaculaires en plus de celles analysées par défaut. Dans ce contexte, ce sont les langues qui sont parlées dans les lieux où se produisent les événements. Pour évaluer cette hypothèse, un processus d'évaluation est mis en place.

5.2.1 Description des données

Pour comparer l'influence des langues, nous avons construit deux jeux de données distincts contenant les mêmes événements. Du fait de la nature ambiguë [Spr18] de ce qu'est un événement, on qualifie d'indiscutable un événement considéré comme tel dans de multiples domaines de recherche (histoire [Sha10], psychologie [Min75] ou traitement du langage naturel [MBC19 ; XW19]). Comme nous l'avons montré en état de l'art, au chapitre 2, cela restreint à seulement certains types, dont six que l'on prend en exemple, répartis en trois groupes : les attentats et assassinats, les catastrophes naturelles et les événements politiques. Sur Wikidata, les deux premiers groupes décrivent des événements soudains, imprévus. Le dernier évoque des événements avec prémisse, annoncés et planifiés.

— **Attentats et assassinats** : assassinat politique (`Q1139665`) et attaque terroriste (`Q2223653`) ;
— **Catastrophes naturelles** : séisme (`Q7944`) et éruption volcanique (`Q7692360`) ;
— **Événements politiques** : cérémonie (`Q2627975`) et élection (`Q40231`).

Wikidata n'est pas une base exhaustive et ne prétend pas l'être. Le nombre d'événements qu'elle contient n'est pas uniformément réparti au fil des ans, mais a tendance à croître depuis le début du XXI^e siècle. Cette augmentation ne doit pas être interprétée comme un accroissement du nombre d'occurrences d'événements dans le monde, mais comme liée à l'amélioration de la qualité des données. Parmi les nombreux événements survenus depuis le début du siècle, tous ou presque contiennent un champ de date qui les ancre temporellement [Rud+19]. Par conséquent, nous décidons de nous concentrer uniquement sur les événements survenus au cours des cinquante dernières années, de janvier 1970 à décembre 2019. Cela assure d'exclure les événements mal documentés. Dans l'intérêt des expériences, il est nécessaire de traiter les événements qui sont décrits dans au moins un article Wikipédia. Dans le cas contraire, aucune comparaison n'est possible entre les deux versions qui utilisent ou non les langues vernaculaires. Par conséquent sont exclus les événements sans aucun article. Leur nombre est reporté en tableau 5.3.

Le processus de collecte des entités participantes décrit en section 5.1 peut être lent, en fonction de leur nombre. Il implique de nombreuses requêtes vers les *API* de Wikidata et l'analyse d'un article peut prendre de quelques secondes à plusieurs minutes, en fonction du nombre d'entités présentes dans les en-têtes des articles. En conséquence, seul un sous-ensemble d'au plus cinquante événements sont sélectionnés aléatoirement pour chaque type d'événement. Ils satisfont les contraintes mentionnées précédemment : tous ont une propriété de type date et sont liés à au moins un article Wikipédia.

Groupe	Type d'événement	Événements	Avec article	Proportion
Attentats et assassinats	Assassinat politique	44	24	54,55%
	Attaque terroriste	905	806	89,06%
Catastrophes naturelles	Séisme	1 102	987	89,56%
	Éruption volcanique	23	18	78,26%
Événements politiques	Cérémonie	11 428	11 233	98,29%
	Élection	29 236	24 488	82,10%

TABLEAU 5.3 – Nombre et proportion d'événements décrits par au moins un article Wikipédia, quelle que soit la langue (même hors des langues pivots), de 1970 à 2019.

Langues vernaculaires

La différence entre les deux jeux de données réside dans les éditions linguistiques de Wikipédia qui sont analysées pour collecter les entités participantes. Le premier corpus est appelé « langues de base » et est construit en analysant les éditions de Wikipédia dans les cinq langues choisies. Pour l'autre, nommé « toutes les langues », les éditions rédigées dans les langues vernaculaires sont analysées en plus de celles de base. Dans le cas où un projet linguistique n'existerait pas, c'est le dialecte source qui est utilisé. Par exemple, l'anglais américain `en-us` n'existe pas sur Wikipédia, mais est un dialecte de l'édition en anglais, `en`.

Ce processus est à nouveau biaisé par la notion de temporalité. Par exemple, le français était une langue officielle en Algérie jusqu'en 1950. Cette information n'est plus disponible sur Wikidata : les données sont mises à jour continuellement au gré des changements, ce qui entrave la recherche des langues parlées historiquement sur un lieu. Pour contrebalancer ce problème, Wikidata fournit en plus des langues officielles, celles employées par les personnes. Nous formulons l'hypothèse que davantage que les langues officielles, ce sont les langues employées au quotidien par les personnes de ces territoires qui ont de l'importance. L'autre hypothèse qui en découle est que les événements sont mieux décrits sur Wikipédia dans l'une des langues parlées dans les lieux où ils se produisent.

Il peut également arriver que les entités géopolitiques changent au fil des ans. L'insurrection de Pâques en 1916 à Dublin (Irlande) s'est déroulée dans l'entité *United Kingdom of Great Britain and Ireland*. Ce pays n'existe plus, remplacé par le *United Kingdom* depuis 1922. Ce n'est néanmoins pas un problème puisque Wikidata fournit les langues officielles ou employées. Dans tous les cas, ce risque est limité par la sélection d'événements s'étant produits entre l'année 1970 et 2019.

Synthèse

Ce corpus de données est publié [Ber+21a] et contient 241 événements divisés en six types et trois groupes, comme présentés ici. Le premier contient des descriptions d'événements obtenues suivant le procédé présenté en section 5.1 avec uniquement les cinq langues de base. Le second corpus contient des représentations d'événements en y

ajoutant les langues vernaculaires. Les données collectées, il est nécessaire de définir un cadre d'évaluation pour valider ou invalider les hypothèses formulées. Ce cadre nécessite la recherche de métriques donnant une relation d'ordre entre deux représentations d'événements : ainsi, nous serons en mesure de déterminer quelle représentation est la meilleure.

5.2.2 Métriques expérimentales

La comparaison entre les deux jeux de données est basée sur quatre métriques. Elles évaluent l'hypothèse selon laquelle l'analyse des langues vernaculaires sur Wikipédia fournit des informations supplémentaires et plus précises à propos des entités participant aux événements. Ces métriques sont respectivement le nombre d'entités participantes trouvées dans les en-têtes des articles Wikpédia (**M1**), le nombre de termes dans l'en-tête (**M2**), le ratio entre les deux précédentes métriques (**M3**) ainsi que le nombre de noms alternatifs trouvés sur Wikidata pour chaque entité participante (**M4**), comme mentionné à la sous-section 5.1.3.

Le tableau 5.4 décrit ces métriques pour l'assassinat de John Fitzgerald Kennedy en 1963 [3]. Il a eu lieu aux États-Unis d'Amérique, un pays anglophone. Par conséquent, aucune autre langue n'est analysée en supplément des langues de base. On note ici que la version en anglais fournit plus d'informations à propos de cet événement. Il y a davantage d'entités dans l'en-tête, celle-ci est plus longue et il y a plus de noms alternatifs les décrivant en anglais que dans n'importe quelle autre langue.

Caractéristique	Langue				
	Italien	Espagnol	Allemand	Anglais	Français
M1 : entités participantes dans l'en-tête	40	42	13	**47**	35
M2 : termes de l'en-tête	218	125	179	**316**	279
M3 : ratio	0.183	**0.336**	0.073	0.149	0.125
M4 : noms alternatifs pour chaque entité	83	130	166	**224**	115

Tableau 5.4 – Les quatre métriques pour les langues décrivant l'assassinat de John Fitzgerald Kennedy.

Dans un second temps, les langues sont triées par ordre décroissant, la plus haute valeur dans le tableau 5.4 étant la meilleure. Ces tris sont présentés dans le tableau 5.5 et montrent que la version en anglais de Wikipédia est la mieux adaptée pour décrire les entités participant à l'événement. L'anglais est à la première place en termes d'entités participantes, de termes dans l'en-tête et de noms alternatifs. Le ratio entre ces informations peut parfois être aberrant, peut-être à cause des différents styles de rédaction adoptés par les communautés linguistiques de Wikipédia. Cela peut expliquer pourquoi

3. `https://www.wikidata.org/wiki/Q193484`

l'espagnol est devant les autres langues sur cette métrique. Il peut être une pratique commune dans certaines éditions de n'écrire qu'un court en-tête, mais riche de liens internes. Cela permet au lecteur ou à la lectrice de naviguer dans l'encyclopédie et connaître les notions essentielles pour comprendre l'article qu'il ou elle lit.

Caractéristique	1er	2ème	3ème	4ème	5ème
M1 : entités participantes dans l'en-tête	**anglais**	espagnol	italien	français	allemand
M2 : termes de l'en-tête	**anglais**	français	italien	allemand	espagnol
M3 : ratio	espagnol	italien	**anglais**	français	allemand
M4 : noms alternatifs pour chaque entité	**anglais**	allemand	espagnol	français	italien

TABLEAU 5.5 – Langues triées dans lesquelles l'assassinat de John Fitzgerald Kennedy est le mieux décrit pour chacune des métriques.

5.2.3 Évaluation

Bien que nous ayons exclu les événements sans aucun article sur Wikipédia lors de la construction des corpus (sous-section 5.2.1), il se peut que pour certains les pages qui existent soient présentes dans des éditions linguistiques qui ne sont pas analysées. Ils n'ont pas été filtrés à la première étape parce que ces événements sont décrits par au moins un article, mais dans une langue qui n'est ni une langue de base ni une langue vernaculaire. Il est également nécessaire d'exclure les événements pour lesquels nous ne connaissons aucune des langues officielles ou employées. Les deux se recoupent dans la plupart des cas. Nous reportons dans le tableau tableau 5.6 le nombre d'événements exclus par cette sélection finale. Lorsque l'on considère les langues vernaculaires, le nombre d'événements à analyser augmente. C'est un premier argument en faveur de notre hypothèse selon laquelle les événements sont mieux décrits dans leurs langues vernaculaires.

Type d'événement	Événements	Sans référence à une langue	Sans article	
			Langues de base	*Toutes les langues*
Assassinat politique	24	0	1	0
Attaque terroriste	50	13	5	2
Séisme	50	11	17	11
Éruption volcanique	17	1	3	2
Cérémonie	50	29	9	6
Élection	50	10	11	2

TABLEAU 5.6 – Nombre d'événements exclus à cause de manque de données dans les langues traitées.

Afin de comparer la description des événements dans les deux ensembles de données, nous appliquons, pour chaque événement, les mêmes calculs que ceux présentés dans

les tableaux 5.4 et 5.5. Pour chaque métrique, nous vérifions si l'une des langues officielles parlées sur le lieu de l'événement figure parmi les trois langues qui le caractérisent le mieux. Les résultats présentés dans le tableau 5.7 montrent, pour chaque métrique, pour combien d'événements une langue vernaculaire est parmi les trois langues qui le représentent le mieux. Les résultats sont significatifs avec seulement la première langue, mais la sélection des trois meilleures tend à limiter le problème décrit dans l'exemple de l'assassinat de Kennedy. Ce faisant, nous affirmons que les cinq langues que nous avions précédemment identifiées comme langues principales ne sont pas suffisantes pour extraire avec précision les qualificatifs des événements. Cette affirmation confirme l'hypothèse de l'importance des langues vernaculaires pour représenter des événements. Nous avons précédemment retenu cinq langues pivots seulement pour accélérer le processus de description d'événement. En ajoutant uniquement les langues vernaculaires, le nombre de langues est toujours restreint, mais la description s'améliore.

Type d'événement	Jeu de données	Nb. événements	Nombre d'événements mieux décrit par une langue vernaculaire			
			***M1** : entités*	***M2** : termes*	***M3** : ratio*	***M4** : noms alt.*
Assassinat politique	Base	23	14	14	13	12
	Toutes		**22**	**22**	**20**	**18**
Attaque terroriste	Base	34	25	25	25	27
	Toutes		**30**	**30**	**29**	**28**
Séisme	Base	25	13	13	12	11
	Toutes		**24**	**24**	**23**	**20**
Éruption volcanique	Base	14	12	12	12	12
	Toutes		**13**	**13**	**13**	**13**
Cérémonie	Base	18	12	13	13	15
	Toutes		**15**	**15**	**15**	**17**
Élection	Base	31	27	27	26	26
	Toutes		**30**	**30**	**29**	26

TABLEAU 5.7 – Comparaison du nombre d'événements mieux décrits dans une langue vernaculaire. Les trois meilleures langues qui décrivent l'événement sont prises en compte.

5.2.4 Analyse des erreurs

Pour la majorité des événements qui contredisent l'hypothèse de l'importance des langues vernaculaires, la raison principale est liée à l'absence de ressources dans ces langues. Il manque des articles sur Wikipedia et nous ne pouvons donc pas analyser d'en-tête. Les articles manquants sont dus par exemple à une petite communauté de locuteurs et de locutrices, ce qui réduit la taille de l'édition de Wikipédia dans cette langue. Ils peuvent s'expliquer par un biais culturel ou politique. Ce dernier cas est principalement vrai pour les événements de type assassinats et attentats ou liés à des conflits et qui sont traités, ou non, différemment en fonction des éditions linguistiques de Wikipédia. Pour quelques événements de ce type, la langue vernaculaire est en quatrième position ou même plus loin et n'est pas la meilleure pour représenter l'événement donné. Cette dernière affirmation est vraie dans des cas de censure étatique, par exemple.

Les langues traitées sont principalement indo-européennes et basées sur l'alphabet latin. Le processus d'extraction de termes (*tokenization*) dans ces langues est assez uniforme et donc comparable. Nous ne présumons pas des résultats sur des langues d'un

autre type comme le mandarin bien qu'il soit improbable que cette limite affecte les présentes conclusions.

5.2.5 Synthèse

Dans cette section est conduite une analyse critique de la méthode présentée précédemment pour caractériser des événements à partir de sources ouvertes. Il peut être tentant de ne sélectionner que des langues pivots [RM12] pour extraire des représentations d'événements à partir de sources comme Wikipédia, mais cela n'est pas suffisant. L'expérience montre que l'analyse des éditions linguistiques dans les langues vernaculaires en plus des pivots améliore systématiquement, pour les types d'événements choisis, la représentation. Les analyses réalisées ici [Ber+21a] sont également publiées pour être réutilisées et vérifiées.

Une fois les événements décrits de la meilleure façon possible, en prenant en compte certaines spécificités linguistiques comme la nécessité d'analyser les éditions en langue vernaculaire, il est nécessaire de concevoir et d'évaluer un outil qui associe cette description d'événement aux documents qui le mentionnent. Pour cette évaluation, des annotations d'événements sont capitales : un lien doit être établi entre les groupes d'événements des jeux de données et les entités Wikidata qui décrivent ces groupes, donc ces événements.

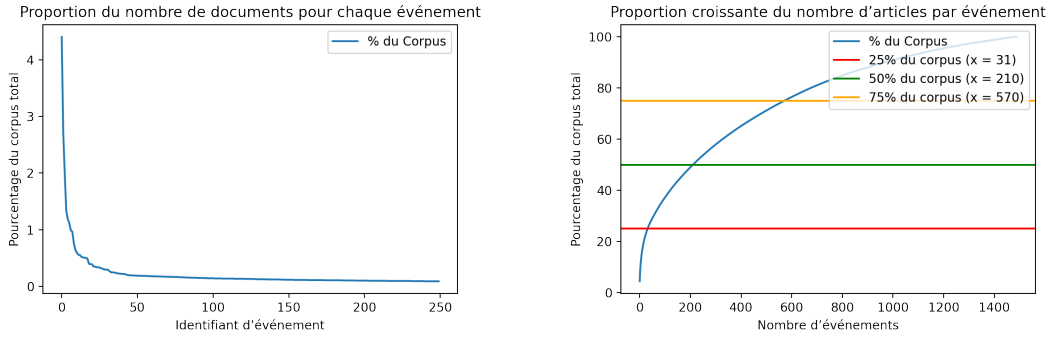
5.3 Identification des événements nommés dans du texte

Le système développé dans ce chapitre est détaillé dans la section section 5.5 qui suit. Pour le moment, considérons-le comme une boîte noire. En entrée de ce processus se trouve un événement, identifié dans le graphe de Wikidata, par exemple `Q274498`, qui représente le séisme du Kantō (Japon) de 1923. La sortie est une liste de documents qui décrivent l'événement et dans ce cas précis ses conséquences. Pour l'évaluer, il faut être en mesure de savoir quels documents sont associés à quelle entité sur Wikidata. Ainsi, le processus expérimental complet consistera à exécuter le système sur l'ensemble des événements connus et à l'évaluer d'après les documents qu'il renvoie. Malheureusement, de telles annotations ne sont pas fournies dans les jeux de données sélectionnées en section 2.3, c'est à nous de les créer, manuellement, et de les faire valider par un collectif d'annotateurs et d'annotatrices. *Event Registry* est le seul corpus multilingue, le seul donc à comporter des événements bilingues ou trilingues ainsi que des articles complets. Le choix est d'analyser uniquement ce jeu de données et d'y annoter chaque événement avec son identifiant sur Wikidata. Puisque des données d'entraînement ne sont pas utiles dans ce contexte, les répartitions en corpus d'entraînement et de test sont rassemblées en un unique jeu de données.

5.3.1 Sélection des événements annotés

Le corpus de documents que nous avons sélectionné, *Event Registry*, contient plus de trente mille documents (tableau 2.4). Face à ce volume de données, nous n'en sélec-

tionnons qu'un échantillon. Nous nous focalisons sur des événements de grande ampleur ainsi que de plus petits événements. C'est ce qui est présenté dans cette section.



(a) Pourcentage du corpus occupé par chaque événement.

(b) Somme cumulée croissante du nombre d'articles par événement.

FIGURE 5.1 – Statistiques décrivant les choix opérés pour la sélection des événements à annoter.

Dans *Event Registry*, comme le montre la figure 5.1, certains événements sont de taille très importante, jusqu'à 4 % de l'ensemble (environ 1 500 documents). Ils sont six à peser plus de 1 %, représentant ensemble 12 % des trente-trois mille documents du corpus. Viser l'annotation des 25 % plus gros événements revient à n'en annoter que 31, comme le montre la figure 5.1b. Pour augmenter le nombre d'événements annotés et conserver la règle des 25 % de données que nous souhaitons voir annotées, nous sélectionnons les six plus gros événements et x autres choisis aléatoirement jusqu'à ce que le nombre d'articles sélectionné atteigne un quart du corpus. Dans ce cas-ci, $x = 108$ et les 114 événements sélectionnés englobent précisément 25,02 % (soit 8 459 documents) de tous les articles d'*Event Registry*.

5.3.2 Annotation des événements

Pour associer à chaque événement d'*Event Registry* un identifiant Wikidata, nous parcourons les événements du jeu de données un à un. Après avoir analysé les articles de chaque événement, nous associons ces derniers à un unique identifiant sur Wikidata. Cette tâche est fastidieuse et une première tentative a cherché à automatiser le processus. Les entités nommées et dates étaient extraites des articles et utilisées dans le moteur de recherche de Wikidata. La vérification des propositions faites aurait été opérée *a posteriori* par des évaluateurs et évaluatrices volontaires. Face à la médiocre qualité initiale des résultats, la décision est prise d'annoter ces événements manuellement.

Les articles associés à un même événement le décrivent tous d'une manière similaire, avec les mêmes entités et sont publiés sur la même période. Pour simplifier la recherche des événements décrits dans les articles, un tableau du même type que le 5.8 est créé, contenant tous les événements sélectionnés. Seuls les titres des articles sont conservés,

de même que la date du premier article publié et celle du dernier. L'événement est temporellement borné par ces deux dates.

#	Début	Fin	Anglais	Espagnol	Allemand	Wikidata	Réel	Spécifique
746645	2014-05-15	2014-05-16		'Al menos cinco muertos por las inundaciones en Serbia y Bosnia', 'Loli consigue una cama y una silla eléctrica gracias al pueblo', 'En 18 meses podría instalarse la quinta turbina de "El Cajón"'	'Hier droht nun Hochwasser', 'Schwere Überschwemmungen : Ausnahmezustand in Serbien und Bosnien', '120 Liter Regen in 24 Stunden'	Q16879871	✓	✓

TABLEAU 5.8 – Annotation d'un événement du corpus *Event Registry* avec son concept Wikidata.

Nous annotons chaque événement avec trois informations. La première est l'identifiant de l'événement sur Wikidata. Dans le tableau 5.8, l'événement 746 645 du corpus *Event Registry*, les inondations de 2014 en Serbie [AFP14], est décrit par des articles publiés en espagnol et en allemand. Une recherche manuelle sur Wikidata permet d'identifier le concept `Q16879871` [4] comme étant le bon candidat pour l'annotation. La seconde annotation détermine si l'événement est réel, c'est-à-dire s'il correspond à la définition d'une « action ancrée dans le temps, dans l'espace et qui implique des participants » (définition 10, page 40). La dernière indique si le concept Wikidata identifié est spécifique à cet événement ou non. Pour cet exemple, l'élément Wikidata `Q16879871` décrit bien *2014 Southeast Europe floods* (Inondations du sud-est de l'Europe en 2014). Un événement non spécifique serait par exemple un but marqué par un joueur de football au sein d'une compétition internationale. Le match pourrait être référencé sur Wikidata, mais pas l'action du joueur de marquer. Dans ce cas, l'annotation indiquerait que l'événement n'est pas décrit par un concept spécifique.

Comme indiqué à la section précédente, nous annotons 114 événements du corpus *Event Registry*. Ils représentent 7 151 documents, soit 21 % du corpus total et 84 % de l'ensemble des articles annotés sont associés à une entité Wikidata. 71 % des événements sont considérés comme réels et 32 % sont spécifiques. Sur l'ensemble des 93 % d'événements rattachés à une entité Wikidata, 45 % sont réels. Au total, 81 événements sont annotés par un identifiant Wikidata.

Les événements sont tous différents et leurs représentations varient dans le jeu de données. La catégorie avec le plus de documents relève du judiciaire (2 événements, 1 544 articles), puis les conflits (guerre, conflit géopolitique, etc.) avec 9 événements et 1 207 documents. Viennent ensuite le sport (32 événements, 1120 articles) et la politique (10 événements, 943 articles). Les autres événements de ce corpus annoté sont présentés en figure 5.2.

Event Registry est un corpus de documents rédigés en anglais, en espagnol et en allemand. Nous avons pu commettre des erreurs d'interprétation ou de compréhension en annotant les documents. Nous ne sommes en effet natif dans aucune de ces langues. La conformité des annotations doit donc être évaluée par des personnes extérieures. L'évaluation a pour objectif d'attester l'exactitude des annotations (d'événement, de son caractère réel ou spécifique) et des entités Wikidata associées aux événements.

4. `https://www.wikidata.org/wiki/Q16879871`

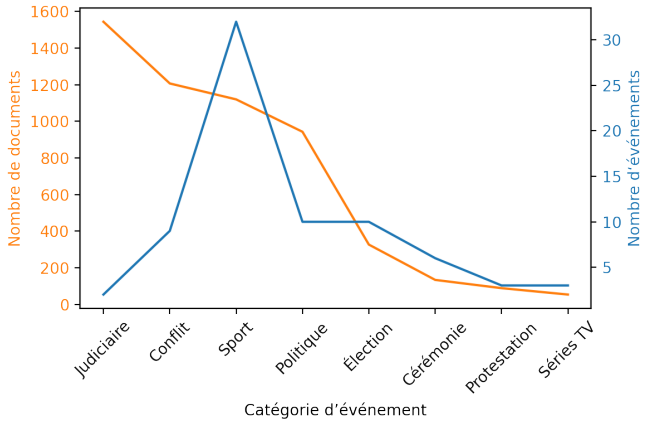


FIGURE 5.2 – Types d'événements annotés par des identifiants Wikidata au sein du corpus *Event Registry*.

Pour cette tâche, nous avons créé une application *ad hoc* (figure 5.3) qui présente les titres des articles associés à un événement, dans chacune des trois langues d'*Event Registry*. Elle est conçue pour être utilisée par un groupe extérieur qui évalue nos propres annotations. L'évaluateur ou l'évaluatrice commence par prendre connaissance de l'événement. Notre application propose la même interface que celle développée au sein du projet *TDT* [All02b] où chaque sujet était décrit par une fiche spécifique. Une fois qu'il ou elle estime avoir compris l'événement dont il est question, une page Wikipédia lui est présentée. Elle est liée à l'événement par son identifiant Wikidata que nous avons annoté manuellement. La première action d'évaluation consiste à déterminer si l'identifiant Wikidata et les articles décrivant l'événement sont liés. On valide ou invalide ensuite les deux annotations binaires selon que l'événement est réel et spécifique sur Wikidata ou non.

Nous fixons des contraintes pour la constitution d'une telle campagne d'évaluation. Les évaluateurs et évaluatrices doivent être natifs en anglais, espagnol et allemand, avec si possible une compétence bilingue pour ces trois langues. Ils doivent comprendre précisément les textes, les éventuelles subtilités stylistiques et identifier des annotations erronées. Il peut s'agir par exemple de documents qui ne rapportent pas le même événement dans les deux ou trois langues de rédaction. Nous proposons enfin, comme cela a été fait auparavant par Mele et coll. [MBC19], d'estimer la qualité des évaluations en présentant régulièrement des événements pour lesquels les annotations sont sûres.

5.3.3 Synthèse

Nous avons annoté plus de 8 000 articles d'*Event Registry* avec des identifiants Wikidata et enrichi le jeu de données existant pour permettre une recherche basée sur des événements. La problématique d'annotation de données est soulevée, de même que son évaluation. Les annotations qui en résultent sont partagées librement [Ber22j] pour

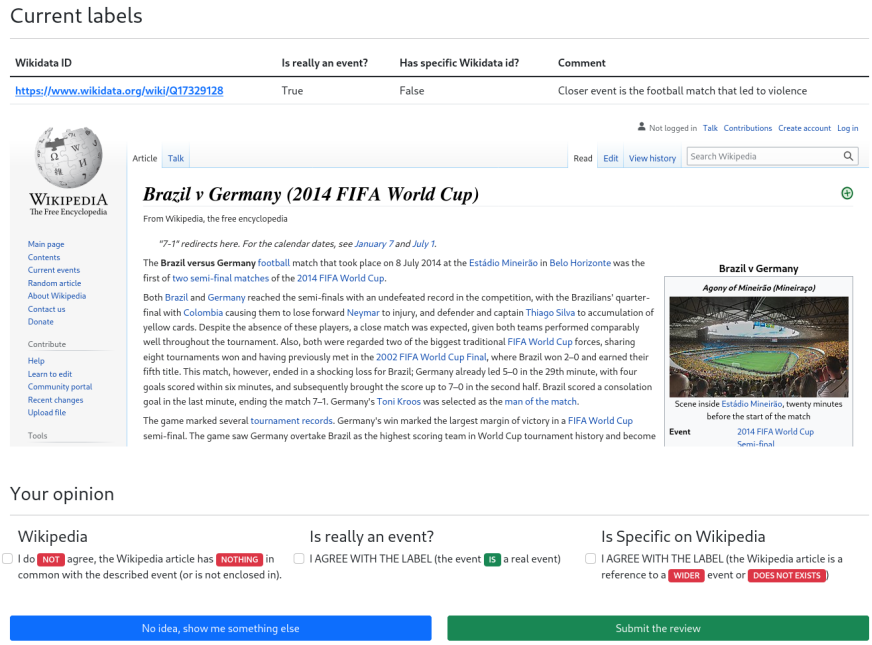


FIGURE 5.3 – Application d'évaluation des annotations des événements.

permettre à chacun de se les approprier. Les analyses de ces données sont également disponibles aux côtés de l'outil d'annotation et d'évaluation présenté précédemment [Ber22c]. Pour cette thèse, nous n'avons pas pu recruter d'évaluateurs ou évaluatrices pour certifier nos annotations. Afin de fournir de premiers résultats expérimentaux, nous considérons que nos annotations peuvent être utilisées pour évaluer la méthode présentée dans ce chapitre.

5.4 Moteur de recherche d'événements

La seconde méthode de suivi d'événements proposée dans le cadre de ce projet de thèse fonctionne sur les principes d'un moteur de recherche, mais adapté aux événements. C'est la boîte noire que nous avons mentionnée précédemment, l'outil qui établit un lien entre un événement nommé donné et l'ensemble des documents qui le rapportent dans la presse. Nous avons vu dans la section 5.1 une méthode d'extraction d'événements qui sert de base à ce processus : c'est la représentation qui est utilisée pour forger la requête. L'autre problématique est celle de l'infrastructure de ce moteur de recherche. Son objectif est de fouiller le corps des articles (stockés selon le format présenté à la section 3.5) pour identifier ceux qui correspondent à l'événement. Pour indexer et rechercher les documents, nous utilisons le moteur Apache LUCENE [Fou22a]. Nous utilisons l'une de ses surcouches au sein du projet *Elasticsearch* [BV22]. Cette architecture est celle utilisée dans des projets récents de fouille de texte dans des documents [WJY22 ; Wan+21a]. Le processus complet est schématisé en figure 5.4.

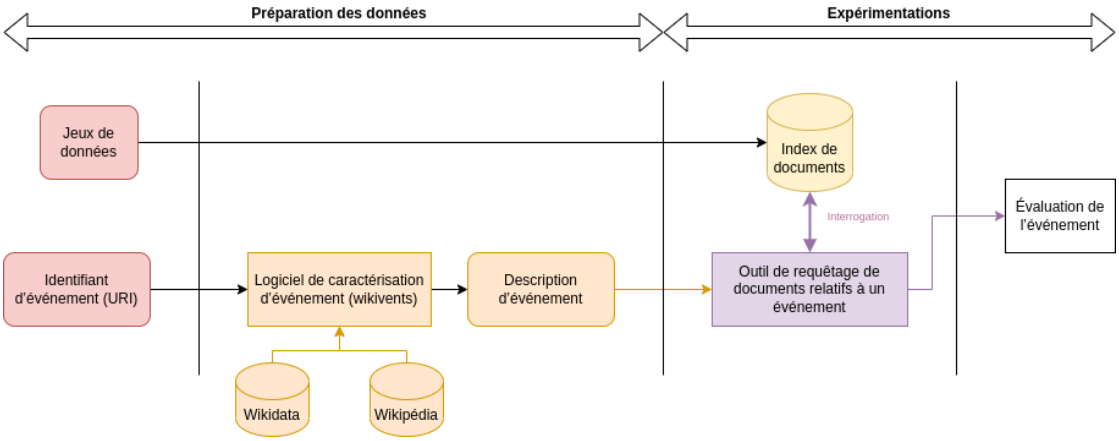


FIGURE 5.4 – Description du processus de traitement global, depuis l'identifiant de l'événement jusqu'à l'évaluation de la qualité de détection des documents.

Un tel moteur de recherche est adapté pour l'analyse d'événements historiques. Dans des travaux de recherche en sociologie ou en histoire par exemple, la consultation numérique d'archives est une activité récurrente. La conception et l'utilisation d'un moteur de recherche réduiraient les temps dédiés à la quête de documents pour les reporter sur des temps d'analyse, de mise en contexte ou autre. En premier lieu, nous avons créé une preuve de concept (figure 5.5) démontrant la faisabilité technique de cet outil et dont l'interface est adaptée à ce public en humanités numériques et à ce besoin de rechercher à partir d'événements. Le point d'entrée est l'identifiant de l'événement. L'outil affiche sa description, des informations sur les entités qui y sont impliquées et tous les articles de presse qui y sont liés. Les documents proviennent du corpus *NewsEye* (section 3.2).

La formulation de la requête est l'élément critique d'un moteur de recherche. Celle-ci doit contenir les informations nécessaires et aussi précises que possible pour sélectionner le maximum de documents pertinents sur l'ensemble de ceux disponibles, tout en limitant l'introduction d'articles qui ne sont pas directement associés à l'événement recherché. Pour déterminer la meilleure requête possible, nous proposons d'en tester plusieurs et d'identifier celle donnant les meilleurs résultats en les évaluant et en les comparant.

5.4.1 Création d'une requête à partir d'un événement

Au sein des documents indexés dans la base de données d'*Elasticsearch*, les dates de publication des articles, le texte ainsi que les entités nommées sont interrogeables. Le but des requêtes est d'exploiter ces trois types de données à partir de la représentation de l'événement dans une langue spécifique. La date de l'article et de l'événement sont comparables, les entités nommées mentionnées sur Wikidata et Wikipédia retrouvables au sein des textes des articles. Le moteur de recherche textuel interroge et analyse alors directement ce texte.

En outre, trois éléments complètent ce mécanisme d'interrogation basique. Une re-

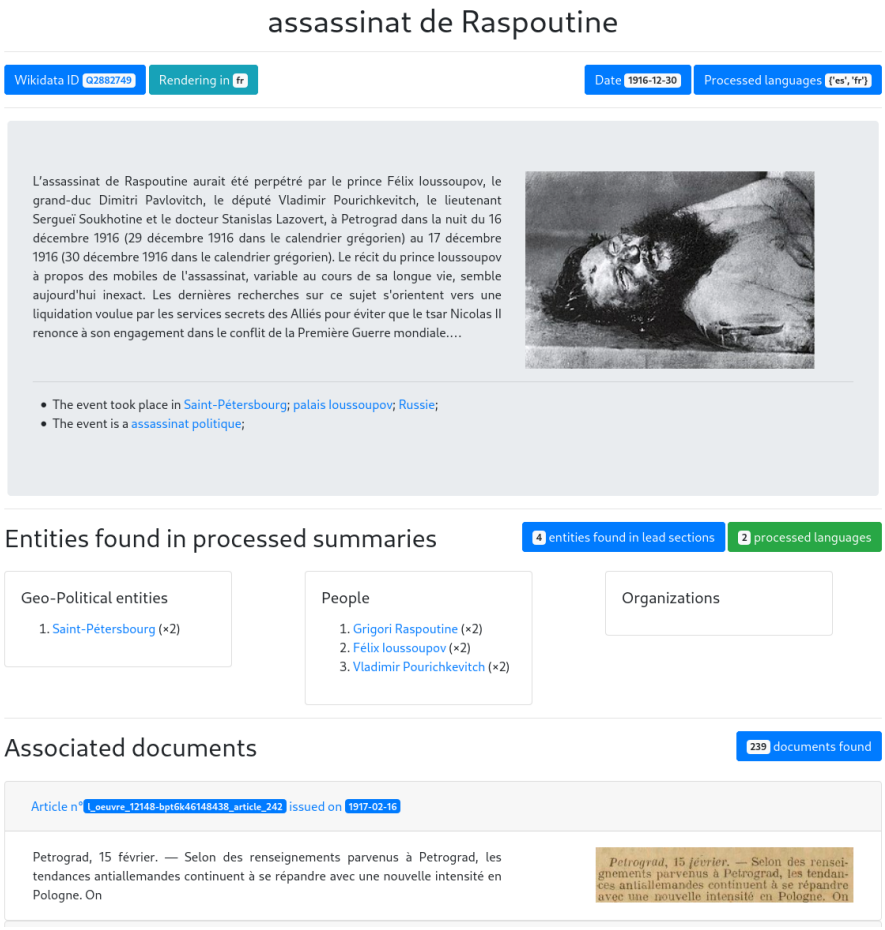


FIGURE 5.5 – Moteur de recherche basé sur les événements

cherche floue est utilisée lorsqu'on tolère une certaine distance d'édition entre le terme de la requête et le terme recherché dans les textes. La distance d'édition est celle de Levenshtein [Lev66]. Dans *Elasticsearch*, ce paramètre dépend de la longueur du terme : en dessous de deux caractères, les termes doivent correspondre exactement, contenir une erreur au maximum en dessous de cinq caractères et deux sont tolérées au-delà. On pondère aussi les termes de la recherche pour donner plus de poids à certains au détriment d'autres.

La requête est une association de plusieurs termes de recherche issus de la description de l'événement, qui peuvent être pondérés selon l'importance donnée au terme recherché. Par exemple, pour poursuivre avec l'assassinat de Raspoutine, la requête limite la recherche à des documents publiés à partir du 30 décembre 1916, jour de l'événement. Un assassinat est un événement brutal, sans prémisse, il n'est pas utile de rechercher de documents antérieurement à son occurrence. Pour celui-ci, « Saint-Petersbourg », « Grigori Raspoutine » ou « Félix Youssoupov » sont des entités participantes et doivent par

conséquent faire partie des termes à rechercher dans les articles. Le paramètre de pondération module l'importance de chaque terme d'après son poids relatif tel que décrit en section 5.1.

Ces paramètres et le contenu utilisé (date de l'événement, entités participantes) influent sur les résultats de chaque requête. Pour évaluer l'importance de chacun, nous pouvons différencier trois catégories de requêtes. L'objectif est d'identifier celles les plus pertinentes et efficaces selon le contexte linguistique ou le type et le niveau de dégradation des documents. La première catégorie de requêtes évalue l'importance des entités participantes qui sont recherchées dans le corps du texte de chaque article. S'il l'événement est désigné par un nom (par exemple « Manifestations d'Euromaïdan »), celui-ci est également utilisé. Une seconde catégorie teste les différentes options possibles pour la pondération de termes de recherche et de la tolérance aux erreurs par des recherches floues. Une première pondération, statique, fixe des valeurs selon le type des entités : les lieux (LOC) et entités géopolitiques (GPE) ont un poids plus important que les personnes (PER). Cette hypothèse s'inscrit dans la lignée des définitions d'événements : les lieux et les institutions discrimineraient davantage les événements que les personnes [ALJ00]. L'autre pondération est dynamique et se base sur les décomptes d'entités participantes calculés par `wikivents`, présentée en section 5.1. La troisième et dernière catégorie joue avec les paramètres de dates et leur combinaison. L'usage de recherches floues introduit davantage de documents non pertinents que désiré. Le choix est fait de ne pas exploiter cette possibilité et de ne rechercher que des termes exacts, sans tolérance aux erreurs.

Il en résulte une requête donnant systématiquement de meilleurs résultats que les autres : les entités nommées de l'événement sont recherchées dans les titres et corps des articles, la pondération des entités est dynamique et basée sur l'importance de chacune et les termes exacts sont recherchés. L'implémentation fait également l'objet d'une publication [BB22c]. L'évaluation des requêtes se fait en analysant la liste des documents renvoyés par le moteur de recherche. Elle est nécessaire pour déterminer la meilleure requête et conduire l'ensemble des expériences sur les corpus de données.

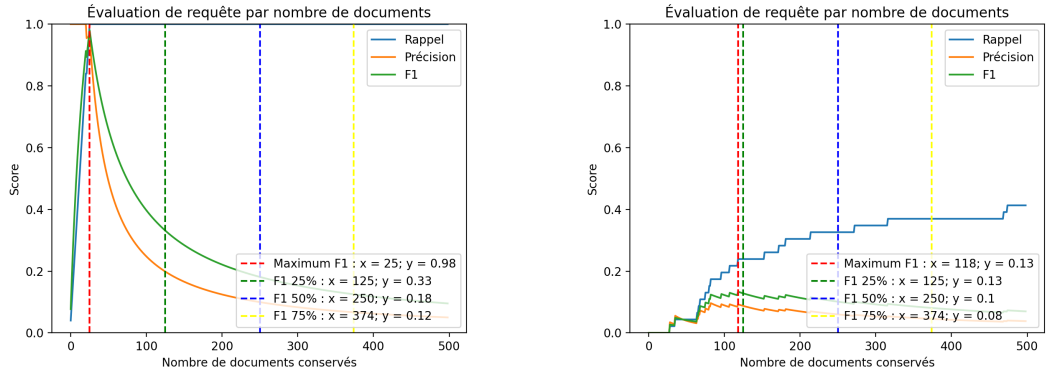
5.4.2 Évaluation des résultats de la recherche

Pour chaque événement donné est renvoyée par le moteur de recherche une liste de documents. Pour toute requête et chaque document est calculé un score de pertinence basé sur la méthode probabiliste *Okapi BM25* [Rob+97], très utilisée en recherche d'information. Les documents sont triés d'après ce score, du plus au moins pertinent.

Dans ce contexte, la précision et le rappel dépendent logiquement du nombre de documents retenus. Ces scores sont reportés dans la figure 5.6a, en fonction du nombre de documents sélectionnés. Il est notable sur ce graphique que le rappel croît au fur et à mesure que des documents sont analysés (maximum atteint pour cent articles), et la précision diminue au-delà, liée à l'introduction de faux positifs. Nous considérons que, dans notre cas d'utilisation, c'est l'expertise de l'utilisateur ou utilisatrice créant la requête qui détermine à partir de quel document les données ne sont plus pertinentes. Par analogie, l'expertise susmentionnée est représentée par le maximum de la *F1* pour

chaque événement : c'est cette information que nous conservons.

Les deux résultats de requêtes en figure 5.6a et 5.6b sont clairement opposés. Pour le premier, la mesure *F1* maximale est très haute vers 25 articles sélectionnés. L'événement est bien décrit, par une date et de nombreuses entités participantes. Dans l'autre cas, très peu d'informations sont extraites de la représentation de l'événement, décrit par seulement quelques entités nommées. Ce sont aussi des entités très courantes, ici « Barack Obama », très présent dans d'autres documents. La recherche doit être affinée par d'autres entités pour être plus pertinente. La description de l'événement influe en premier lieu sur les résultats. Sans données, il est impossible de créer une requête de qualité, et donc d'obtenir les résultats désirés.



(a) Événement `Q1190093` en anglais.

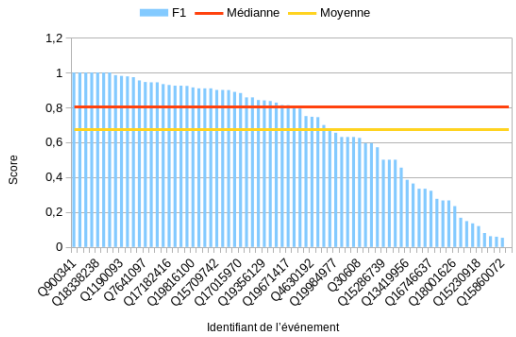
(b) Événement `Q15894506` en anglais.

FIGURE 5.6 – Évaluation de la précision et du rappel des résultats de la requête pour un événement donné.

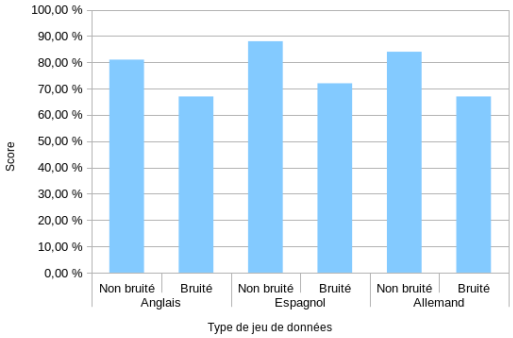
Dès lors, chaque évaluation est spécifique à un événement. Pour avoir un aperçu global du fonctionnement d'une requête sur la totalité d'un jeu de données, et pour les 81 événements annotés avec un identifiant sur Wikidata, il faut synthétiser ces évaluations. Nous proposons d'utiliser la médiane, moins sensible que la moyenne aux valeurs aberrantes. L'évaluation globale du système est la médiane de toutes les valeurs maximales de *F1* calculées pour chacun des événements, présentée en figure 5.7a.

Seuls 25 % des documents sont annotés, mais la totalité peut être utilisée. Pour déterminer l'efficacité du processus, nous distinguons deux situations. Dans la première, seuls les documents annotés sont utilisés, c'est le contexte que l'on nomme « non bruité ». L'autre, dit « bruité », contient en plus des documents annotés tous les autres articles d'*Event Registry*. Par cette technique, on évalue la qualité des résultats dans un corpus de plus grande taille, plus proche de la réalité.

L'évaluation de référence (figure 5.7b) sur le corpus *Event Registry* non dégradé, non segmenté donne une médiane de *F1* à 81 % en anglais, 88 % en espagnol et 84 % en allemand. Si l'on tient compte en plus des documents non annotés (le « bruit »), les résultats sont réduits à 67 % en anglais, 72 % en espagnol et 67 % en allemand.



(a) Profil des *F1* maximales par événement en anglais, non bruité.



(b) Médiane des *F1* maximales par jeu de données.

FIGURE 5.7 – Résultat de l'évaluation sur les 81 événements du corpus *Event Registry*.

5.5 Recherche d'événements dans la presse ancienne

Tout comme les expériences menées au chapitre 4, celles conduites ici se basent sur le jeu de données *Event Registry*. À cause de limitations posées par l'annotation des événements (section 5.3), les corpus *CoAID* et *FibVid* ne sont pas exploitables. L'objectif des expériences est de tester 81 requêtes, une pour chaque événement annoté et de collecter des résultats pour tous les jeux de données, dégradés ou non, segmentés ou non. Tout comme précédemment, nous proposons d'évaluer l'impact des dégradations introduites par le mécanisme *OCR* ainsi que celles liées au phénomène de sursegmentation du texte.

Conjointement à ce chapitre, les résultats de toutes les expérimentations sont publiés [BB22b]. Nous avons évoqué précédemment plusieurs catégories de requêtes qui ont mené à la création d'une dernière, plus efficace que toutes les autres. Les détails d'implémentation de ces dix requêtes intermédiaires, et les résultats obtenus sont également partagés publiquement pour faciliter la reproduction et l'extension de ces expériences.

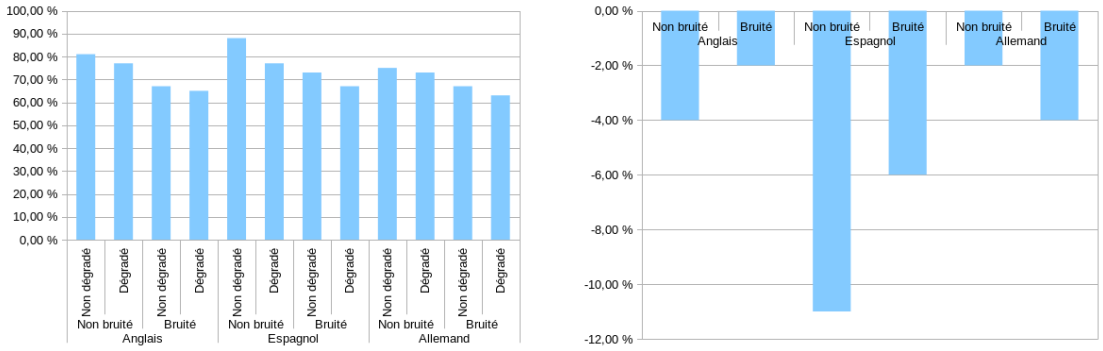
Question ressources, l'infrastructure fonctionne sur un ordinateur doté d'un processeur *Intel® Core™ i7-7820HQ CPU @ 2.90 GHz*. Le processus gérant l'index de documents basé sur *Elasticsearch*, celui qui exécute la requête, est limité à une consommation de 4 Go de mémoire vive. Les temps de traitement pour chaque requête n'excèdent jamais les cinq secondes, pour trente mille documents indexés. L'infrastructure est également reproductible par le biais de conteneurs logiciels [BB22a]. La durée de traitement la plus importante est liée à l'analyse par `wikivents` des articles Wikipédia et des pages Wikidata. Ces données sont mises en cache pour économiser à la fois de la bande passante et de la charge processeur, cette dernière délocalisée sur les serveurs de la fondation Wikimédia via les appels *API*. Cet environnement expérimental intègre bien les contraintes fixées de limitation de la consommation des ressources informatiques.

Nous allons analyser successivement, comme pour le chapitre 4, les résultats des expériences pour *Event Registry* d'abord non segmenté puis les versions segmentées en deux puis en trois portions. Enfin, nous comparerons les dégâts causés par les segmentations

sur la performance générale du système.

5.5.1 *Event Registry*

Les résultats bruts des expérimentations sont présentés dans la figure 5.8. Les valeurs pour chaque catégorie sont les médianes des *F1* maximales de tous les événements recherchés.



(a) Résultats bruts.

(b) Exécution sur articles dégradés contre articles non dégradés.

FIGURE 5.8 – Résultats sur *Event Registry* dans les trois langues, en fonction du niveau de dégradation.

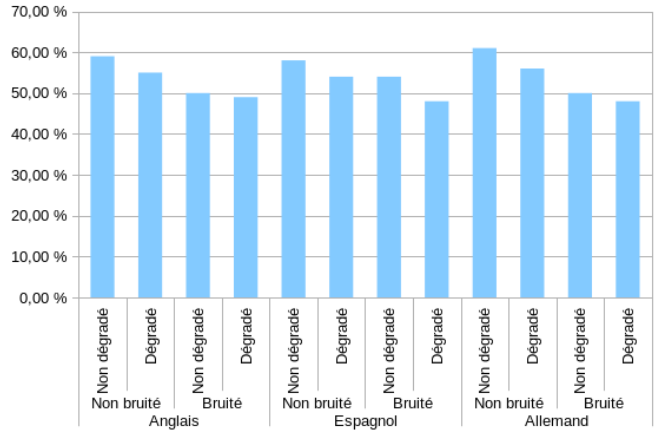
La figure 5.8 présente des résultats plutôt bons dans les trois langues, avec 80 % de réussite en anglais, 88 % en espagnol et 75 % en allemand. Si l'on compare ces résultats à ceux obtenus sur le corpus *Event Registry* d'origine (figure 5.7b), on note d'abord une très faible dégradation pour l'anglais (un point) et plus importante pour l'allemand (neuf points). Pour rappel, la version non dégradée du corpus contient des textes analysés par *OCR*, mais pour lesquelles les images sources n'ont subi aucune des dégradations présentées en section 3.4. Par l'ajout de bruit dans ces données, c'est-à-dire les 75 % de documents d'*Event Registry* non annotés d'événements, les résultats se dégradent logiquement comme expliqué précédemment. Ils sont respectivement de 67 %, 73 % et 67 % pour l'anglais, l'espagnol et l'allemand. Ce sont les mêmes résultats que pour la variante bruitée du jeu de données original, sans traitement des textes par *OCR* ni segmentation.

Les dégradations par *OCR* ont un impact différent en fonction des langues : -4 points pour l'anglais, -11 pour l'espagnol et -2 pour l'allemand. Leur impact est plus marqué que celui observé au chapitre 4 où les expériences dans le même contexte menaient à une diminution des résultats de l'ordre de 1 point en moyenne. Aucun modèle ne se dégage dans la figure 5.8b : certaines dégradations sont plus importantes dans un contexte bruité (allemand) que dans l'autre (anglais et espagnol). Les requêtes fonctionnent en recherchant les termes exacts au sein des textes (comme expliqué précédemment, l'utilisation des recherches floues peut dégrader les résultats), des erreurs introduites par

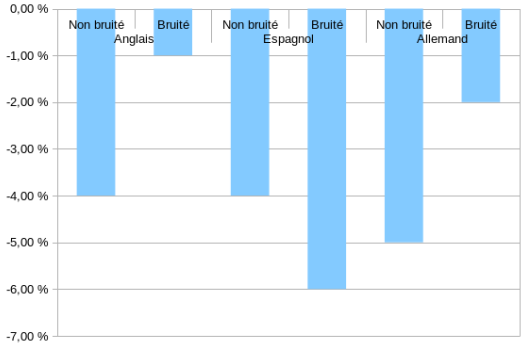
l'*OCR* à certaines positions critiques peuvent expliquer la forte dégradation visible pour l'espagnol, comparée à la dégradation dans les deux autres langues.

5.5.2 *Event Registry*, segmenté en deux

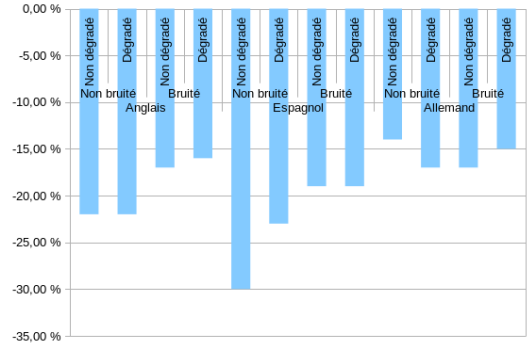
Les résultats bruts des expérimentations sont présentés dans la figure 5.9. Les valeurs pour chaque catégorie sont les médianes des *F1* maximales de tous les événements recherchés.



(a) Résultats bruts.



(b) Exécution sur articles dégradés contre articles non dégradés.



(c) Comparaison des résultats non segmentés avec *Event Registry* segmenté en deux.

FIGURE 5.9 – Résultats sur *Event Registry* segmenté en deux, dans les trois langues, en fonction du niveau de dégradation.

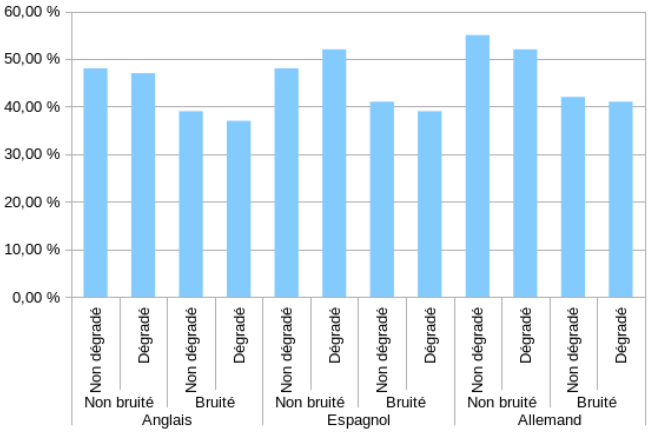
Avec un premier niveau de segmentation, c'est-à-dire des textes divisés en deux portions, la qualité globale des résultats se dégrade, comme présenté en figure 5.9a. La segmentation a un impact très fort, avec des diminutions d'au moins 15 points, et jusque 30 pour l'allemand, non bruité et non dégradé. L'ajout de « bruit » dans les données ne semble pas avoir davantage de conséquences. Ces derniers sont plus proches de la réalité

de notre cas d'utilisation. Par conséquent, seule l'évaluation dans les données « bruitées » fait sens. La dégradation est d'entre 15 et 20 points pour chacune des langues, des niveaux de diminution que l'on observe également en sous-section 4.3.1 pour des documents segmentés en deux avec les algorithmes de *clustering* du chapitre 4.

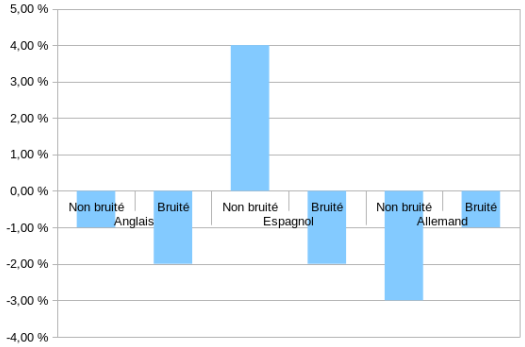
La baisse de qualité des résultats causée par les erreurs d'*OCR* est plus faible que pour le corpus non segmenté, quel que soit le contexte ou la langue. La segmentation est donc le facteur le plus important, comparé aux erreurs d'*OCR*. Les niveaux de dégradations sont toutefois toujours relativement élevés par rapport aux expériences menées au chapitre 4.

5.5.3 *Event Registry*, segmenté en trois

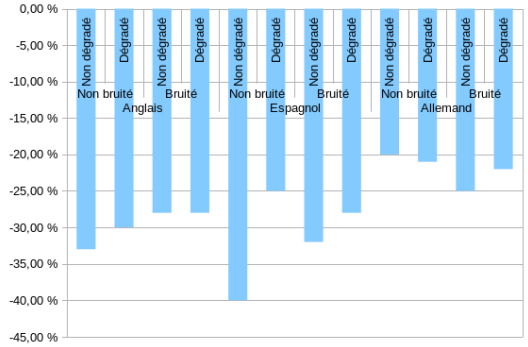
Les résultats bruts des expérimentations sont présentés dans la figure 5.10. Les valeurs pour chaque catégorie sont les médianes des *F1* maximales de tous les événements recherchés.



(a) Résultats bruts.



(b) Exécution sur articles dégradés contre articles non dégradés.



(c) Comparaison des résultats non segmentés avec *Event Registry* segmenté en trois.

FIGURE 5.10 – Résultats sur *Event Registry* segmenté en trois, dans les trois langues, en fonction du niveau de dégradation.

Le second niveau de segmentation qui implique une division des articles de presse en trois parties pose logiquement des problèmes au moteur de recherche documentaire. Les résultats présentés à la figure 5.10 sont plus faibles en général, ne dépassant pas les 60 % de réussite. C'est légèrement inférieur, de quelques points, aux résultats obtenus par les algorithmes du chapitre 4 qui donnaient, pour ce niveau de segmentation, des résultats de *F1 BCubed* (cette métrique est présentée en section 4.3 à la page 112) dans toutes les langues approchant les 70 %.

La tendance évoquée précédemment se confirme : à partir d'un certain niveau de segmentation du texte, l'effet des erreurs introduites par l'*OCR* est négligeable, comme indiqué dans la figure 5.10b. Entre les versions dégradées et non dégradées des documents, la baisse de performance de l'outil est de l'ordre de quelques points, moins de deux en général. Sur le corpus non bruité et non dégradé en espagnol, c'est l'inverse qui se produit : sur les textes dégradés, l'outil fonctionne mieux à ce niveau de segmentation que sur les textes non dégradés.

Comparée aux données non segmentées d'*Event Registry* et tout comme dans les autres expérimentations à ce sujet, la segmentation en trois portions a l'impact le plus fort sur les résultats du moteur de recherche. Là où la segmentation en deux entraîne des baisses de l'ordre de 15 points, une segmentation en trois diminue l'efficacité du système de dix points supplémentaires avec une réduction quasi générale d'au moins 25 points. Les dégradations introduites par *OCR* ne semblent pas se surajouter aux diminutions causées par la segmentation du texte, leur impact est de plus en plus faible au fur et à mesure des segmentations successives.

5.6 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle approche de suivi de mentions d'événements basée sur un moteur de recherche documentaire. Celle-ci est conçue avec l'objectif de diminuer les ressources nécessaires pour opérer un suivi d'événements, contrainte à laquelle nous nous sommes plié tout au long de ce chapitre.

Un moteur de recherche basé sur des événements se doit d'utiliser des descriptions presque exhaustives de ces événements. Nous proposons un processus, implémenté dans un outil librement partagé, `wikivents` [Ber20] qui fournit des réponses aux questions de base sur un événement (quoi, quand, où et qui) à partir de sources publiques comme Wikidata et Wikipédia. Nous avons en outre montré que les langues pivots ne doivent pas être utilisées seules pour décrire un événement : mobiliser les langues vernaculaires permet de caractériser plus finement un événement historique. Pour évaluer cette approche expérimentale, des annotations associant les articles de presse d'*Event Registry* étaient nécessaires. Nous avons réalisé ce travail, en associant à chaque article l'identifiant Wikidata de l'événement auquel il est rattaché. En première intention, nous avons utilisé ce jeu de données. Il fait l'objet d'une campagne d'évaluation au moment où ces lignes sont écrites. Nous avons enfin présenté le fonctionnement du moteur de recherche, en insistant sur la nécessité de définir des requêtes précises, et expérimenté son fonctionnement sur le corpus *Event Registry* et toutes ses variantes dégradées que nous avons

synthétisées en section 3.4.

L'objectif de ce chapitre était d'évaluer l'utilité d'un tel moteur de recherche documentaire, et l'impact des dégradations que présentent les documents historiques (*OCR*, sursegmentation) sur la performance de cet outil. Nous tirons plusieurs conclusions de ces travaux.

— Le principe de moteur de recherche documentaire fonctionne : certes, il est moins fiable que peuvent l'être des algorithmes entraînés ou non, tels que présentés au chapitre 4. Cependant, il ne nécessite aucun entraînement et est peu consommateur de temps de processeur ou de mémoire vive. Dans notre cas, nous avons fixé des limites à l'exploitation des ressources de l'ordinateur pour valider cet aspect. L'infrastructure répond systématiquement en des temps raisonnables, inférieurs à quelques secondes. Pour des bases de données de millions d'articles, comme notamment le corpus *NewsEye*, c'est probablement le seul outil capable de fournir des réponses rapides et avec des résultats relativement précis tout en limitant les ressources consommées.
— Les dégradations de contenu introduites par l'*OCR* ont un faible impact sur les capacités du moteur de recherche à fournir des résultats pertinents.
— Les dégradations introduites par la segmentation du texte sont cruciales. Ce sont elles qui ont l'impact le plus important sur la baisse de qualité des résultats du moteur de recherche, menant à négliger l'impact des erreurs *OCR* d'autant que la segmentation est importante. À partir d'un certain niveau de segmentation, les articles de presse sont tellement divisés qu'il est complexe d'y retrouver toutes les entités qui décrivent les événements (par quoi, quand, où et comment). Ces informations sont éclatées dans les différents segments de texte, complexifiant la recherche.

Tout comme précédemment, une limite posée à cette analyse est la quantité d'articles annotés par des événements. Pour tendre vers un contexte réel, nous n'avons annoté que 25 % des documents soit 114 événements, dont les six plus importants et les autres choisis aléatoirement. Pour permettre une meilleure comparaison avec les autres algorithmes présentés au chapitre 4, il serait idéal de disposer d'annotations complètes sur ce jeu de données, mais cela reviendrait à rechercher 1 400 identifiants Wikidata pour chacun des événements du corpus *Event Registry*.

Les débouchés de ces travaux sont prometteurs, à la fois en termes de résultats (sur des documents non dégradés, environ 70 % de précision) et en termes de réduction de l'empreinte du numérique. Cette voie de recherche ouverte, nous proposons d'explorer plus avant cette mécanique pour la fiabiliser et l'évaluer plus intensivement. L'ajout de nouvelles bases de connaissance comme *EventKG* [GD19 ; AGD20] permettrait d'obtenir des descriptions d'événements encore plus fiables. Il serait souhaitable de les ajouter au sein de `wikivents` [Ber20]. De nouvelles expérimentations devraient être menées dès lors que de nouveaux jeux de données répondant aux contraintes fixées dans section 2.3 sont publiés. Il s'agit de continuer à améliorer ce processus tout en contribuant à réduire l'impact du numérique sur notre environnement [FDM13 ; 21a ; Bon+22].

Chapitre 6

Conclusion et perspectives

Sommaire

Pour conclure les travaux menés durant cette thèse, nous allons rappeler tout d'abord l'ensemble des problématiques étudiées et présentées en introduction de ce document. Ces travaux ont fait émerger des propositions nouvelles. Ils débouchent également sur des contributions scientifiques et techniques. Enfin, nous les mettrons en perspective avec des travaux futurs et identifierons leurs avantages ainsi que leurs limites.

6.1 Les problématiques traitées

La détection et le suivi de mentions d'événement dans la presse ancienne recoupent tout un ensemble de problématiques. Celles-ci intègrent la notion d'événement et la question du suivi que nous avons mentionnées au chapitre 2. Nous isolons trois approches : la première est basée sur l'analyse des textes à la recherche d'une méthode de *clustering* efficace. Une autre fonctionne de façon analogue à un moteur de recherche documentaire et une dernière utilise les graphes pour représenter la propagation de l'information. Ce sont les deux premières solutions que nous avons utilisées dans le cadre de cette thèse. La notion d'événement est elle-même assez floue et nous nous sommes concentré sur une approche simple, liée à une représentation ontologique des événements. Elle décrit un événement comme une « action qui se déroule dans un lieu donné, à un instant donné et qui fait intervenir des entités ». Cette définition répond à des questions de base sur l'événement : quelle est l'action qui a eu lieu, quand s'est-elle déroulée, ou/et qui y a pris part. La pertinence de ce raisonnement provient du style d'écriture journalistique de la brève de presse ou de la dépêche, tels les télégrammes que nous avons présentés en chapitre 3. Les jeux de données annotés selon cette définition d'événement sont rares,

mais nécessaires pour tout processus d'évaluation expérimental. Pour anticiper les biais dans l'analyse des méthodes présentées dans les chapitres suivants, nous produisons en état de l'art une analyse des jeux de données retenus.

Dans le chapitre 3 nous avons mis en avant la complexité et les spécificités des documents de presse historique numérisée. Les altérations physiques des supports en papier entraînent des dégradations numériques : la retranscription des textes des documents est fatalement imparfaite tout comme la reconnaissance des articles. Parmi plusieurs projets, *NewsEye* a contribué à faire progresser la recherche sur ces aspects. Cependant, nous ne pouvons ignorer que les erreurs induites par *OCR* et par sursegmentation du texte posent un problème crucial pour le suivi d'événements dans la presse historique. C'est en prévision de travaux futurs que nous proposons d'évaluer les méthodes des chapitres 4 et 5 sur des documents dégradés de ce type. Pour ce faire, nous produisons des variations de tous les jeux de données en générant artificiellement des erreurs typiquement induites par *OCR* et par sursegmentation du texte.

Nous présentons dans les chapitres 4 et 5 deux méthodes différentes de suivi d'événements. La problématique traitée est toujours la même et correspond à un cas d'utilisation que nous avons utilisé comme fil rouge de cette thèse. Comment, en tant qu'historien, historienne ou sociologue qui analysent des événements dans des articles de presse, peut-on obtenir tous les documents relatant un même événement historique ?

La première solution que nous proposons en chapitre 4 est d'utiliser des algorithmes de suivi adaptés à la presse nativement numérique. Nous les appliquons aux documents historiques numérisés. Ils conviennent aux contextes où les événements potentiellement mentionnés dans les textes sont inconnus ou que l'on cherche à faire émerger des événements non référencés jusqu'alors (comparables à des signaux faibles). Ces processus fonctionnent en encodant les textes des articles de presse en vecteurs utilisés par des outils de *Machine Learning*. Nous avons proposé deux algorithmes, adaptés à deux situations différentes. Le premier est supervisé, basé sur les travaux d'autres auteurs [Mir+18] et utile lorsque des données d'entraînement sont disponibles et que des résultats doivent être obtenus rapidement. L'autre solution propose d'exploiter un algorithme de K-Moyennes [Mac67], adapté au suivi d'événements dans la presse. Nous avons analysé les comportements de ces algorithmes en fonction du type d'encodage utilisé (ceux obtenus par pondération *TF-IDF* ou dense, calculés à partir de modèles d'apprentissage profond). Nous avons également comparé les capacités de ces procédés selon les différents niveaux de dégradation des documents, étudiant l'impact de chacun d'eux.

L'autre alternative est la conception d'un moteur de recherche d'événements, présenté au chapitre 5. Adapté aux situations où les événements sont connus et où l'on cherche tous les documents qui s'y rapportent, il fournit de bons résultats en général, à la condition de renseigner une description de l'événement suffisamment précise. Afin d'obtenir cette description pour un cadre expérimental, nous avons développé un outil capable d'extraire une représentation d'événement historique à partir de connaissances publiées sur Wikidata et Wikipédia. Cette description nourrit le moteur de recherche pour formuler une requête qui renvoie tous les documents mentionnant cet événement. Tout comme pour le chapitre 4, nous avons analysé l'impact des différentes dégradations

sur les capacités de cet outil.

Dans les deux cas, nous proposons une série de recommandations pour prolonger ce travail. Elles concernent l'amélioration des algorithmes, les contextes dans lesquels chacun est le plus efficace, leurs limites et les solutions qu'elles permettent d'envisager.

6.2 Contributions

Ces travaux de thèse se concrétisent avec plusieurs productions, à la fois scientifiques et techniques. Nous avons, tout au long de ce projet de recherche, travaillé sur des corpus de documents de presse, les avons analysés et utilisés pour construire des expériences. Systématiquement, la transparence et la publicité des corpus, des données calculées et des algorithmes étaient au cœur des problématiques de ce projet scientifique.

Les contributions principales de cette thèse sont, parmi d'autres :

— **Publication d'algorithmes de suivi de mentions d'événements**. L'état de l'art décrit des algorithmes de suivi de mentions d'événements pour la presse nativement numérique et peu d'implémentations sont librement disponibles et réutilisables. Ce problème est très probablement lié à la forte valeur ajoutée qu'ont ces outils en ce qui concerne la veille documentaire, avec des applications dans les milieux bancaires [LH20] ou journalistiques. Pour ces publics, ces outils offrent un avantage concurrentiel sur leur marché. Ce travail de thèse enrichit également l'état de l'art d'une méthode supplémentaire, un moteur de recherche adapté à l'analyse de documents historiques. L'ensemble des algorithmes conçus et tous les outils intermédiaires mentionnés dans ce document sont publiés.

— **Une première étude d'impact de ces algorithmes pour les documents historiques**. Les algorithmes et méthodes de l'état de l'art sont conçus et testés pour la presse nativement numérique et récente. Rien ne permet de présager de leurs comportements en présence de documents historiques. En effet, ces derniers sont altérés, la langue dans laquelle ils sont écrits a évolué, l'orthographe ou les modes de diffusion de l'information également. Nous avons produit la première analyse, à notre connaissance, de l'impact de deux types de dégradations (induites par *OCR* et sursegmentation) sur le suivi de mentions d'événements dans du texte. Avec ces résultats, nous proposons une série de recommandations pour améliorer le suivi de mentions d'événements dans la presse ancienne.

— **Une procédure et une analyse de jeux de données exploitables pour le suivi d'événements**. En première partie de ce document, nous avons proposé une analyse des jeux de données *Event Registry*, *CoAID* et *FibVid*. Elle s'intéresse aux défauts à éviter pour la constitution de jeux de données ainsi que soulève de potentiels biais d'analyse que ces données peuvent engendrer. Nous proposons un cadre pour l'analyse de corpus de documents de presse utilisables dans le cadre de travaux similaires.

— **Mise en avant de l'*Open Science* et la reproductibilité des expérimentations**. Ce travail s'inscrit dans la lignée des mouvements pour la science ouverte et des logiciels libres. Les implémentations de tous les algorithmes et de tous les

outils jusqu'à ceux générant les moindres graphiques de cet ouvrage sont librement partagées sur des archives numériques. L'objectif est de valider ou étendre ces travaux de recherche. Les données intermédiaires et corpus utilisés sont également diffusés librement sur l'archive Zenodo [EO13].

6.3 Limites

Ces travaux de thèse ont certes permis d'appréhender le fonctionnement des algorithmes de suivi d'événements dans la presse historiques, mais nous pouvons identifier certaines limites à ce travail.

La première d'entre elles est liée au manque de données. Nous l'avons mentionné à plusieurs reprises, les documents de presse annotés suivant les événements qu'ils mentionnent sont rares et, en l'état actuel de nos connaissances, inexistants pour la presse ancienne. Notre proposition de dégradation synthétique de documents de presse ne rend donc pas compte de la réalité : le style de rédaction, le vocabulaire ou les entités nommées ne sont pas les mêmes que celles de la presse récente. Également, les processus de vectorisation présentés au chapitre 4 apprennent à partir de données récentes. Entraîner les modèles d'apprentissage profond et de pondération avec de la presse ancienne devrait avoir une incidence sur des résultats expérimentaux. Les annotations d'événements Wikidata utilisées au chapitre 5 ne sont, de la même façon, pas validées par un groupe d'annotateurs et d'annotatrices répondant à nos exigences de qualité.

Dans ce dernier chapitre, le chapitre 5, nous avons conçu un moteur de recherche d'événements qui se base sur des descriptions automatiquement collectées à partir de Wikidata et Wikipédia. En fonction de celles-ci, les résultats peuvent différer. Un événement très bien décrit, de dates et d'entités ne devrait pas donner les mêmes résultats qu'un événement mal décrit pour lequel par exemple seul le titre est disponible. La définition d'un « score de qualité d'événement » pourrait être une solution de contournement à ce problème, permettant d'établir un lien potentiel entre la qualité des événements représentés et les résultats fournis par le moteur de recherche.

6.4 Perspectives

Dans cette thèse, nous avons évalué l'impact des dégradations typiques de documents de presse historique, celles introduites par les outils d'*OCR* et de segmentation de contenu. Nous avons évalué l'impact des algorithmes sur des données de référence ainsi que sur des données dégradées. Pour poursuivre ces travaux de recherche, nous proposons les pistes d'amélioration suivantes :

— **Réduction de la sensibilité aux erreurs de segmentation**. Nous l'avons vu, les plus importantes baisses de performance de toutes les solutions envisagées sont liées en premier lieu à la segmentation du texte puis aux erreurs *OCR* dans un second temps. Notre première suggestion est de limiter ces problèmes liés à la sursegmentation. Une première possibilité est de se focaliser sur la réduction de la dégradation à la source, en prenant en compte les dernières avancées dans ce

domaine [Pal+12 ; Mic+21 ; Zhu+22]. Des travaux récents [MWL22] proposent des solutions à la sursegmentation du texte et améliorent l'identification d'articles complets. L'autre solution est de rendre plus robustes à ce type d'erreur les processus développés aux chapitres 4 et 5.

— **Validation de la vérité terrain des annotations d'événements**. En chapitre 5, nous avons annoté des documents du corpus *Event Registry* par des identifiants d'événements Wikidata. Nous avons également fourni un outil d'évaluation de ces annotations. Pour améliorer le moteur de recherche conçu et présenté dans ce chapitre, des travaux supplémentaires sont à prévoir. Davantage d'événements doivent être annotés et validés par un groupe d'annotateurs et d'annotatrices plus conséquent et plus qualifié que celui que nous avons pu mobiliser.

— **Développement d'outils pour les scientifiques en humanités numériques**. Bien que l'objectif ultime de ces travaux soit de fournir des outils d'analyse pour des scientifiques en humanités numériques ou les utilisateurs et utilisatrices de bibliothèques numériques, nous n'avons rien produit qui soit directement utilisable. Nous avons fait le choix de ne pas traiter en priorité l'aspect de mise en production et de développement d'application. Les algorithmes et outils développés ne sont pas utilisables par un public non spécialiste. Une interface personne-machine adaptée doit être conçue pour simplifier l'accès à ces travaux de recherche et à leurs résultats. Cette application peut par exemple projeter sur une carte l'ensemble des articles associés à un même événement et identifier leurs interconnexions. Elles seraient multiples : un article A est publié avant un autre article B, l'article C est partiellement ou totalement inspiré d'un article D, lui-même rédigé dans une autre langue, etc. Nous proposons, en figure 6.1, une proposition d'interface de navigation, issu de la préparation au concours Ma Thèse en 180 secondes.

— **Réduction de l'empreinte numérique globale**. L'une des motivations principales des recherches menées au chapitre 5 est la sobriété numérique à des fins écologiques. Nous y avons proposé une approche qui cherche à diminuer l'usage de ressources numériques, qui limite les consommations de temps de processeur et de mémoire vive. Les futurs travaux de recherche dédiés à l'analyse et au suivi de mentions d'événements pourraient s'en inspirer, avec l'intention de réduire l'empreinte numérique causée par les systèmes développés en recherche.

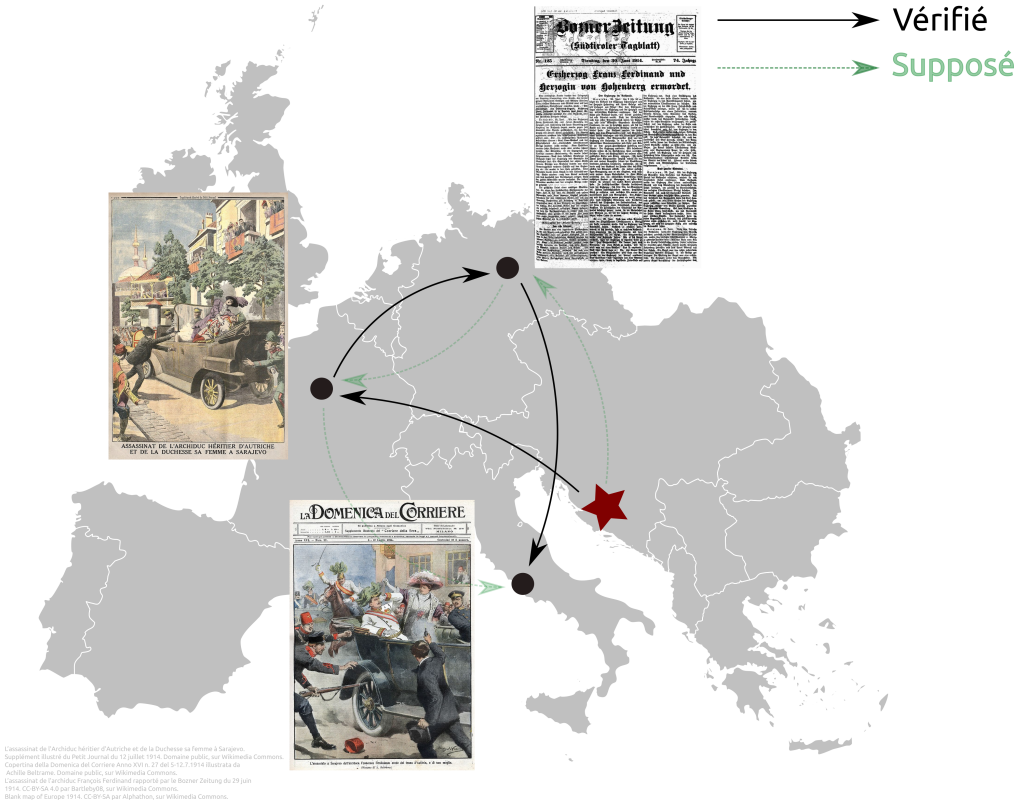


FIGURE 6.1 – Vue possible d'une application de suivi d'événements. Visuel réalisé dans le cadre du concours Ma Thèse en 180 secondes, le 16 mars 2021.

Publications issues de ces travaux

Publications scientifiques

— Guillaume Bernard et al. « A Comprehensive Extraction of Relevant Real-World-Event Qualifiers for Semantic Search Engines ». In : *Proceedings of the 25th International Conference on Theory and Practice of Digital Libraries*. 25th International Conference on Theory and Practice of Digital Libraries. T. 12866. Online, 15 sept. 2021, p. 153-164. DOI : 10.1007/978-3-030-86324-1_19

— Guillaume Bernard et al. « Event Related Document Retrieval with Multilingual Real World Event Representation ». In : *Proceedings of the ISWC 2021 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice Co-Located with 20th International Semantic Web Conference*. 20th International Semantic Web Conference. T. 2980. Online, 27 oct. 2021, p. 5. ISBN : 1613-0073. URL : http://ceur-ws.org/Vol-2980/paper309.pdf

— Guillaume Bernard et al. « Tracking News Stories in Short Messages in the Era of Infodemic ». In : *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. CLEF 2022 Conference and Labs of the Evaluation Forum. T. 13390. Lecture Notes in Computer Science. Bologna, Italy, 5 sept. 2022, p. 18-32. ISBN : 978-3-031-13642-9. DOI : 10.1007/978-3-031-13643-6_2

Jeux de données

— Guillaume Bernard. *Resources to compute TF-IDF weightings on press articles and tweets*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6610406. URL : https://doi.org/10.5281/zenodo.6610406

— Guillaume Bernard. *Event Registry dataset with multiple extracted features (both sparse and dense)*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6630367. URL : https://doi.org/10.5281/zenodo.6630367

— Guillaume Bernard. *CoAID dataset with multiple extracted features (both sparse and dense)*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6630405. URL : https://doi.org/10.5281/zenodo.6630405

— Guillaume Bernard. *Fibvid dataset with multiple extracted features (both sparse and dense)*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6630409. URL : https://doi.org/10.5281/zenodo.6630409

— Guillaume Bernard. *Event Registry titles only dataset with multiple extracted features (both sparse and dense)*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6630447. URL : https://doi.org/10.5281/zenodo.6630447

— Guillaume Bernard. *Event Registry dataset with multiple extracted features (both sparse and dense) and degraded by OCR*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6631267. URL : https://doi.org/10.5281/zenodo.6631267

— Guillaume Bernard. *CoAID dataset with multiple extracted features (both sparse and dense) and degraded by OCR*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.66

30966. URL : https://doi.org/10.5281/zenodo.6630966

— Guillaume Bernard. *FibVid dataset with multiple extracted features (both sparse and dense) and degraded by OCR*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6631070. URL : https://doi.org/10.5281/zenodo.6631070

— Guillaume Bernard. *Event Registry titles dataset with multiple extracted features (both sparse and dense) and degraded by OCR*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6631082. URL : https://doi.org/10.5281/zenodo.6631082

— Guillaume Bernard. *Event Registry dataset texts with OCR degradations and synthesised segmentation*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6631305. URL : https://doi.org/10.5281/zenodo.6631305

— Guillaume Bernard. *CoAID dataset texts with OCR degradations*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6630710. URL : https://doi.org/10.5281/zenodo.6630710

— Guillaume Bernard. *FibVid dataset texts with OCR degradations*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6630758. URL : https://doi.org/10.5281/zenodo.6630758

— Guillaume Bernard. *Event Registry titles dataset texts with OCR degradations*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6630828. URL : https://doi.org/10.5281/zenodo.6630828

— Guillaume Bernard. *Event Registry events associated to Wikidata entities*. Version 1. Zenodo, juin 2022. DOI : 10.5281/zenodo.6683770. URL : https://doi.org/10.5281/zenodo.6683770

Code et paquets logiciels

Divers

— [Log.] Guillaume Bernard, *synthesise_ocr_and_segmentation_errors_in_texts* version 1.0.0, fév. 2022. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/datasets/dataset_manipulation_tools/damage_datasets, SWHID : ⟨swh:1:dir:8847db56967b8110ab30c99e3e272e29ad86fbd5⟩

Chapitre 4 : Des documents de presse aux événements

Les bibliothèques logicielles et outils en ligne de commande utilisés pour les expériences présentées en chapitre 4, à la section 4.2.

— [Log.] Guillaume Bernard, *document_tracking* version 1.0.1, oct. 2021. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-documents-to-events/document_tracking, SWHID : ⟨swh:1:dir:3d4095a5bf4a021818152097741c6541430771cb⟩

— [Log.] Guillaume Bernard, *document_tracking_resources* version 1.0.1, 5 oct. 2021. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-documents-to-events/documents_tracking_resources, SWHID : ⟨swh:1:dir:337cec7f3b1ce92155490364e6fce581e2126dbf⟩

— [Log.] Guillaume Bernard, *document_processing* oct. 2021. Laboratoire L3i. LIC : GPLv3.

VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-documents-to-events/document_processing, SWHID : ⟨swh:1:dir:ae7da97b10a3cd3552fff2e7d86581f55e7187c0⟩

— [Log.] Guillaume BERNARD, *news_tracking* version 1.0.1, oct. 2021. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-documents-to-events/news_clustering_in_multiple_languages, SWHID : ⟨swh:1:dir:e28ca550b6faa36a7255e49c7df5b86f40cf6b14⟩

— [Log.] Guillaume BERNARD, *from_documents_to_events_experiments* avr. 2022. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-documents-to-events/news_tracking_experiments

Les utilitaires de vectorisation de documents présentés en chapitre 4, à la section 4.1.

— [Log.] Guillaume BERNARD, *compute_dense_vectors* oct. 2021. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/datasets/dataset_manipulation_tools/compute_dense_vectors, SWHID : ⟨swh:1:dir:2067958624a98644d4d3448056af542e3400453e⟩

— [Log.] Guillaume BERNARD, *compute_tf_idf_weights* oct. 2021. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/datasets/dataset_manipulation_tools/compute_tf_idf_weights, SWHID : ⟨swh:1:dir:a0638337f4eec1c4776746cd618703995bb6a936⟩

Chapitre 5 : Des événements aux documents de presse

— [Log.] Guillaume BERNARD, *wikivents* 9 juill. 2020. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-event-to-documents/wikivents-projects/wikivents, SWHID : ⟨swh:1:dir:0f9b87bd8fd89080dd4e0477d6759d275e4cf132⟩

— [Log.] Guillaume BERNARD, *annotate_events_with_wikidata_identifiers* mai 2022. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-event-to-documents/annotate_events_with_wikidata_identifiers, SWHID : ⟨swh:1:dir:d52f6f7ea67707a53628acedeb3935b4a7533869⟩

— [Log.] Thomas BLOT et Guillaume BERNARD, *database_infrastructure_text_mining* mai 2022. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-event-to-documents/database_infrastructure_text_mining, SWHID : ⟨swh:1:dir:3a9c6ca794d2f6fb00acd1fd48ee6ded9496891d⟩

— [Log.] Thomas BLOT et Guillaume BERNARD, *request_documents_based_on_events* mai 2022. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-event-to-documents/request_documents_based_on_events, SWHID : ⟨swh:1:dir:eee0c1d6db9bd1f34b7b55ac16a7a578c1658a29⟩

— [Log.] Thomas BLOT et Guillaume BERNARD, *events_to_documents_experiments* mai 2022. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-event-to-documents/experiments

Bibliographie

[05] *Reuters 2, Volume 2, Multilingual Corpus*. 31 mai 2005. URL : https://trec.nist.gov/data/reuters/reuters.html.

[21a] *Impact Environnemental Du Numérique à l'EPFL*. 2021. 40 p.

[21b] *Le Murmure des Mondes, « une Histoire captivante »*. Documentaire. Août 2021. URL : https://www.cite-telecoms.com/ (visité le 08/04/2022).

[81] *Loi Du 29 Juillet 1881 Sur La Liberté de La Presse*. Avec la coll. d'ASSEMBLÉE NATIONALE. 29 juill. 1881. URL : https://www.legifrance.gouv.fr/loda/id/LEGITEXT000006070722/ (visité le 23/07/2022).

[AA05] E. ADAR et L.A. ADAMIC. « Tracking Information Epidemics in Blogspace ». In : *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05). Compiegne, France, 2005, p. 207-214. ISBN : 978-0-7695-2415-3. DOI : 10.1109/WI.2005.151.

[Abb+20] Sajjad ABBASI et al. « Modeling Teacher-Student Techniques in Deep Neural Networks for Knowledge Distillation ». In : *2020 International Conference on Machine Vision and Image Processing (MVIP)*. 2020 International Conference on Machine Vision and Image Processing (MVIP). Iran, fév. 2020, p. 1-6. ISBN : 978-1-72816-832-6. DOI : 10.1109/MVIP49855.2020.9116923.

[AC18] Chantal AMRHEIN et Simon CLEMATIDE. « Supervised OCR Error Detection and Correction Using Statistical and Neural Machine Translation Methods ». In : *Journal for Language Technology and Computational Linguistics (JLCL)* 33.1 (1 2018), p. 49-76. ISSN : 0175-1336. DOI : 10.5167/uzh-162394.

[ACP21] ACPM. *Communiqué de Presse OneNext 2021 V1*. 2021. URL : https://www.acpm.fr/Media/Files/CP-ONENEXT-2021-V1-21-JANVIER-20213.

[AFP14] AFP. *Inondations catastrophiques en Bosnie et Serbie, au moins 40 morts*. Libération. 17 mai 2014. URL : https://www.liberation.fr/planete/2014/05/17/des-inondations-font-11-morts-en-bosnie-15000-evacues-en-serbie_1019583/ (visité le 21/06/2022).

[AG22] Iana Atanassova et Nicolas Gutehrlé. « Processing the Structure of Documents: Logical Layout Analysis of Historical Newspapers in French ». In : *Journal of Data Mining & Digital Humanities* NLP4DH (30 mai 2022). doi : 10.46298/jdmdh.9093.

[AGD20] Sara Abdollahi, Simon Gottschalk et Elena Demidova. *EventKG+Click: A Dataset of Language-specific Event-centric User Interaction Traces*. 23 oct. 2020. arXiv : 2010.12370 [cs]. url : http://arxiv.org/abs/2010.12370 (visité le 12/07/2022).

[Age+16] Rodrigo Agerri et al. « Multilingual Event Detection Using the NewsReader Pipelines ». In : International Conference on Language Resources and Evaluation (LREC). Portorož, Slovenia, 16 mai 2016, p. 42-46.

[Age22] Agence France Presse. *Factuel*. Factuel. 2 mai 2022. url : https://factuel.afp.com/ (visité le 02/05/2022).

[Agu+14] Jacqueline Aguilar et al. « A Comparison of the Events and Relations Across ACE, ERE, TAC-KBP,and FrameNet Annotation Standards ». In : *Proceedings of the The 2nd Workshop on EVENTS: Definition, Detection,Coreference, and Representation*. The 2nd Workshop on EVENTS: Definition, Detection,Coreference, and Representation. Baltimore, Maryland, USA, 22 juin 2014, p. 55-63. isbn : 978-1-941643-14-3. url : https://www.aclweb.org/anthology/W14-29.pdf#page=55 (visité le 29/01/2020).

[Ai21] Mts Ai. « BERT for Russian News Clustering ». In : *Proceedings of the International Conference "Dialogue 2021"*. Moscow, Russia, juin 2021, p. 6.

[AK15] Farzindar Atefeh et Wael Khreich. « A Survey of Techniques for Event Detection in Twitter: Techniques for Event Detection in Twitter ». In : *Computational Intelligence* 31.1 (1 fév. 2015), p. 132-164. issn : 08247935. doi : 10.1111/coin.12017.

[Ale+12] Bea Alex et al. « Digitised Historical Text: Does It Have to Be mediOCRe? » In : LThist 2012 Workshop at KONVENS 2012. Vienna, Switzerland, jan. 2012, p. 10.

[ALJ00] James Allan, Victor Lavrenko et Hubert Jin. « First Story Detection in TDT Is Hard ». In : *Proceedings of the Ninth International Conference on Information and Knowledge Management - CIKM '00*. The Ninth International Conference. McLean, Virginia, United States, 2000, p. 374-381. isbn : 978-1-58113-320-2. doi : 10.1145/354756.354843.

[All02a] James Allan. « Introduction to Topic Detection and Tracking ». In : *Topic Detection And Tracking: Event-based Information Organization*. 2002, p. 1-16. isbn : 978-1-4613-5311-9.

[All02b] James Allan. *Topic Detection And Tracking: Event-based Information Organization*. 2002. 272 p. isbn : 978-1-4613-5311-9.

[Ami+07] Enrique Amigo et al. « A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints ». In : (2007), p. 33.

[Amm+16] Waleed Ammar et al. *Massively Multilingual Word Embeddings*. 21 mai 2016. arXiv : 1602.01925 [cs]. URL : http://arxiv.org/abs/1602.01925 (visité le 24/07/2022).

[Ans+09] Scherp Ansgar et al. « F : A Model of Events Based on the Foundational Ontology dolce+DnS Ultralight ». In : *Proceedings of the Fifth International Conference on Knowledge Capture*. K-CAP '09: Fifth International Conference on Knowledge Capture. Redondo Beach, CA, USA, 1er sept. 2009, p. 137-144. ISBN : 978-1-60558-658-8.

[APS14] Rami Al-Rfou, Bryan Perozzi et Steven Skiena. *Polyglot: Distributed Word Representations for Multilingual NLP*. 27 juin 2014. arXiv : 1307.1662 [cs]. URL : http://arxiv.org/abs/1307.1662 (visité le 24/07/2022).

[AS19] Mikel Artetxe et Holger Schwenk. « Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond ». In : *Transactions of the Association for Computational Linguistics* 7 (nov. 2019), p. 597-610. ISSN : 2307-387X. DOI : 10.1162/tacl_a_00288. arXiv : 1812.10464 [cs].

[Asg+21] Meysam Asgari-Chenaghlu et al. « Topic Detection and Tracking Techniques on Twitter: A Systematic Review ». In : *Complexity* 2021 (18 juin 2021), e8833084. ISSN : 1076-2787. DOI : 10.1155/2021/8833084.

[Aue+07] Sören Auer et al. « DBpedia: A Nucleus for a Web of Open Data ». In : *The Semantic Web*. Réd. par David Hutchison et al. T. 4825. Lecture Notes in Computer Science. Berlin, Heidelberg, 2007, p. 722-735. ISBN : 978-3-540-76297-3 978-3-540-76298-0. DOI : 10.1007/978-3-540-76298-0_52.

[AY06] Charu C. Aggarwal et Philip S. Yu. « A Framework for Clustering Massive Text and Categorical Data Streams ». In : *Proceedings of the 2006 SIAM International Conference on Data Mining*. Proceedings of the 2006 SIAM International Conference on Data Mining. 20 avr. 2006, p. 479-483. ISBN : 978-0-89871-611-5 978-1-61197-276-4. DOI : 10.1137/1.9781611972764.44.

[AY10] Charu C. Aggarwal et Philip S. Yu. « On Clustering Massive Text and Categorical Data Streams ». In : *Knowledge and Information Systems* 24.2 (2 août 2010), p. 171-196. ISSN : 0219-1377, 0219-3116. DOI : 10.1007/s10115-009-0241-z.

[Bac86] Emmon Bach. « The Algebra of Events ». In : *Linguistics and Philosophy* 9.1 (1 fév. 1986), p. 5-16. DOI : 10.1002/9780470758335.ch13.

[Ban+21] Yejin Bang et al. « Model Generalization on COVID-19 Fake News Detection ». In : *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Communications in Computer and Information Science. Cham, 2021, p. 128-140. ISBN : 978-3-030-73696-5. DOI : 10.1007/978-3-030-73696-5_13.

[Ban+22] Juan M. Banda et al. *A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration*. Version 123. This dataset will be updated bi-weekly at least with additional tweets, look at the github repo for these updates. Release: We have standardized the name of the resource to match our pre-print manuscript and to not have to update it every week. Zenodo, juill. 2022. DOI : 10.5281/zenodo.6855183. URL : https://doi.org/10.5281/zenodo.6855183.

[BB22a] [Log.] Thomas Blot et Guillaume Bernard, *database_infrastructure_text_mining* mai 2022. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-event-to-documents/database_infrastructure_text_mining, SWHID : ⟨swh:1:dir:3a9c6ca794d2f6fb00acd1fd48ee6ded9496891d⟩.

[BB22b] [Log.] Thomas Blot et Guillaume Bernard, *events_to_documents_experiments* mai 2022. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-event-to-documents/experiments.

[BB22c] [Log.] Thomas Blot et Guillaume Bernard, *request_documents_based_on_events* mai 2022. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-event-to-documents/request_documents_based_on_events, SWHID : ⟨swh:1:dir:eee0c1d6db9bd1f34b7b55ac16a7a578c1658a29⟩.

[BD21] Emanuela Boros et Antoine Doucet. « Transformer-Based Methods for Recognizing Ultra Fine-grained Entities (RUFES) ». 13 avr. 2021. arXiv : 2104.06048 [cs]. URL : http://arxiv.org/abs/2104.06048 (visité le 12/05/2021).

[Bee+13] Roel Beetsma et al. « Spread the News: The Impact of News on the European Sovereign Bond Markets during the Crisis ». In : *Journal of International Money and Finance* 34 (avr. 2013), p. 83-101. ISSN : 02615606. DOI : 10.1016/j.jimonfin.2012.11.005.

[Ben+20] Domenico Benvenuto et al. « The 2019-New Coronavirus Epidemic: Evidence for Virus Evolution ». In : *Journal of Medical Virology* 92.4 (2020), p. 455-459. ISSN : 1096-9071. DOI : 10.1002/jmv.25688.

[Ber+21a] Guillaume Bernard et al. *Event representation on Wikidata and Wikipedia with, and without the analysis of vernacular languages*. Version 1. Zenodo, avr. 2021. DOI : 10.5281/zenodo.4733507. URL : https://doi.org/10.5281/zenodo.4733507.

[Ber+21b] Guillaume BERNARD et al. « Event Related Document Retrieval with Multilingual Real World Event Representation ». In : *Proceedings of the ISWC 2021 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice Co-Located with 20th International Semantic Web Conference*. 20th International Semantic Web Conference. T. 2980. Online, 27 oct. 2021, p. 5. ISBN : 1613-0073. URL : http://ceur-ws.org/Vol-2980/paper309.pdf.

[Ber+21c] Guillaume BERNARD et al. « A Comprehensive Extraction of Relevant Real-World-Event Qualifiers for Semantic Search Engines ». In : *Proceedings of the 25th International Conference on Theory and Practice of Digital Libraries*. 25th International Conference on Theory and Practice of Digital Libraries. T. 12866. Online, 15 sept. 2021, p. 153-164. DOI : 10.1007/978-3-030-86324-1_19.

[Ber+22] Guillaume BERNARD et al. « Tracking News Stories in Short Messages in the Era of Infodemic ». In : *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. CLEF 2022 Conference and Labs of the Evaluation Forum. T. 13390. Lecture Notes in Computer Science. Bologna, Italy, 5 sept. 2022, p. 18-32. ISBN : 978-3-031-13642-9. DOI : 10.1007/978-3-031-13643-6_2.

[Ber20] [Log.] Guillaume BERNARD, *wikivents* 9 juill. 2020. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-event-to-documents/wikivents-projects/wikivents, SWHID : ⟨swh:1:dir:0f9b87bd8fd89080dd4e0477d6759d275e4cf132⟩.

[Ber21a] [Log.] Guillaume BERNARD, *compute_dense_vectors* oct. 2021. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/datasets/dataset_manipulation_tools/compute_dense_vectors, SWHID : ⟨swh:1:dir:2067958624a98644d4d3448056af542e3400453e⟩.

[Ber21b] [Log.] Guillaume BERNARD, *compute_tf_idf_weights* oct. 2021. Laboratoire L3i. LIC : GPLv3. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/datasets/dataset_manipulation_tools/compute_tf_idf_weights, SWHID : ⟨swh:1:dir:a0638337f4eec1c4776746cd618703995bb6a936⟩.

[Ber21c] [Log.] Guillaume BERNARD, *Corpus de presse multilingue pour calcul de pondérations TF-IDF* oct. 2021. Laboratoire L3i. VCS : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/datasets/datasets/tf_idf_datasets/news_tf_idf_dataset.

[Ber21d] [Log.] Guillaume BERNARD, *Corpus de tweets multilingue pour calcul de pondérations TF-IDF* oct. 2021. Laboratoire L3i. VCS : https://gitlab

.univ-lr.fr/cross-lingual-event-tracking/datasets/datasets/tf_idf_datasets/twitter_tf_idf_dataset.

[Ber21e] [Log.] Guillaume Bernard, *document_processing* oct. 2021. Laboratoire L3i. lic : GPLv3. vcs : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-documents-to-events/document_processing, swhid : ⟨swh:1:dir:ae7da97b10a3cd3552fff2e7d86581f55e7187c0⟩.

[Ber21f] [Log.] Guillaume Bernard, *document_tracking* version 1.0.1, oct. 2021. Laboratoire L3i. lic : GPLv3. vcs : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-documents-to-events/document_tracking, swhid : ⟨swh:1:dir:3d4095a5bf4a021818152097741c6541430771cb⟩.

[Ber21g] [Log.] Guillaume Bernard, *document_tracking_resources* version 1.0.1, 5 oct. 2021. Laboratoire L3i. lic : GPLv3. vcs : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-documents-to-events/documents_tracking_resources, swhid : ⟨swh:1:dir:337cec7f3b1ce92155490364e6fce581e2126dbf⟩.

[Ber21h] [Log.] Guillaume Bernard, *news_tracking* version 1.0.1, oct. 2021. Laboratoire L3i. lic : GPLv3. vcs : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-documents-to-events/news_clustering_in_multiple_languages, swhid : ⟨swh:1:dir:e28ca550b6faa36a7255e49c7df5b86f40cf6b14⟩.

[Ber22a] [Log.] Guillaume Bernard, *Analyse des jeux de données* avr. 2022. Laboratoire L3i. lic : GPLv3. vcs : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/datasets/analysis.

[Ber22b] [Log.] Guillaume Bernard, *Analyse des jeux de données NewsEye* mai 2022. Laboratoire L3i. lic : GPLv3. vcs : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/datasets/analysis_newseye.

[Ber22c] [Log.] Guillaume Bernard, *annotate_events_with_wikidata_identifiers* mai 2022. Laboratoire L3i. lic : GPLv3. vcs : https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-event-to-documents/annotate_events_with_wikidata_identifiers, swhid : ⟨swh:1:dir:d52f6f7ea67707a53628acedeb3935b4a7533869⟩.

[Ber22d] Guillaume Bernard. *CoAID dataset texts with OCR degradations*. Version 1.0. Zenodo, juin 2022. doi : 10.5281/zenodo.6630710. url : https://doi.org/10.5281/zenodo.6630710.

[Ber22e] Guillaume Bernard. *CoAID dataset with multiple extracted features (both sparse and dense)*. Version 1.0. Zenodo, juin 2022. doi : 10.5281/zenodo.6630405. url : https://doi.org/10.5281/zenodo.6630405.

[Ber22f] Guillaume Bernard. *CoAID dataset with multiple extracted features (both sparse and dense) and degraded by OCR*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6630966. URL : https://doi.org/10.5281/zenodo.6630966.

[Ber22g] Guillaume Bernard. *Event Registry dataset texts with OCR degradations and synthesised segmentation*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6631305. URL : https://doi.org/10.5281/zenodo.6631305.

[Ber22h] Guillaume Bernard. *Event Registry dataset with multiple extracted features (both sparse and dense)*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6630367. URL : https://doi.org/10.5281/zenodo.6630367.

[Ber22i] Guillaume Bernard. *Event Registry dataset with multiple extracted features (both sparse and dense) and degraded by OCR*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6631267. URL : https://doi.org/10.5281/zenodo.6631267.

[Ber22j] Guillaume Bernard. *Event Registry events associated to Wikidata entities*. Version 1. Zenodo, juin 2022. DOI : 10.5281/zenodo.6683770. URL : https://doi.org/10.5281/zenodo.6683770.

[Ber22k] Guillaume Bernard. *Event Registry titles dataset texts with OCR degradations*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6630828. URL : https://doi.org/10.5281/zenodo.6630828.

[Ber22l] Guillaume Bernard. *Event Registry titles dataset with multiple extracted features (both sparse and dense) and degraded by OCR*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6631082. URL : https://doi.org/10.5281/zenodo.6631082.

[Ber22m] Guillaume Bernard. *Event Registry titles only dataset with multiple extracted features (both sparse and dense)*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6630447. URL : https://doi.org/10.5281/zenodo.6630447.

[Ber22n] Guillaume Bernard. *FibVid dataset texts with OCR degradations*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6630758. URL : https://doi.org/10.5281/zenodo.6630758.

[Ber22o] Guillaume Bernard. *Fibvid dataset with multiple extracted features (both sparse and dense)*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6630409. URL : https://doi.org/10.5281/zenodo.6630409.

[Ber22p] Guillaume Bernard. *FibVid dataset with multiple extracted features (both sparse and dense) and degraded by OCR*. Version 1.0. Zenodo, juin 2022. DOI : 10.5281/zenodo.6631070. URL : https://doi.org/10.5281/zenodo.6631070.

[Ber22q] [Log.] Guillaume BERNARD, *from_documents_to_events_experiments* avr. 2022. Laboratoire L3i. LIC : GPLv3. VCS : `https://gitlab.univ-lr.fr/cross-lingual-event-tracking/developpement/from-documents-to-events/news_tracking_experiments`.

[Ber22r] Guillaume BERNARD. *Resources to compute TF-IDF weightings on press articles and tweets*. Version 1.0. Zenodo, juin 2022. DOI : `10.5281/zenodo.6610406`. URL : `https://doi.org/10.5281/zenodo.6610406`.

[Ber22s] [Log.] Guillaume BERNARD, *synthesise_ocr_and_segmentation_errors_in_texts* version 1.0.0, fév. 2022. Laboratoire L3i. LIC : GPLv3. VCS : `https://gitlab.univ-lr.fr/cross-lingual-event-tracking/datasets/dataset_manipulation_tools/damage_datasets`, SWHID : ⟨`swh:1:dir:8847db56967b8110ab30c99e3e272e29ad86fbd5`⟩.

[BGC14] Igor BRIGADIR, Derek GREENE et Pádraig CUNNINGHAM. « Adaptive Representations for Tracking Breaking News on Twitter ». 28 nov. 2014. arXiv : `1403.2923 [cs]`. URL : `http://arxiv.org/abs/1403.2923` (visité le 28/01/2022).

[BGK15] Janusz BRZESZCZYŃSKI, Jerzy GAJDKA et Ali M. KUTAN. « Investor Response to Public News, Sentiment and Institutional Trading in Emerging Markets: A Review ». In : *International Review of Economics & Finance* 40 (nov. 2015), p. 338-352. ISSN : 10590560. DOI : `10.1016/j.iref.2015.10.042`.

[Bie+16] Ann BIES et al. « A Comparison of Event Representations in DEFT ». In : *Proceedings of the Fourth Workshop on Events*. Proceedings of the Fourth Workshop on Events. San Diego, California, 2016, p. 27-36. DOI : `10.18653/v1/W16-1004`.

[Bit+11] Andre BITTAR et al. « French TimeBank: An ISO-TimeML Annotated Reference Corpus ». In : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Shortpapers*. Portland, Oregon, USA, 19 juin 2011, p. 130-134.

[BJY03] David M BLEI, Michael I. JORDAN et Andrew Y. NG. « Latent Dirichlet Allocation ». In : *The Journal of Machine Learning Research* 3 (Jan jan. 2003), p. 993-1022.

[BL06] David M. BLEI et John D. LAFFERTY. « Dynamic Topic Models ». In : *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*. The 23rd International Conference. Pittsburgh, Pennsylvania, 2006, p. 113-120. ISBN : 978-1-59593-383-6. DOI : `10.1145/1143844.1143859`.

[Blo+08] Vincent D. BLONDEL et al. « Fast Unfolding of Communities in Large Networks ». In : *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[BMD21] Emanuela Boros, Jose Moreno et Antoine Doucet. « Event Detection as Question Answering with Entity Information ». In : 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Mexico City, Mexico, 6 juin 2021, p. 9.

[BNG11] Hila Becker, Mor Naaman et Luis Gravano. « Beyond Trending Topics: Real-World Event Identification on Twitter ». In : *Proceedings of the Fifth International Conference on Weblogs and Social*. International Conference on Weblogs and Social. Barceolona, Catalonio, Spain, 17 juill. 2011, p. 4. URL : http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2745.

[Bol19] Liza Bolz. « L'émergence de la dépêche télégraphique d'agence comme nouveau format d'écriture dans la presse française et allemande du XIXe siècle (1849-1870) ». 12 fév. 2019.

[Bon+22] Bonamy et al. « L'écoconception d'un service numérique : des actions pour réduire l'impact environnemental du numérique: » in : *Bulletin 1024* 19 (avr. 2022), p. 59-68. ISSN : 22701419. DOI : 10.48556/SIF.1024.19.59.

[Bor+20a] Emanuela Boros et al. « Alleviating Digitization Errors in Named Entity Recognition for Historical Documents ». In : *Proceedings of the 24th Conference on Computational Natural Language Learning*. Proceedings of the 24th Conference on Computational Natural Language Learning. Online, 2020, p. 431-441. DOI : 10.18653/v1/2020.conll-1.35.

[Bor+20b] Emanuela Boros et al. « Event Extraction over Digitised and Historical Documents ». In : 1.1 (nov. 2020), p. 17.

[Bor+20c] Emanuela Boros et al. « Robust Named Entity Recognition and Linking on Historical Multilingual Documents ». In : *Conference and Labs of the Evaluation Forum (CLEF 2020)*. T. 2696. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. Thessaloniki, Greece, sept. 2020, p. 1-17. URL : https://hal.archives-ouvertes.fr/hal-03026969 (visité le 12/05/2021).

[Bor+22a] Emanuela Boros et al. « Knowledge-Based Contexts for Historical Named Entity Recognition & Linking ». In : *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. CLEF 2022 Conference and Labs of the Evaluation Forum. Bologna, Italy, sept. 2022.

[Bor+22b] Emanuela Boros et al. *NewsEye : D3.8 Event Detection (Final)*. 31 jan. 2022.

[Bor+98] Andrew Borthwick et al. « NYU: Description of the MENE Named Entity System as Used in MUC-7 ». In : *Proceedings of a Conference Held in Fairfax, Virginia*. Seventh Message Understanding Conference. Fairfax, Virginia, 29 jan. 1998, p. 6.

[Bor18] Emanuela Boros. « Neural Methods for Event Extraction ». Paris, France : Université Paris-Saclay, 27 sept. 2018. 153 p. URL : https://tel.archives-ouvertes.fr/tel-01943841.

[BPL18] Desmond Bala Bisandu, Rajesh Prasad et Musa Muhammad Liman. « Clustering News Articles Using Efficient Similarity Measure and N-grams ». In : *International Journal of Knowledge Engineering and Data Mining* 5 (jan. 2018), p. 333-348. DOI : 10.1504/IJKEDM.2018.095525.

[Bri+13] Romain Brixtel et al. « Any Language Early Detection of Epidemic Diseases from Web News Streams ». In : *2013 IEEE International Conference on Healthcare Informatics*. 2013 IEEE International Conference on Healthcare Informatics (ICHI). Philadelphia, PA, USA, sept. 2013, p. 159-168. ISBN : 978-0-7695-5089-3. DOI : 10.1109/ICHI.2013.94.

[Bro+15] Ofer Bronstein et al. « Seed-Based Event Trigger Labeling: How Far Can Event Descriptions Get Us? » In : *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Beijing, China, 2015, p. 372-376. DOI : 10.3115/v1/P15-2061.

[Bru81] Étienne Brunet. « Le vocabulaire français de 1789 à nos jours d'après les données du Trésor de la Langue Française ». In : Congrès Informatique et Sciences humaines. Nov. 1981, p. 111-119.

[Bru95] Étienne Brunet. « L'évolution du lexique: approche statistique ». In : (1995), p. 29.

[BV22] [Log.] Elasticsearch B.V., *Elastic Search* version 8.1.2, 31 mars 2022. URL : https://www.elastic.co, VCS : https://github.com/elastic/elasticsearch, SWHID : ⟨swh:1:dir:892f855b90dc45ab6652b8d4d17a53e0af8ff7f1;origin=https://github.com/elastic/elasticsearch;⟩.

[C F46] Clarke C. F. O. *The Times: A revolution in Newspaper Printing*. Amstutz & Herdeg Graphis Press. Graphis 15. Zürich, Switzerland, 1946. 362-375. URL : https://magazines.iaddb.org/issue/GR/1946-05-01/edition/15/page/1.

[Cas+11] Tommaso Caselli et al. « Annotating Events, Temporal Expressions and Relations in Italian: The It-Timeml Experience for the Ita-TimeBank ». In : *Proceedings of the Fifth Linguistic Annotation Workshop*. Fifth Linguistic Annotation Workshop. Portland, Oregon, USA, 23 juin 2011, p. 143-151. DOI : 10.5555/2018966.2018984.

[CCG17] A. Castellanos, J. Cigarrán et A. García-Serrano. « Formal Concept Analysis for Topic Detection: A Clustering Quality Experimental Analysis ». In : *Information Systems* 66 (juin 2017), p. 24-42. issn : 03064379. doi : 10.1016/j.is.2017.01.008.

[Cer+18] Daniel Cer et al. « Universal Sentence Encoder ». 12 avr. 2018. arXiv : 1803.11175 [cs]. url : http://arxiv.org/abs/1803.11175 (visité le 28/01/2022).

[Che+15] Yubo Chen et al. « Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks ». In : *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China, 2015, p. 167-176. doi : 10.3115/v1/P15-1017.

[Chi+17] Guillaume Chiron et al. « Impact of OCR Errors on the Use of Digital Libraries ». In : *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*. Joint Conference on Digital Libraries. Toronto, ON, Canada, 19 juin 2017, p. 4. isbn : 978-1-5386-3861-3. doi : 10.1109/JCDL.2017.7991582.

[Chi+19] Muthu Chidambaram et al. « Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model ». In : *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. Florence, Italy, août 2019, p. 250-259. doi : 10.18653/v1/W19-4330.

[Chi98] Nancy A Chinchor. « Overview of MUC-7/MET-2 ». In : *Proceedings of a Conference Held in Fairfax, Virginia*. Seventh Message Understanding Conference. Fairfax, Virginia, 29 jan. 1998, p. 5.

[CHK12] Ivan Chupin, Nicolas Hubé et Nicolas Kaciaf. « II. L'« âge d'or » de la presse (1870-1939) ». In : t. 2e éd. Repères. Paris, 2012, p. 35-52. isbn : 978-2-7071-7371-3. url : https://www.cairn.info/histoire-politique-et-economique-des-medias-en-fra--9782707173713-p-35.htm (visité le 23/07/2022).

[Cie+02] Christopher Cieri et al. « Corpora for Topic Detection and Tracking ». In : *Topic Detection And Tracking: Event-based Information Organization*. 2002, p. 33-66. isbn : 978-1-4613-5311-9.

[CL02] Christopher Cieri et Mark Liberman. « TIDES Language Resources: A Resource Map for Translingual Information Access ». In : *In Proceedings of the Third International Language Resources and Evaluation Conference (Las Palmas)* (2002), p. 1334-1339.

[CL20] Limeng Cui et Dongwon Lee. « CoAID: COVID-19 Healthcare Misinformation Dataset ». 3 nov. 2020. arXiv : 2006.00885 [cs]. URL : http://arxiv.org/abs/2006.00885 (visité le 13/10/2021).

[CLH93] Nancy Chinchor, David D. Lewis et Lynette Hirschman. « Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3) ». In : *Computational Linguistics* 19.3 (3 sept. 1993), p. 409-449.

[CM98] Nancy Chinchor et Elaine Marsh. « Appendix D: MUC-7 Information Extraction Task Definition (Version 5.1) ». In : *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Seventh Message Understanding Conference (MUC-7). Fairfax, Virginia, 29 mai 1998, p. 53.

[Cou] [Log.] David Cournapeau, *scikit-learn* version 1.0.2. LIC : BSD 3-Clause. URL : https://scikit-learn.org/, VCS : https://github.com/scikit-learn/scikit-learn, SWHID : ⟨swh:1:dir:6fdfb293849a2b491e0302ff046b06848fd5efc8;origin=https://github.com/scikit-learn/scikit-learn;⟩.

[Dee+90] Scott Deerwester et al. « Indexing by Latent Semantic Analysis ». In : *Journal of the American Society for Information Science* 41.6 (1990), p. 391-407. ISSN : 1097-4571. DOI : 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[Des+19] Loïc Desquilbet et al. *Vers une recherche reproductible*. Mai 2019, p. 1. ISBN : 979-10-97595-05-0. URL : https://hal.archives-ouvertes.fr/hal-02144142 (visité le 31/05/2022).

[Dev+19] Jacob Devlin et al. « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ». 24 mai 2019. arXiv : 1810.04805 [cs]. URL : http://arxiv.org/abs/1810.04805 (visité le 28/11/2019).

[Die20] Nadine Dieudonné-Glad. *Numériser et mettre en ligne des fonds patrimoniaux anciens. Et apr...* Calendrier des sciences humaines et sociales. 30 avr. 2020. URL : https://calenda.org/750066 (visité le 23/07/2022).

[Dod+04] George Doddington et al. « The Automatic Content Extraction (ACE) Program. Tasks, Data and Evaluation ». In : *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Fourth International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal, mai 2004, p. 837-840. URL : http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf (visité le 28/01/2020).

[DOS07] Martin Doerr, Christian-Emil Ore et Stephen Stead. « The CIDOC Conceptual Reference Model - A New Standard for Knowledge Sharing ER2007 Tutorial ». In : *Tutorials, Posters, Panels and Industrial Contributions at the 26th International Conference on Conceptual Modeling*. 26th International Conference on Conceptual Modeling. T. 83. Auckland, New Zealand, nov. 2007, p. 51-56.

[Dow12] D. R. Dowty. *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. 6 déc. 2012. 441 p. isbn : 978-94-009-9473-7. Google Books : SxhtCQAAQBAJ.

[DZ17] Roberto Di Cosmo et Stefano Zacchiroli. « Software Heritage: Why and How to Preserve Software Source Code ». In : *iPRES 2017 - 14th International Conference on Digital Preservation*. Kyoto, Japan, sept. 2017, p. 1-10. url : https://hal.archives-ouvertes.fr/hal-01590958 (visité le 31/05/2022).

[EGC21] Eberhard, David M., Gary F. Simons et Charles D. Fennig. *Ethnologue: Languages of the World*. 2021. url : https://www.ethnologue.com/ (visité le 07/04/2021).

[Ehr+20] Maud Ehrmann et al. « Introducing the CLEF 2020 HIPE Shared Task: Named Entity Recognition and Linking on Historical Newspapers ». In : *Lecture Notes in Computer Science*. 42nd European Conference on IR Research (ECIR). Lisbon, Portugal, 8 avr. 2020, p. 524-532. doi : 10.1007/978-3-030-45442-5_68.

[Ehr+22] Maud Ehrmann et al. « Introducing the HIPE 2022 Shared Task: Named Entity Recognition and Linking in Multilingual Historical Documents ». In : *Advances in Information Retrieval*. Lecture Notes in Computer Science. Cham, 2022, p. 347-354. isbn : 978-3-030-99739-7. doi : 10.1007/978-3-030-99739-7_44.

[EMN92] Bernhard E. Boser, Isabelle M. Guyon et Vladimir N. Vapnik. « A Training Algorithm for Optimal Margin Classifiers ». In : *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. Annual Workshop on Computational Learning Theory. 1992, p. 144-152.

[EN11] Peter Exner et Pierre Nugues. « Using Semantic Role Labeling to Extract Events from Wikipedia ». In : *DeRiVE@ ISWC*. 2011, p. 38-47.

[EO13] European Organization For Nuclear Research et OpenAIRE. « Zenodo ». In : 2013. doi : 10.25495/7GXK-RD71.

[Eve20] Event Registry. *Event Registry - Use the Power of AI to Turn Raw News Content into Actionable Insights*. Event Registry. 3 mars 2020. url : https://eventregistry.org (visité le 03/03/2020).

[Fan+21] Wentao Fan et al. « Clustering-Based Online News Topic Detection and Tracking Through Hierarchical Bayesian Nonparametric Models ». In : *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual Event Canada, 11 juill. 2021, p. 2126-2130. isbn : 978-1-4503-8037-9. doi : 10.1145/3404835.3462982.

[Fär+17] Michael Färber et al. « Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO ». In : *Semantic Web* 9.1 (30 nov. 2017), p. 77-129. ISSN : 22104968, 15700844. DOI : 10.3233/SW-170275.

[FDM13] Fabrice Flipo, Michelle Dobré et Marion Michot. *La face cachée du numérique*. Pour en finir avec. 2013. ISBN : 978-2-915830-77-4. URL : https://journals.openedition.org/lectures/12270 (visité le 25/06/2022).

[FH17] Erik Faessler et Udo Hahn. « Semedico: A Comprehensive Semantic Search Engine for the Life Sciences ». In : *Proceedings of ACL 2017, System Demonstrations*. Proceedings of ACL 2017, System Demonstrations. Vancouver, Canada, 2017, p. 91-96. DOI : 10.18653/v1/P17-4016.

[FHM06] Elena Filatova, Vasileios Hatzivassiloglou et Kathleen McKeown. « Automatic Creation of Domain Templates ». In : *Proceedings of the COLING/ACL on Main Conference Poster Sessions* -. The COLING/ACL. Sydney, Australia, 2006, p. 207-214. DOI : 10.3115/1273073.1273100.

[Fou22a] [Log.] Apache Foundation, *Apache Lucene* version 9.2.0, 23 mai 2022. URL : https://lucene.apache.org/, VCS : https://github.com/apache/lucene, SWHID : ⟨swh:1:dir:fc2c593e3e805776d462ce3fda8c738049b884a4;origin=https://github.com/apache/lucene;⟩.

[Fou22b] [Log.] Apache Foundation, *Apache SOLR* version 9.0.0, mai 2022. LIC : Apache License, Version 2.0. URL : https://solr.apache.org, VCS : https://github.com/apache/solr, SWHID : ⟨swh:1:dir:6b81fe2966acce31b56511caa6f82a0ca8953e97;origin=https://github.com/apache/solr;⟩.

[FQL18] Xiaocheng Feng, Bing Qin et Ting Liu. « A Language-Independent Neural Network for Event Detection ». In : *Science China Information Sciences* 61.9 (sept. 2018), p. 092106. ISSN : 1674-733X, 1869-1919. DOI : 10.1007/s11432-017-9359-x.

[Fra21] FranceArchives. *Numérisation du patrimoine*. FranceArchives. 21 déc. 2021. URL : https://francearchives.fr/fr/article/37769 (visité le 23/07/2022).

[Fre+22] Simona Frenda et al. « The Unbearable Hurtfulness of Sarcasm ». In : *Expert Systems with Applications* 193 (1er mai 2022), p. 116398. ISSN : 0957-4174. DOI : 10.1016/j.eswa.2021.116398.

[FZ02] Richard Fikes et Qing Zhou. « A Reusable Time Ontology ». In : *Proceedings of the AAAI 2002 National Conference*. AAAI 2002 National Conference. Juill. 2002, p. 6.

[Gan+20] Sahaj Gandhi et al. « Event-Related Query Classification with Deep Neural Networks ». In : *Companion Proceedings of the Web Conference 2020*. WWW '20: The Web Conference 2020. Taipei Taiwan, 20 avr. 2020, p. 324-330. ISBN : 978-1-4503-7024-0. DOI : 10.1145/3366424.3382183.

[GC09] Robert Gwadera et Fabio Crestani. « Mining and Ranking Streams of News Stories Using Cross-Stream Sequential Patterns ». In : *Proceeding of the 18th ACM Conference on Information and Knowledge Management - CIKM '09*. Proceeding of the 18th ACM Conference. Hong Kong, China, 2009, p. 1709. isbn : 978-1-60558-512-3. doi : 10.1145/1645953.1646210.

[GD02] Jonathan G. Fiscuss et George Doddington. « Topic Detection and Tracking Evaluation Overview ». In : *Topic Detection And Tracking: Event-based Information Organization*. 2002, p. 17-32. isbn : 978-1-4613-5311-9.

[GD19] Simon Gottschalk et Elena Demidova. « EventKG - the Hub of Event Knowledge on the Web - and Biographical Timeline Generation ». In : *Semantic Web* 10.6 (28 oct. 2019), p. 1039-1070. doi : 10.3233/SW-190355. arXiv : 1905.08794.

[GF15] Adrien Guille et Cecile Favre. « Event Detection, Tracking, and Visualization in Twitter: A Mention-Anomaly-Based Approach ». In : *Social Network Analysis and Mining* 5.1 (déc. 2015), p. 18. issn : 1869-5450, 1869-5469. doi : 10.1007/s13278-015-0258-0. arXiv : 1505.05657 [cs].

[Gha+18] Reza Ghaeini et al. *Event Nugget Detection with Forward-Backward Recurrent Neural Networks*. 15 fév. 2018. arXiv : 1802.05672 [cs]. url : http://arxiv.org/abs/1802.05672 (visité le 11/07/2022).

[Gha+20] Bilal Ghanem et al. « Irony Detection in a Multilingual Context ». In : *Advances in Information Retrieval*. Lecture Notes in Computer Science. Cham, 2020, p. 141-149. isbn : 978-3-030-45442-5. doi : 10.1007/978-3-030-45442-5_18.

[GLM16] Salvatore Gaglio, Giuseppe Lo Re et Marco Morana. « A Framework for Real-Time Twitter Data Analysis ». In : *Computer Communications* 73 (jan. 2016), p. 236-242. issn : 01403664. doi : 10.1016/j.comcom.2015.09.021.

[Got+18] Simon Gottschalk et al. « Towards Better Understanding Researcher Strategies in Cross-Lingual Event Analytics ». 5 sept. 2018. doi : 10.1007/978-3-030-00066-0_12. arXiv : 1809.08084 [cs].

[Gra+21] Kacper T Gradoń et al. « Countering Misinformation: A Multidisciplinary Approach ». In : *Big Data & Society* 8.1 (jan. 2021), p. 205395172110138. issn : 2053-9517, 2053-9517. doi : 10.1177/20539517211013848.

[Gri+03] Thomas Griffiths et al. « Hierarchical Topic Models and the Nested Chinese Restaurant Process ». In : *Advances in Neural Information Processing Systems*. T. 16. 2003. url : https://proceedings.neurips.cc/paper/2003/hash/7b41bfa5085806dfa24b8c9de0ce567f-Abstract.html (visité le 25/07/2022).

[GS03] R Gaizauskas et A Setzer. « The TimeBank Corpus ». In : *Proceedings of Corpus Linguistics*. Corpus Linguistics. Mars 2003, p. 647-656.

[GS96] Ralph Grishman et Beth Sundheim. « Message Understanding Conference-6: A Brief History ». In : *Proceedings of the 16th Conference on Computational Linguistics*. COLING'96. T. 1. Copenhagen, Denmark, 5 août 1996, p. 466-471. doi : 10.3115/992628.992709.

[Ham+19] Ahmed Hamdi et al. « An Analysis of the Performance of Named Entity Recognition over OCRed Documents ». In : *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). Champaign, IL, USA, juin 2019, p. 333-334. isbn : 978-1-72811-547-4. doi : 10.1109/JCDL.2019.00057.

[HHD20] Vinh-Nam Huynh, Ahmed Hamdi et Antoine Doucet. « When to Use OCR Post-correction for Named Entity Recognition? » In : *Digital Libraries at Times of Massive Societal Transition*. T. 12504. Lecture Notes in Computer Science. Cham, 2020, p. 33-42. isbn : 978-3-030-64451-2 978-3-030-64452-9. doi : 10.1007/978-3-030-64452-9_3.

[Hof+13] Johannes Hoffart et al. « YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia ». In : *Artificial Intelligence* 194 (jan. 2013), p. 28-61. issn : 00043702. doi : 10.1016/j.artint.2012.06.001.

[Hon+11] Yu Hong et al. « Using Cross-Entity Inference to Improve Event Extraction ». In : Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon, USA, 19 juin 2011, p. 1127-1136.

[Hon+18] Yu Hong et al. « Self-Regulation: Employing a Generative Adversarial Network to Improve Event Detection ». In : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, 2018, p. 515-526. doi : 10.18653/v1/P18-1048.

[Hua+15] Kejun Huang et al. « Translation Invariant Word Embeddings ». In : *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015, p. 1084-1088. doi : 10.18653/v1/D15-1127.

[Hua08] Anna Huang. « Similarity Measures for Text Document Clustering ». In : *Proceedings of the Sixth New Zealand Computer Science Research Student Conference*. Christchurch, New Zealand, 2008, p. 49-56.

[Huy+19] Charles Huyghues-Despointes et al. « Weaving Information Propagation: Modeling the Way Information Spreads in Document Collections ». In : *Advances in Artificial Intelligence 32nd Canadian Conference on Artificial Intelligence*. 32nd Canadian Conference on Artificial Intelligence. Kingston, Canada, 24 avr. 2019, p. 394-399. url : https://doi.org/10.1007/978-3-030-18305-9_35.

[Int13] Intergovernmental Panel on Climate Change. *Climate Change 2013: The Physical Science Basis*. 2013. 1552 p.

[Int21] International Council on Archives - Expert Group on Archival Description. *Records in Contexts Conceptual Model*. Brouillon 0.2. Juill. 2021.

[IQO21] Muhammad Imran, Umair Qazi et Ferda Ofli. « TBCOV: Two Billion Multilingual COVID-19 Tweets with Sentiment, Entity, Geo, and Gender Labels ». 4 oct. 2021. arXiv : 2110.03664 [cs]. URL : http://arxiv.org/abs/2110.03664 (visité le 14/10/2021).

[ISF21] ISFJ. *Qu'est-ce que la règle des 5w en journalisme ?* Institut Supérieur de Formation au Journalisme. 19 mai 2021. URL : https://www.isfj.fr/actualites/2021-journalisme-5w/ (visité le 25/07/2022).

[JD20] Axel Jean-Caurant et Antoine Doucet. « Accessing and Investigating Large Collections of Historical Newspapers with the NewsEye Platform ». In : *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. New York, NY, USA, 1er août 2020, p. 531-532. ISBN : 978-1-4503-7585-6. URL : https://doi.org/10.1145/3383583.3398627 (visité le 02/08/2022).

[Jeo+16] Young-Seob Jeong et al. « Korean TimeML and Korean TimeBank ». In : (2016), p. 356-359.

[Jia+21] Hang Jiang et al. *Topic Detection and Tracking with Time-Aware Document Embeddings*. 12 déc. 2021. arXiv : 2112.06166 [cs]. URL : http://arxiv.org/abs/2112.06166 (visité le 12/07/2022).

[Joa02] Thorsten Joachims. « Optimizing Search Engines Using Clickthrough Data ». In : *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton Alberta Canada, juill. 2002, p. 133-142. ISBN : 978-1-58113-567-1. URL : https://dl.acm.org/doi/proceedings/10.1145/775047.

[Joa06] Thorsten Joachims. « Training Linear SVMs in Linear Time ». In : *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '06*. The 12th ACM SIGKDD International Conference. Philadelphia, PA, USA, 2006, p. 217. ISBN : 978-1-59593-339-3. DOI : 10.1145/1150402.1150429.

[Joa09] [Log.] Thorsten Joachims, *Support Vector Machine for Ranking* version 1.00, 21 mars 2009. Cornell University. URL : https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html.

[Jou+17] Nicholas Journet et al. « DocCreator: A New Software for Creating Synthetic Ground-Truthed Document Images ». In : *Journal of Imaging* 3.4 (11 déc. 2017), p. 62. ISSN : 2313-433X. DOI : 10.3390/jimaging3040062.

[JSS11] Alan Jackoway, Hanan Samet et Jagan Sankaranarayanan. « Identification of Live News Events Using Twitter ». In : *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks - LBSN '11*. The 3rd ACM SIGSPATIAL International Workshop. Chicago, Illinois, 2011, p. 1. isbn : 978-1-4503-1033-8. doi : 10.1145/2063212.2063224.

[JY16] Abhyuday N Jagannatha et Hong Yu. « Bidirectional RNN for Medical Event Detection in Electronic Health Records ». In : *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, 2016, p. 473-482. doi : 10.18653/v1/N16-1056.

[Kan81] Emanuel Kant. *Critique de La Raison Pure*. Quadrige. 1781. 624 p. isbn : 978-2-13-060871-4.

[Ken03] Anthony Kenny. *Action, Emotion and Will*. 2 sept. 2003. isbn : 978-0-203-71146-0. doi : 10.4324/9780203711460.

[Khr+15] Anton S. Khritankov et al. « Discovering Text Reuse in Large Collections of Documents: A Study of Theses in History Sciences ». In : *2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT)*. 2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT). St. Petersburg, Russia, nov. 2015, p. 26-32. isbn : 978-952-68397-0-7. doi : 10.1109/AINL-ISMW-FRUCT.2015.7382965.

[KIF07] Ken Kaneiwa, Michiaki Iwazume et Ken Fukuda. « An Upper Ontology for Event Classifications and Relations ». In : *AI 2007: Advances in Artificial Intelligence*. T. 4830. Lecture Notes in Computer Science. Berlin, Heidelberg, 2007, p. 394-403. isbn : 978-3-540-76926-2. doi : 10.1007/978-3-540-76928-6_41.

[Kim+09] Jin-Dong Kim et al. « Overview of BioNLP'09 Shared Task on Event Extraction ». In : *Proceedings of the Workshop on BioNLP Shared Task - BioNLP '09*. The Workshop. Boulder, Colorado, 2009, p. 1. isbn : 978-1-932432-44-2. doi : 10.3115/1572340.1572342.

[Kim+21] Jisu Kim et al. « FibVID: Comprehensive Fake News Diffusion Dataset during the COVID-19 Period ». In : *Telematics and Informatics* 64 (nov. 2021), p. 101688. issn : 07365853. doi : 10.1016/j.tele.2021.101688.

[KRS21] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan et Sivanesan Sangeetha. *AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing*. 28 août 2021. arXiv : 2108.05542 [cs]. url : http://arxiv.org/abs/2108.05542 (visité le 09/07/2022).

[LBH21] Philippe Laban, Lucas Bandarkar et Marti A. Hearst. « News Headline Grouping as a Challenging NLU Task ». In : *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online, 2021, p. 3186-3198. doi : 10.18653/v1/2021.naacl-main.255.

[LCB12] Jey Han Lau, Nigel Collier et Timothy Baldwin. « Online Trend Analysis with Topic Models: #twitter Trends Detection Topic Model Online ». In : *Proceedings of COLING 2012* (2012), p. 1519-1534.

[Le 22] Le Figaro avec AFP. « Les marchés un peu rassurés par la réaction de la Fed face à l'inflation ». In : *Le Figaro. Flash Eco* (15 juill. 2022). url : https://www.lefigaro.fr/flash-eco/les-marches-un-peu-rassures-par-la-reaction-de-la-fed-face-a-l-inflation-20220715 (visité le 17/07/2022).

[Leb+14] Gregor Leban et al. « Event Registry: Learning about World Events from News ». In : *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*. The 23rd International Conference. Seoul, Korea, 2014, p. 107-110. isbn : 978-1-4503-2745-9. doi : 10.1145/2567948.2577024.

[Lej+15] Gaël Lejeune et al. « Multilingual Event Extraction for Epidemic Detection ». In : *Artificial Intelligence in Medicine* 65.2 (2 oct. 2015), p. 131-143. issn : 09333657. doi : 10.1016/j.artmed.2015.06.005.

[Lej13] Gaël Lejeune. « Veille épidémiologique multilingue : une approche parcimonieuse au grain caractère fondée sur le genre textuel ». Caen, Basse-Normandie : Université de Caen Basse-Normandie, 16 oct. 2013. 204 p. url : https://hal.archives-ouvertes.fr/tel-01074940/document (visité le 03/12/2019).

[Les22a] Les Contributeurs de Wikipédia. *Loi de Moore*. In : *Wikipédia*. 29 mars 2022. url : https://fr.wikipedia.org/w/index.php?title=Loi_de_Moore&oldid=192355797 (visité le 17/06/2022).

[Les22b] Les Contributeurs de Wikipédia. *Lsjbot*. In : *Wikipédia*. 17 avr. 2022. url : https://fr.wikipedia.org/w/index.php?title=Lsjbot&oldid=192920081 (visité le 20/06/2022).

[Lev66] Vladimir Levenshtein. « Binary Codes Capable of Correcting Deletions, Insertions, and Reversals ». In : *Soviet physics doklady* 10.8 (fév. 1966), p. 707-710.

[LH17] Philippe Laban et Marti Hearst. « newsLens: Building and Visualizing Long-Ranging News Stories ». In : *Proceedings of the Events and Stories in the News Workshop*. Proceedings of the Events and Stories in the News Workshop. Vancouver, Canada, 2017, p. 1-9. doi : 10.18653/v1/W17-2701.

[LH20] Mathis Linger et Mhamed Hajaiej. « Batch Clustering for Multilingual News Streaming ». In : *Proceedings of Text2Story - Third Workshop on Narrative Extraction From Texts Co-Located with 42nd European Conference on Information Retrieval*. Third Workshop on Narrative Extraction From Texts Co-Located with 42nd European Conference on Information Retrieval, Text2Story@ECIR 2020. T. 2593. CEUR Workshop Proceedings. Lisbon, Portugal, 14 avr. 2020, p. 55-61. arXiv : 2004.08123. URL : http://ceur-ws.org/Vol-2593/paper7.pdf (visité le 17/06/2021).

[LHP01] Carl Lagoze, Jane Hunter et DSTC Pty. « The ABC Ontology and Model ». In : *DC 2001 Proceedings*. DCMI International Conference on Dublin Core and Metadata Applications. Tokyo, Japan, 24 oct. 2001, p. 17. ISBN : 4-924600-98-9.

[Li+20] Yichuan Li et al. « MM-COVID: A Multilingual and Multimodal Data Repository for Combating COVID-19 Disinformation ». 23 nov. 2020. arXiv : 2011.04088 [cs]. URL : http://arxiv.org/abs/2011.04088 (visité le 02/05/2022).

[Lig21] Anne-Laure Ligozat. *Consommation énergétique de l'utilisation de l'IA – EcoInfo*. 12 juin 2021. URL : https://ecoinfo.cnrs.fr/2021/06/12/consommation-energetique-de-lutilisation-de-lia/ (visité le 20/06/2022).

[Lin+19] Elvys Linhares Pontes et al. « Impact of OCR Quality on Named Entity Linking ». In : *Digital Libraries at the Crossroads of Digital Information for the Future*. T. 11853. Lecture Notes in Computer Science. Cham, 2019, p. 102-115. ISBN : 978-3-030-34057-5 978-3-030-34058-2. DOI : 10.1007/978-3-030-34058-2_11.

[Lin+20] Elvys Linhares Pontes et al. « Entity Linking for Historical Documents: Challenges and Solutions ». In : *Digital Libraries at Times of Massive Societal Transition*. T. 12504. Lecture Notes in Computer Science. Cham, 2020, p. 215-231. ISBN : 978-3-030-64451-2 978-3-030-64452-9. DOI : 10.1007/978-3-030-64452-9_19.

[Lin02] Linguistic Data Consortium. *TDT Pilot Study Corpus*. 2002. URL : https://catalog.ldc.upenn.edu/LDC98T25 (visité le 10/12/2019).

[Lin05] Linguistic Data Consortium. *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*. 1er juill. 2005. URL : http://www.ldc.upenn.edu/Projects/ACE/ (visité le 10/12/2019).

[Liu+20] Wei Liu et al. « Topic Detection and Tracking Based on Event Ontology ». In : *IEEE Access* 8 (2020), p. 98044-98056. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2020.2995776.

[LJC19] Chae-Gyun Lim, Young-Seob Jeong et Ho-Jin Choi. « Survey of Temporal Information Extraction ». In : *Journal of Information Processing Systems* 15.4 (31 août 2019), p. 931-956. DOI : 10.3745/JIPS.04.0129.

[LJH13] Qi Li, Heng Ji et Liang Huang. « Joint Event Extraction via Structured Prediction with Global Features ». In : *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. 51st Annual Meeting of the Association for Computational Linguistics. T. 1. 2013, p. 73-82.

[LLC99] [Log.] ImageMagick Studio LLC, 1999. lic : ImageMagick License. url : `https://imagemagick.org/index.php`, vcs : `https://github.com/imagemagick/imagemagick`, swhid : ⟨`swh:1:dir:022108bbad7f1de460627e469a914a2f4919bc6c;origin=https://github.com/ImageMagick/ImageMagick`⟩.

[LLY21] Jialu Liu, Tianqi Liu et Cong Yu. « NewsEmbed: Modeling News through Pre-trained Document Representations ». In : *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 14 août 2021, p. 1076-1086. doi : `10.1145/3447548.3467392`. arXiv : `2106.00590 [cs]`.

[LM14] Quoc V. Le et Tomas Mikolov. « Distributed Representations of Sentences and Documents ». In : 22 mai 2014. arXiv : `1405.4053 [cs]`. url : `http://arxiv.org/abs/1405.4053` (visité le 07/06/2021).

[LMD20] Elvys Linhares Pontes, Jose G. Moreno et Antoine Doucet. « Linking Named Entities across Languages Using Multilingual Word Embeddings ». In : *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. JCDL '20: The ACM/IEEE Joint Conference on Digital Libraries in 2020. Virtual Event China, août 2020, p. 329-332. isbn : 978-1-4503-7585-6. doi : `10.1145/3383583.3398597`.

[LS17] Xiaoyan Lu et Boleslaw Szymanski. « Predicting Viral News Events in Online Media ». In : *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. 2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). Orlando / Buena Vista, FL, USA, mai 2017, p. 1447-1456. isbn : 978-1-5386-3408-0. doi : `10.1109/IPDPSW.2017.82`.

[LSS02] Tim Leek, Richard Schwartz et Srinivasa Sista. « Probabilistic Approaches to Topic Detection and Tracking ». In : *Topic Detection And Tracking: Event-based Information Organization*. 2002, p. 67-82.

[Mac67] J Macqueen. « Some Methods for Classification and Analysis of Multivariate Observations ». In : *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley Symposium on Mathematical Statistics and Probability. T. 1. Berkeley, CA, USA, juin 1967, p. 281-297.

[Mak03] Juha Makkonen. « Investigations on Event Evolution ». In : *Proceedings of the HLT-NAACL 2003 Student Research Workshop*. HLT-NAACL 2003 Student Research Workshop. Edmonton Alberta Canada, 2003, p. 43-48.

[Mar+21] Michał Marcińczuk et al. « Text Document Clustering: Wordnet vs. TF-IDF vs. Word Embeddings ». In : *Proceedings of the 11th Global Wordnet Conference*. Global Wordnet Conference. University of South Africa (UNISA), jan. 2021, p. 207-214. URL : https://www.aclweb.org/anthology/2021.gwc-1.24.pdf (visité le 07/06/2021).

[MAS03] Juha Makkonen, Helena Ahonen-Myka et Marko Salmenkivi. « Topic Detection and Tracking with Spatio-Temporal Evidence ». In : *Advances in Information Retrieval*. Réd. par Gerhard Goos, Juris Hartmanis et Jan van Leeuwen. T. 2633. Lecture Notes in Computer Science. Berlin, Heidelberg, 2003, p. 251-265. ISBN : 978-3-540-01274-0 978-3-540-36618-8. DOI : 10.1007/3-540-36618-0_18.

[MAS04] Juha Makkonen, Helena Ahonen-Myka et Marko Salmenkivi. « Simple Semantics in Topic Detection and Tracking ». In : *Information Retrieval* 7.3/4 (sept. 2004), p. 347-368. ISSN : 1386-4564. DOI : 10.1023/B:INRT.0000011210.12953.86.

[MBC17] Ida Mele, Seyed Ali Bahrainian et Fabio Crestani. « Linking News across Multiple Streams for Timeliness Analysis ». In : *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*. The 2017 ACM. Singapore, Singapore, 2017, p. 767-776. ISBN : 978-1-4503-4918-5. DOI : 10.1145/3132847.3132988.

[MBC19] Ida Mele, Seyed Ali Bahrainian et Fabio Crestani. « Event Mining and Timeliness Analysis from Heterogeneous News Streams ». In : *Information Processing & Management* 56.3 (3 mai 2019), p. 969-993. ISSN : 03064573. DOI : 10.1016/j.ipm.2019.02.003.

[Mei+17] Benjamin Meier et al. « Fully Convolutional Neural Networks for Newspaper Article Segmentation ». In : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto, nov. 2017, p. 414-419. ISBN : 978-1-5386-3586-5. DOI : 10.1109/ICDAR.2017.75.

[Mic+21] Johannes Michael et al. « ICPR 2020 Competition on Text Block Segmentation on a NewsEye Dataset ». In : *Pattern Recognition. ICPR International Workshops and Challenges*. T. 12668. Lecture Notes in Computer Science. Cham, 2021, p. 405-418. ISBN : 978-3-030-68792-2 978-3-030-68793-9. DOI : 10.1007/978-3-030-68793-9_30.

[Mil95] George A Miller. « WordNet: A Lexical Database for English ». In : *Communications of the ACM* 38.11 (nov. 1995), p. 3.

[Min22] Ministère de la Culture. *Gérer un fonds photographique - La numérisation*. Ministère de la Culture - République Française. 2022. URL : https://www.culture.gouv.fr/Thematiques/Photographie/Gerer-u

n-fonds-photographique/Valorisation/La-numerisation (visité le 23/07/2022).

[Min75] Marvin Minsky. « A Framework for Representing Knowledge ». In : *The Psychology of Computer Vision* (1975).

[Mir+18] Sebastião Miranda et al. « Multilingual Clustering of Streaming News ». In : *2018 Conference on Empirical Methods in Natural Language Processing*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 31 oct. 2018, p. 4535-4544. arXiv : 1809.00540. URL : https://www.aclweb.org/anthology/D18-1483/ (visité le 22/06/2021).

[Mot+21] Zeynab Mottaghinia et al. « A Review of Approaches for Topic Detection in Twitter ». In : *Journal of Experimental & Theoretical Artificial Intelligence* 33.5 (3 sept. 2021), p. 747-773. ISSN : 0952-813X. DOI : 10.1080/0952813X.2020.1785019.

[MRŠ12] Andrej Muhič, Jan Rupnik et Primož Škraba. « Cross-Lingual Document Similarity ». In : *Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces*. Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces. Cavtat / Dubrovnik, Croatia, juin 2012, p. 387-392. ISBN : 978-1-4673-1629-3. DOI : 10.2498/iti.2012.0467.

[MS02] I Scott MacKenzie et R William Soukoreff. « A Character-level Error Analysis Technique for Evaluating Text Entry Methods ». In : *Proceedings of the second Nordic conference on Human-computer interaction* (oct. 2002), p. 243-246.

[MSM11] David McClosky, Mihai Surdeanu et Christopher Manning. « Event Extraction as Dependency Parsing for BioNLP 2011 ». In : *In Proceedings of BioNLP Shared Task 2011 Workshop*. BioNLP Shared Task 2011 Workshop. Juin 2011, p. 41-45.

[Mut+18] Stephen Mutuvi et al. « Evaluating the Impact of OCR Errors on Topic Modeling ». In : *Maturity and Innovation in Digital Libraries*. T. 11279. Lecture Notes in Computer Science. Cham, 2018, p. 3-14. ISBN : 978-3-030-04256-1 978-3-030-04257-8. DOI : 10.1007/978-3-030-04257-8_1.

[Mut+21a] Stephen Mutuvi et al. « Token-Level Multilingual Epidemic Dataset for Event Extraction ». In : *Linking Theory and Practice of Digital Libraries*. T. 12866. Lecture Notes in Computer Science. Cham, 2021, p. 55-59. ISBN : 978-3-030-86323-4 978-3-030-86324-1. DOI : 10.1007/978-3-030-86324-1_6.

[Mut+21b] Stephen Mutuvi et al. « Multilingual Epidemic Event Extraction ». In : *Towards Open and Trustworthy Digital Societies*. T. 13133. Lecture Notes in Computer Science. Cham, 2021, p. 139-156. ISBN : 978-3-030-91668-8 978-3-030-91669-5. DOI : 10.1007/978-3-030-91669-5_12.

[MWL22] Johannes MICHAEL, Max WEIDEMANN et Roger LABAHN. *NewsEye : D2.7 Article Separation*. 6. 31 jan. 2022, p. 54. URL : https://www.newseye.eu/fileadmin/deliverables/NewsEye-T23-D27-ArticleSeparation-c-final-Submitted-v6.0.pdf (visité le 31/07/2022).

[Na17] [Log.] Muriel Visani NICHOLAS JOURNET et AL., 2017. LIC : GPLv3. URL : https://doc-creator.labri.fr/, VCS : https://github.com/DocCreator/DocCreator, SWHID : ⟨swh:1:dir:1737b8fc2176454623fa8e5e63b4a9f59d100d9a;origin=https://github.com/DocCreator/DocCreator;⟩.

[NAK15] Nic NEWMAN, David A. L. LEVY et Rasmus KLEIS NIELSEN. *Digital News Report 2015*. Reuters Institute for the Study of Journalism, 2015, p. 112. URL : https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/files/Reuters%2520Institute%2520Digital%2520News%2520Report%25202015_Full%2520Report.pdf (visité le 12/07/2022).

[Nal+04] Ramesh NALLAPATI et al. « Event Threading within News Topics ». In : *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management - CIKM '04*. The Thirteenth ACM Conference. Washington, D.C., USA, 2004, p. 446. ISBN : 978-1-58113-874-0. DOI : 10.1145/1031171.1031258.

[NCG16] Thien Huu NGUYEN, Kyunghyun CHO et Ralph GRISHMAN. « Joint Event Extraction via Recurrent Neural Networks ». In : *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, 2016, p. 300-309. DOI : 10.18653/v1/N16-1034.

[Ned+13] Claire NEDELLEC et al. « Overview of BioNLP Shared Task 2013 ». In : *Proceedings of the BioNLP Shared Task 2013 Workshop*. BioNLP Shared Task 2013 Workshop. Août 2013, p. 7.

[New+21] Nic NEWMAN et al. *Digital News Report 2021*. 10. Reuters Institute for the Study of Journalism, 2021, p. 164. URL : https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf (visité le 12/07/2022).

[New12] Nic NEWMAN. *Digital News Report 2012*. Reuters Institute for the Study of Journalism, 2012, p. 68. URL : https://s3-eu-west-1.amazonaws.com/media.digitalnewsreport.org/wp-content/uploads/2012/05/Reuters-Institute-Digital-News-Report-2012.pdf (visité le 12/07/2022).

[New18] NEWSEYE. *NewsEye: A Digital Investigator for Historical Newspapers | NewsEye Project | Fact Sheet | H2020*. CORDIS | European Commission. 2018. URL : https://cordis.europa.eu/project/id/770299 (visité le 12/04/2022).

[NG15] Thien Huu Nguyen et Ralph Grishman. « Event Detection and Domain Adaptation with Convolutional Neural Networks ». In : *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Beijing, China, 2015, p. 365-371. doi : 10.3115/v1/P15-2060.

[Ngu+19] Thi-Tuyet-Hai Nguyen et al. « Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing ». In : *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). Champaign, IL, USA, juin 2019, p. 29-38. isbn : 978-1-72811-547-4. doi : 10.1109/JCDL.2019.00015.

[Ngu+20] Thi Tuyet Hai Nguyen et al. « Neural Machine Translation with BERT for Post-OCR Error Detection and Correction ». In : *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. JCDL '20: The ACM/IEEE Joint Conference on Digital Libraries in 2020. Virtual Event China, août 2020, p. 333-336. isbn : 978-1-4503-7585-6. doi : 10.1145/3383583.3398605.

[Ngu20] Thi Tuyet Hai Nguyen. « Facilitating Access to Historical Documents by Improving Digitisation Results ». Thèse de doct. Université de La Rochelle, 6 avr. 2020. url : https://tel.archives-ouvertes.fr/tel-03176609 (visité le 02/08/2022).

[NIS10] NIST. *The AQUAINT Project*. 22 déc. 2010. url : https://www-nlpir.nist.gov/projects/aquaint/ (visité le 05/07/2022).

[NNC19] Andrew Naoum, Joel Nothman et James Curran. « Article Segmentation in Digitised Newspapers with a 2D Markov Model ». In : *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 2019 International Conference on Document Analysis and Recognition (ICDAR). Sept. 2019, p. 1007-1014. doi : 10.1109/ICDAR.2019.00165.

[NNT12] Dung T. Nguyen, Nam P. Nguyen et My T. Thai. « Sources of Misinformation in Online Social Networks: Who to Suspect? » In : *MILCOM 2012 - 2012 IEEE Military Communications Conference*. MILCOM 2012 - 2012 IEEE Military Communications Conference. Orlando, FL, USA, oct. 2012, p. 1-6. isbn : 978-1-4673-1731-3 978-1-4673-1729-0 978-1-4673-1730-6. doi : 10.1109/MILCOM.2012.6415780.

[Oce17] Oceanic Exchanges Project Team. *Oceanic Exchanges: Tracing Global Information Networks In Historical Newspaper Repositories, 1840-1914*. 2017. url : https://doi.org/10.17605/OSF.IO/WA94S.

[Oiv+19] Mila Oiva et al. « Spreading News in 1904 ». In : *Media History* (11 août 2019), p. 18. ISSN : 1368-8804. URL : https://doi.org/10.1080/13688804.2019.1652090.

[Org20a] Organisation Mondiale de la santé. *Aplatissons la courbe de l'infodémie*. Sept. 2020. URL : https://www.who.int/fr/news-room/spotlight/let-s-flatten-the-infodemic-curve (visité le 28/04/2022).

[Org20b] Organisation Mondiale de la santé. *Gestion de l'infodémie sur la COVID-19 : Promouvoir des comportements sains et atténuer les effets néfastes de la diffusion d'informations fausses et trompeuses*. 23 sept. 2020. URL : https://www.who.int/fr/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation (visité le 28/04/2022).

[Pal+12] Thomas Palfray et al. « Logical Segmentation for Article Extraction in Digitized Old Newspapers ». In : *Proceedings of the 2012 ACM Symposium on Document Engineering - DocEng '12*. The 2012 ACM Symposium. Paris, France, sept. 2012, p. 129-132. ISBN : 978-1-4503-1116-8. DOI : 10.1145/2361354.2361383.

[Pat+21] Parth Patwa et al. « Overview of CONSTRAINT 2021 Shared Tasks: Detecting English COVID-19 Fake News and Hindi Hostile Posts ». In : *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Communications in Computer and Information Science. Cham, 2021, p. 42-53. ISBN : 978-3-030-73696-5. DOI : 10.1007/978-3-030-73696-5_5.

[Pet+18] Matthew E. Peters et al. « Deep Contextualized Word Representations ». In : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. T. 1. arXiv:1802.05365. New Orleans, 1er juin 2018. ISBN : 978-1-948087-27-8. arXiv : 1802.05365 [cs]. URL : http://arxiv.org/abs/1802.05365 (visité le 08/06/2022).

[Pin20] Roy Pinker. *Fake news & viralité avant Internet*. 4 juin 2020. 232 p. URL : https://www.cnrseditions.fr/catalogue/societe/fake-news-viralite-avant-internet/ (visité le 22/11/2020).

[PLS07] James Pustejovsky, Jessica Littman et Roser Saurí. « Arguments in TimeML: Events and Entities ». In : *Annotating, Extracting and Reasoning about Time and Events*. T. 4795. Lecture Notes in Computer Science. Berlin, Heidelberg, 2007, p. 107-126. ISBN : 978-3-540-75988-1. DOI : 10.1007/978-3-540-75989-8_8.

[PM10] Swit Phuvipadawat et Tsuyoshi Murata. « Breaking News Detection and Tracking in Twitter ». In : *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 2010 IEEE/ACM International Conference on Web Intelligence-Intelligent Agent Technology (WI-IAT). Toronto, AB, Canada, août 2010, p. 120-123. isbn : 978-1-4244-8482-9. doi : 10.1109/WI-IAT.2010.205.

[POL10] Sasa Petrovic, Miles Osborne et Victor Lavrenko. « Streaming First Story Detection with Application to Twitter ». In : *HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, Californi, USA, 2 juin 2010, p. 181-189. url : https://dl.acm.org/citation.cfm?id=1858020.

[Pol22] PolitiFact. *PolitiFact*. 2 mai 2022. url : https://www.politifact.com/ (visité le 02/05/2022).

[Pou+04] Bruno Pouliquen et al. « Multilingual and Cross-Lingual News Topic Tracking ». In : *Proceedings of the 20th International Conference on Computational Linguistics - COLING '04*. The 20th International Conference. Geneva, Switzerland, 2004, 959-es. doi : 10.3115/1220355.1220493.

[Pou+06] Kimler Pouliquen et al. « Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation ». In : *Proceedings of the 5th International Conference on Language Resources And Evaluation*. LREC' 2006. Genoa, Italy, 22 mai 2006, p. 53-58.

[PSD08] Bruno Pouliquen, Ralf Steinberger et Olivier Deguernel. « Story Tracking: Linking Similar News over Time and across Languages ». In : *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization - MMIES '08*. The Workshop. Manchester, United Kingdom, 2008, p. 49. isbn : 978-1-905593-51-4. doi : 10.3115/1613172.1613184.

[PSG19] Telmo Pires, Eva Schlinger et Dan Garrette. « How Multilingual Is Multilingual BERT? » 4 juin 2019. arXiv : 1906.01502 [cs]. url : http://arxiv.org/abs/1906.01502 (visité le 24/11/2020).

[PSM14] Jeffrey Pennington, Richard Socher et Christopher Manning. « Glove: Global Vectors for Word Representation ». In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, 2014, p. 1532-1543. doi : 10.3115/v1/D14-1162.

[PTV12] Pedro C. PINTO, Patrick THIRAN et Martin VETTERLI. « Locating the Source of Diffusion in Large-Scale Networks ». In : *Physical Review Letters* 109.6 (10 août 2012), p. 068702. ISSN : 0031-9007, 1079-7114. DOI : 10.1103/PhysRevLett.109.068702. arXiv : 1208.2534 [physics].

[Pus+03] James PUSTEJOVSKY et al. « TimeML: Robust Specification of Event and Temporal Expressions in Text ». In : New Directions in Question Answering 2003. Stanford, CA, USA, 2003, p. 28-34.

[Pus+04] James PUSTEJOVSKY et al. « The Specification Language TimeML ». In : (23 jan. 2004), p. 545-557.

[Pus+10] James PUSTEJOVSKY et al. « ISO-TimeML: An International Standard for Semantic Annotation ». In : *Proceedings of the International Conference on Language Resources and Evaluation*. International Conference on Language Resources and Evaluation. T. 10. Valetta, Malta, mai 2010, p. 394-397. URL : http://www.lrec-conf.org/proceedings/lrec2010/summaries/55.html.

[QIO20] Umair QAZI, Muhammad IMRAN et Ferda OFLI. « GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information ». 22 mai 2020. arXiv : 2005.11177.

[RA07] Yves RAIMOND et Samer ABDALLAH. *The Event Ontology*. 2007. URL : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.463.126&rep=rep1&type=pdf (visité le 17/11/2020).

[Rai+07] Yves RAIMOND et al. « The Music Ontology ». In : *ISMIR*. International Conference On Music Information Retrieval. Vienna, Austria, 23 sept. 2007, p. 6.

[RBS11] Manuel Gomez RODRIGUEZ, David BALDUZZI et Bernhard SCHÖLKOPF. *Uncovering the Temporal Dynamics of Diffusion Networks*. 3 mai 2011. arXiv : 1105.0697 [physics]. URL : http://arxiv.org/abs/1105.0697 (visité le 19/07/2022).

[Rc75] [Log.] Zdenko Podobny RAY SMITH et CONTRIBUTORS, *Tesseract-OCR* 1975. LIC : Apache License, Version 2.0. URL : https://tesseract-ocr.github.io/, VCS : https://github.com/tesseract-ocr/tesseract, SWHID : ⟨swh:1:dir:0cbf929886ec7efc5e064786407e2d78540a5b3c;origin=https://github.com/tesseract-ocr/tesseract⟩.

[RG] [Log.] Nils REIMERS et Iryna GUREVYCH, *Multilingual Sentence and Image Embeddings with BERT* version 2.2.0. URL : https://www.sbert.net/, VCS : https://github.com/UKPLab/sentence-transformers, SWHID : ⟨swh:1:dir:7114eb8623638895b165690850a1534d9f04aa9b;origin=https://github.com/UKPLab/sentence-transformers;⟩.

[RG19] Nils Reimers et Iryna Gurevych. « Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks ». In : *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, nov. 2019, p. 3982-3992. DOI : 10.18653/v1/D19-1410. arXiv : 1908.10084.

[RG20] Nils Reimers et Iryna Gurevych. *Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation*. 5 oct. 2020. arXiv : 2004.09813 [cs]. URL : http://arxiv.org/abs/2004.09813 (visité le 08/06/2022).

[Rie+11] Sebastian Riedel et al. « Model Combination for Event Extraction in BioNLP 2011 ». In : *Proceedings of BioNLP Shared Task 2011 Workshop*. BioNLP Shared Task 2011 Workshop. Juin 2011, p. 51-55.

[Ril95] Ellen Riloff. « An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains ». In : *Artificial Intelligence* 85 (jan. 1995), p. 101-134.

[Rit+12] Alan Ritter et al. « Open Domain Event Extraction from Twitter ». In : *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '12*. The 18th ACM SIGKDD International Conference. Beijing, China, 2012, p. 1104. ISBN : 978-1-4503-1462-6. DOI : 10.1145/2339530.2339704.

[RM12] Jan Rupnik et Andrej Muhič. « Cross-Lingual Document Retrieval through Hub Languages ». In : *Proceedings of xLiTe: Cross-Lingual Technologies, NIPS 2012 Workshop*. xLiTe: Cross-Lingual Technologies, NIPS 2012 Workshop. Lake Tahoe, Nevada, USA, jan. 2012, p. 5. URL : https://www.researchgate.net/publication/281581540_Cross-Lingual_Document_Retrieval_through_Hub_Languages.

[Rn17] Aliza Rosen et nabokov. *Giving You More Characters to Express Yourself*. 27 sept. 2017. URL : https://blog.twitter.com/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself (visité le 17/05/2022).

[Rob+97] Steven Robertson et al. « Okapi at TREC-5 ». In : *Nist Special Publication SP* (31 juill. 1997), p. 23.

[Rou87] Peter J. Rousseeuw. « Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis ». In : *Journal of Computational and Applied Mathematics* 20 (nov. 1987), p. 53-65. ISSN : 03770427. DOI : 10.1016/0377-0427(87)90125-7.

[Rud+19] Charlotte Rudnik et al. « Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata ». In : *Companion Proceedings of The 2019 World Wide Web Conference on - WWW '19*. Companion The 2019 World Wide Web Conference. San Francisco, USA, 2019, p. 1232-1239. isbn : 978-1-4503-6675-5. doi : 10.1145/3308560.3316761.

[Rup+16] Jan Rupnik et al. « News Across Languages - Cross-Lingual Document Similarity and Event Tracking ». In : *Journal of Artificial Intelligence Research* 55 (30 jan. 2016), p. 283-316. issn : 1076-9757. doi : 10.1613/jair.4780.

[Ryl09] Gilbert Ryle. *The Concept of Mind*. London ; New York, 2009. 314 p. isbn : 978-0-415-48547-0 978-0-203-87585-8.

[SA75] Roger C. Schank et Robert P. Abelson. « Scripts, Plans and Knowledge ». In : *Proceedings of the Fourth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization. T. 75. Tbilisi, Georgia, USSR, sept. 1975, p. 151-157. url : https://www.ijcai.org/Proceedings/75/Papers/021.pdf (visité le 22/03/2020).

[Sal+21] Hannu Salmi et al. « The Reuse of Texts in Finnish Newspapers and Journals, 1771–1920: A Digital Humanities Perspective ». In : *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 54.1 (2 jan. 2021), p. 14-28. issn : 0161-5440, 1940-1906. doi : 10.1080/01615440.2020.1803166.

[Sar+20] Samer Muthana Sarsam et al. « Sarcasm Detection Using Machine Learning Algorithms in Twitter: A Systematic Review ». In : *International Journal of Market Research* 62.5 (2020), p. 578-598.

[SC01] Nicolas Stokes et Joe Carthy. « Combining Semantic and Syntactic Document Classifiers to Improve First Story Detection ». In : *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM SIGIR Conference on Research and Development in Information Retrieval. Sept. 2001, p. 424-425.

[SC93] Beth M. Sundheim et Nancy A. Chinchor. « Survey of the Message Understanding Conferences ». In : *Proceedings of the Workshop on Human Language Technology - HLT '93*. The Workshop. Princeton, New Jersey, 1993, p. 56. isbn : 978-1-55860-324-0. doi : 10.3115/1075671.1075684.

[SCD13] David A. Smith, Ryan Cordell et Elizabeth Maddock Dillon. « Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers ». In : *2013 IEEE International Conference on Big Data*. 2013 IEEE International Conference on Big Data. Silicon Valley, CA, USA, oct. 2013, p. 86-94. isbn : 978-1-4799-1293-3. doi : 10.1109/BigData.2013.6691675.

[Sch+09] Ansgar SCHERP et al. « A Model of Events Based on a Foundational Ontology ». In : International Conference on Knowledge Capturing. Redondo Beach, CA, USA, 30 jan. 2009, p. 27.

[SGJ11] Jannik STRÖTGEN, Michael GERTZ et Conny JUNGHANS. « An Event-Centric Model for Multilingual Document Similarity ». In : *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information - SIGIR '11*. The 34th International ACM SIGIR Conference. Beijing, China, 2011, p. 953. ISBN : 978-1-4503-0757-4. DOI : 10.1145/2009916.2010043.

[Sha10] Ryan Benjamin SHAW. « Events and Periods as Concepts for Organizing Historical Knowledge ». UC Berkeley, 2010. URL : https://escholarship.org/uc/item/4111f1fw (visité le 02/12/2020).

[Sha13] Ryan SHAW. « A Semantic Tool for Historical Events ». In : *Proceedings of the The 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. The 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation. Atlanta, Georgia, USA, 13 juin 2013, p. 38-46.

[Shi+18] Bichen SHI et al. « Story Disambiguation: Tracking Evolving News Stories across News and Social Streams ». 16 août 2018. arXiv : 1808.05906 [cs, stat]. URL : http://arxiv.org/abs/1808.05906 (visité le 29/10/2019).

[SL18] Holger SCHWENK et Xian LI. « A Corpus for Multilingual Document Classification in Eight Languages ». In : *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan, 7 mai 2018. ISBN : 979-10-95546-00-9. arXiv : 1805.09821. URL : http://arxiv.org/abs/1805.09821 (visité le 17/06/2021).

[SL20] Lukas STANKEVIČIUS et Mantas LUKOŠEVIČIUS. « Testing Pre-Trained Transformer Models for Lithuanian News Clustering ». 3 avr. 2020. arXiv : 2004.03461 [cs]. URL : http://arxiv.org/abs/2004.03461 (visité le 02/07/2021).

[SL99] J Michael SCHULTZ et Mark LIBERMAN. « Topic Detection and Tracking Using Idf-Weighted Cosine Coefficient ». In : *Proceedings of the DARPA Broadcast News Workshop*. DARPA Broadcast News Workshop. T. 1892192. San Francisco, USA, 1999, p. 4.

[SMM22] João SANTOS, Afonso MENDES et Sebastião MIRANDA. *Simplifying Multilingual News Clustering Through Projection From a Shared Space*. 28 avr. 2022. arXiv : 2204.13418 [cs]. URL : http://arxiv.org/abs/2204.13418 (visité le 09/07/2022).

[SN20] Gautam Kishore Shahi et Durgesh Nandini. « FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19 ». In : *CoRR* abs/2006.11343 (2020), p. 9. URL : https://arxiv.org/abs/2006.11343.

[Sno22] Snopes. *Snopes.Com*. Snopes.com. 2 mai 2022. URL : https://www.snopes.com/ (visité le 02/05/2022).

[Sof22] [Log.] Explosion Software, *spaCy: Industrial-strength Natural Language Processing in Python* avr. 2022. DOI : 10.5281/zenodo.1212303, VCS : https://github.com/explosion/spaCy, SWHID : ⟨swh:1:dir:cfddc17dd28ef2939f38cf3b7350be588f459db1;origin=https://github.com/explosion/spaCy⟩.

[Son+15] Zhiyi Song et al. « From Light to Rich ERE: Annotation of Entities, Relations, and Events ». In : *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation. Denver, Colorado, 2015, p. 89-98. DOI : 10.3115/v1/W15-0812.

[Son+18] Zhiyi Song et al. « Cross-Document, Cross-Language Event Coreference Annotation Using Event Hoppers ». In : *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan, 7 mai 2018, p. 6.

[SPH02] Ralf Steinberger, Bruno Pouliquen et Johan Hagman. « Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC ». In : *Computational Linguistics and Intelligent Text Processing*. Réd. par Gerhard Goos, Juris Hartmanis et Jan van Leeuwen. T. 2276. Lecture Notes in Computer Science. Berlin, Heidelberg, 2002, p. 415-424. ISBN : 978-3-540-43219-7 978-3-540-45715-2. DOI : 10.1007/3-540-45715-1_44.

[Spr18] Rachele Sprugnoli. « Event Detection and Classification for the Digital Humanities ». Trento, Italia : Università degli Studi di Trento, avr. 2018. 229 p. URL : http://eprints-phd.biblio.unitn.it/2865/ (visité le 03/12/2019).

[SPV09] Ralf Steinberger, Bruno Pouliquen et Erik Van der Goot. « An Introduction to the Europe Media Monitor Family of Applications ». In : *SIGIR 2009 Workshop Proceedings*. SIGIR 2009. T. 43. Boston, MA, 14 déc. 2009, p. 24-28. DOI : 10.1145/1670564.1670568.

[SSG14] Miguel A Sanchez-Perez, Grigori Sidorov et Alexander Gelbukh. « The Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014 ». In : CLEF (Working Notes). Sept. 2014, p. 1004-1011.

[SSS02] Yusuke Shinyama, Satoshi Sekine et Kiyoshi Sudo. « Automatic Paraphrase Acquisition from News Articles ». In : *Proceedings of the Second International Conference on Human Language Technology Research* -. The Second International Conference. San Diego, California, 2002, p. 313-318. doi : 10.3115/1289189.1289218.

[ST17] R. Sprugnoli et S. Tonelli. « One, No One and One Hundred Thousand Events: Defining and Processing Events in an Inter-Disciplinary Perspective ». In : *Natural Language Engineering* 23.4 (juill. 2017), p. 485-506. issn : 1351-3249, 1469-8110. doi : 10.1017/S1351324916000292.

[Sta+19] Todor Staykovski et al. « Dense vs. Sparse Representations for News Stream Clustering ». In : *Proceedings of Text2Story - 2nd Workshop on Narrative Extraction From Texts, Co-Located with the 41st European Conference on Information*. T. 2342. Cologne, Germany, 14 avr. 2019, p. 47-52. url : https://ceur-ws.org/Vol-2342/paper6.pdf.

[Ste+04] Ralf Steinberger et al. « Providing Cross-Lingual Information Access with Knowledge-Poor Methods ». In : Informatica 28. T. 28. Slovenia, 2004, p. 415-423.

[Ste+15] Ralf Steinberger et al. « Observing Trends in Automated Multilingual Media Analysis ». In : *Proceedings of the Symposium of New Frontiers of Automated Content Analysis in the Social Sciences*. ACA 2015. Zürich, Switzerland, 1er juill. 2015, p. 20.

[STH09] Ryan Shaw, Raphaël Troncy et Lynda Hardman. « LODE: Linking Open Descriptions of Events ». In : *The Semantic Web*. Réd. par David Hutchison et al. T. 5926. Lecture Notes in Computer Science. Berlin, Heidelberg, 2009, p. 153-167. isbn : 978-3-642-10870-9 978-3-642-10871-6. doi : 10.1007/978-3-642-10871-6_11.

[Suá+22] Avelino Suárez et al. *Changement climatique et biodiversité*. Document technique V du GIEC. Groupe intergouvermental sur l'évolution du climat (GIEC), avr. 2022, p. 89.

[Sun95] Beth M Sundheim. « Overview of Results of the MUC-6 Evaluation ». In : *Proceedings of the 6th Conference on Message Understanding*. 6th Conference on Message Understanding. Nov. 1995, p. 13-31. isbn : 978-1-55860-402-5. doi : 10.3115/1072399.1072402.

[SW17] Sahar Sohangir et Dingding Wang. « Improved Sqrt-Cosine Similarity Measurement ». In : *Journal of Big Data* 4.1 (déc. 2017), p. 25. issn : 2196-1115. doi : 10.1186/s40537-017-0083-6.

[TDP19] Ian Tenney, Dipanjan Das et Ellie Pavlick. « BERT Rediscovers the Classical NLP Pipeline ». 9 août 2019. arXiv : 1905.05950 [cs]. url : http://arxiv.org/abs/1905.05950 (visité le 24/06/2021).

[tea11] The DAnIEL team. *DAnIEL Multilingual Epidemic Surveillance, Annotation Guidelines*. Nov. 2011. URL : https://daniel.greyc.fr/guidelines.pdf (visité le 20/04/2020).

[Tha+21] Nandan Thakur et al. *Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks*. 12 avr. 2021. arXiv : 2010.08240 [cs]. URL : http://arxiv.org/abs/2010.08240 (visité le 08/06/2022).

[The20] The Wikimedia Foundation. *Wikipedia article depth*. 10 sept. 2020. URL : https://meta.wikimedia.org/wiki/Wikipedia_article_depth (visité le 09/12/2020).

[The21a] The Wikimedia Foundation. *List of Wikipedias*. In : *Wikipedia*. 6 avr. 2021. URL : https://en.wikipedia.org/w/index.php?title=List_of_Wikipedias&oldid=1016309550 (visité le 07/04/2021).

[The21b] The Wikimedia Foundation. *Wikipedia:Summary Style*. In : *Wikipedia*. 2 avr. 2021. URL : https://en.wikipedia.org/w/index.php?title=Wikipedia:Summary_style&oldid=1015628666 (visité le 06/04/2021).

[The22] The Wikimedia Foundation. *List of Wikipedias by Speakers per Article - Meta*. 31 juill. 2022. URL : https://meta.wikimedia.org/wiki/List_of_Wikipedias_by_speakers_per_article (visité le 31/07/2022).

[TN12] Mitja Trampuš et Blaž Novak. « The Internals Of An Aggregated Web News Feed ». In : *In Proceedings of the 15th Multiconference on Information Society 2012*. IS-2012. Ljubljana, Slovenia, 2012, p. 221-224. URL : https://www.researchgate.net/profile/Mitja_Trampus/publication/260385767_INTERNALS_OF_AN_AGGREGATED_WEB_NEWS_FEED/links/544a98b00cf2bcc9b1d2f706/INTERNALS-OF-AN-AGGREGATED-WEB-NEWS-FEED.pdf.

[TWH06] Robert Tibshirani, Guenther Walther et Trevor Hastie. « Estimating the Number of Clusters in a Data Set via the Gap Statistic ». In : *Journal of the Royal Statistical Society* Series B.63 (6 jan. 2006), p. 411-423. DOI : 10.1111/1467-9868.00293.

[UzZ+13] Naushad UzZaman et al. « SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations ». In : *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Second Joint Conference on Lexical and Computaional Semantics (*SEM). T. 2. Atlanta, Georgia, USA, 14 juin 2013, p. 1-9. URL : https://www.aclweb.org/anthology/S13-2001.pdf.

[Vas+17] Ashish Vaswani et al. « Attention Is All You Need ». In : *Advances in neurol information processing systems* 30 (2017), p. 11.

[VBG18] Lise Volkart, Bouillon, Pierrette et Girletti, Sabrina. « Statistical vs. Neural Machine Translation: A Comparison of MTH and DeepL at Swiss Post's Language Service ». In : *Proceedings of the 40th Conference Translating and the Computer*. 40th Conference Translating and the Computer. London, United-Kingdom, 15 nov. 2018, p. 145-150. isbn : 978-2-9701095-5-6. url : https://archive-ouverte.unige.ch/unige:111777 (visité le 12/07/2022).

[Ven57] Zeno Vendler. « Verbs and Times ». In : *The Philosophical Review* 66.2 (2 avr. 1957), p. 143-160. url : http://links.jstor.org/sici?sici=0031-8108%28195704%2966%3A2%3C143%3AVAT%3E2.0.CO%3B2-2.

[Ven67] Zeno Vendler. *Linguistics in Philosophy*. 1967. isbn : 978-1-5017-4372-6. doi : 10.7591/9781501743726.

[Ver+07] Marc Verhagen et al. « SemEval-2007 Task 15: TempEval Temporal Relation Identification ». In : *Proceedings of the 4th International Workshop on Semantic Evaluations - SemEval '07*. The 4th International Workshop. Prague, Czech Republic, 2007, p. 75-80. doi : 10.3115/1621474.1621488.

[Ver+10] Marc Verhagen et al. « SemEval-2010 Task 13: TempEval-2 ». In : *SemEval '10: Proceedings of the 5th International Workshop on Semantic Evaluation*. SemEval '10: 5th International Workshop on Semantic Evaluation. Uppsala, Sweden, 15 juill. 2010, p. 57-62. url : https://dl.acm.org/doi/abs/10.5555/1859664.1859674.

[Ves+17] Aleksi Vesanto et al. « Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910 ». In : (2017), p. 5.

[vHag+11] Willem Robert van Hage et al. « Design and Use of the Simple Event Model (SEM) ». In : *Journal of Web Semantics* 9.2 (juill. 2011), p. 128-136. issn : 15708268. doi : 10.1016/j.websem.2011.03.003.

[VK14] Denny Vrandečić et Markus Krötzsch. « Wikidata: A Free Collaborative Knowledgebase ». In : *Communications of the ACM* 57.10 (23 sept. 2014), p. 78-85. issn : 0001-0782, 1557-7317. doi : 10.1145/2629489.

[VLH18] Cynthia Van Hee, Els Lefever et Véronique Hoste. « SemEval-2018 Task 3: Irony Detection in English Tweets ». In : *Proceedings of The 12th International Workshop on Semantic Evaluation*. SemEval 2018. New Orleans, Louisiana, juin 2018, p. 39-50. doi : 10.18653/v1/S18-1005.

[Voo00] Ellen M. Voorhees. « Overview of the TREC-9 Question Answering Track ». In : *In Proceedings of the Ninth Text REtrieval Conference*. The Ninth Text REtrieval Conference (TREC 9). T. 500–249. NIST Special Publication. Gaithersburg, Maryland, USA, 13 nov. 2000. url : http://trec.nist.gov/pubs/trec9/papers/qa%5C_overview.pdf (visité le 23/03/2020).

[Voo99] Ellen M. VOORHEES. « The TREC-8 Question Answering Track Report ». In : *Proceedings of The Eighth Text REtrieval Conference, TREC 1999*. The Eighth Text REtrieval Conference, TREC 1999. NIST Special Publication. Gaithersburg, Maryland, USA, nov. 1999. URL : https://trec.nist.gov/pubs/trec8/papers/qa_report.pdf (visité le 23/03/2020).

[Vos+05] Piek VOSSEN et al. « NewsReader: Recording History from Daily News Streams ». In : (2005), p. 8.

[Vos+16] Piek VOSSEN et al. « NewsReader: Using Knowledge Resources in a Cross-Lingual Reading Machine to Generate More Knowledge from Massive Streams of News ». In : *Knowledge-Based Systems* 110 (oct. 2016), p. 60-85. ISSN : 09507051. DOI : 10.1016/j.knosys.2016.07.013.

[VT11] Pierre-yves VANDENBUSSCHE et Charles TEISSÈDRE. « Events Retrieval Using Enhanced Semantic Web Knowledge ». In : Workshop DeRIVE 2011 (Detection, Representation, and Exploitation of Events in the Semantic Web). Bonn, Germany, oct. 2011, p. 6.

[VVP08] Iraklis VARLAMIS, Vasilis VASSALOS et Antonis PALAIOS. « Monitoring the Evolution of Interests in the Blogosphere ». In : *2008 IEEE 24th International Conference on Data Engineering Workshop*. 2008 IEEE 24th International Conference on Data Engineeing Workshop (ICDE Workshop 2008). Cancun, Mexico, avr. 2008, p. 513-518. ISBN : 978-1-4244-2161-9 978-1-4244-2162-6. DOI : 10.1109/ICDEW.2008.4498371.

[Wan+11] Dashun WANG et al. « Information Spreading in Context ». In : *Proceedings of the 20th International Conference on World Wide Web - WWW '11*. The 20th International Conference. Hyderabad, India, 2011, p. 735. ISBN : 978-1-4503-0632-4. DOI : 10.1145/1963405.1963508.

[Wan+21a] Jiexin WANG et al. « Improving Question Answering for Event-Focused Questions in Temporal Collections of News Articles ». In : *Information Retrieval Journal* 24.1 (fév. 2021), p. 29-54. ISSN : 1386-4564, 1573-7659. DOI : 10.1007/s10791-020-09387-9.

[Wan+21b] Apurva WANI et al. « Evaluating Deep Learning Approaches for Covid19 Fake News Detection ». In : *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Communications in Computer and Information Science. Cham, 2021, p. 153-163. ISBN : 978-3-030-73696-5. DOI : 10.1007/978-3-030-73696-5_15.

[WJD03] Richard WALDINGER, Peter JARVIS et Jennifer DUNGAN. « Program Synthesis for Multi-agent Question Answering ». In : *Verification: Theory and Practice*. Réd. par Gerhard GOOS, Juris HARTMANIS et Jan van LEEUWEN. T. 2772. Lecture Notes in Computer Science. Berlin, Heidelberg, 2003, p. 747-761. ISBN : 978-3-540-21002-3 978-3-540-39910-0. DOI : 10.1007/978-3-540-39910-0_32.

[WJY21] Jiexin Wang, Adam Jatowt et Masatoshi Yoshikawa. « Event Occurrence Date Estimation Based on Multivariate Time Series Analysis over Temporal Document Collections ». In : *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual Event Canada, 11 juill. 2021, p. 398-407. isbn : 978-1-4503-8037-9. doi : 10.1145/3404835.3462885.

[WJY22] Jiexin Wang, Adam Jatowt et Masatoshi Yoshikawa. *ArchivalQA: A Large-scale Benchmark Dataset for Open Domain Question Answering over Historical News Collections*. 21 fév. 2022. arXiv : 2109.03438 [cs]. url : http://arxiv.org/abs/2109.03438 (visité le 23/05/2022).

[WL11] Jianshu Weng et Bu-Sung Lee. « Event Detection in Twitter ». In : *Proceedings of the Fifth International Conference on Weblogs and Social Media*. Fifth International Conference on Weblogs and Social Media. Barcelona, Catalonia, Spain, juill. 2011, p. 401-408. url : http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2767.

[WWC82] Friedrich M Wahl, Kwan Y Wong et Richard G Casey. « Block Segmentation and Text Extraction in Mixed Text/Image Documents ». In : *Computer Graphics and Image Processing* 20.4 (4 1982), p. 375-390. doi : 10.1016/0146-664X(82)90059-4.

[XW19] Wei Xiang et Bang Wang. « A Survey of Event Extraction From Text ». In : *IEEE Access* 7 (nov. 2019), p. 173111-173137. issn : 2169-3536. doi : 10.1109/ACCESS.2019.2956831.

[Yan+00a] Yiming Yang et al. « Improving Text Categorization Methods for Event Tracking ». In : *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '00*. The 23rd Annual International ACM SIGIR Conference. Athens, Greece, 24 juill. 2000, p. 65-72. isbn : 978-1-58113-226-7. doi : 10.1145/345508.345550.

[Yan+00b] Roman Yangarber et al. « Automatic Acquisition of Domain Knowledge for Information Extraction ». In : COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics. 2000.

[Yan+19] Sen Yang et al. « Exploring Pre-trained Language Models for Event Extraction and Generation ». In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019, p. 5284-5294. doi : 10.18653/v1/P19-1522.

[Yan+20] Zhilin Yang et al. « XLNet: Generalized Autoregressive Pretraining for Language Understanding ». 2 jan. 2020. arXiv : 1906.08237 [cs]. url : http://arxiv.org/abs/1906.08237 (visité le 30/07/2021).

[Yan+21] Ziyi YANG et al. *Universal Sentence Representation Learning with Conditional Masked Language Model*. 10 sept. 2021. arXiv : 2012.14388 [cs]. URL : http://arxiv.org/abs/2012.14388 (visité le 24/07/2022).

[YB18] Vikas YADAV et Steven BETHARD. « A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models ». In : *Proceedings of the 27th International Conference on Computational Linguistics* (août 2018), p. 14.

[YPC98] Yiming YANG, Tom PIERCE et Jaime CARBONELL. « A Study on Retrospective and On-Line Event Detection ». In : *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, août 1998, p. 28-36. ISBN : 978-1-58113-015-7. DOI : 10.1145/290941.290953.

[Zar+17] Ali ZAREZADE et al. « Correlated Cascades: Compete or Cooperate ». In : *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1 (10 fév. 2017). ISSN : 2374-3468, 2159-5399. DOI : 10.1609/aaai.v31i1.10483.

[Zha+16] Zi-Ke ZHANG et al. « Dynamics of Information Diffusion and Its Applications on Complex Networks ». In : *Physics Reports* 651 (sept. 2016), p. 1-34. ISSN : 03701573. DOI : 10.1016/j.physrep.2016.07.002.

[Zha+22] Yunyi ZHANG et al. *Unsupervised Key Event Detection from Massive Text Corpora*. 3 juill. 2022. DOI : 10.1145/3534678.3539395. arXiv : 2206.04153 [cs].

[Zhu+22] Wenzhen ZHU et al. *DocBed: A Multi-Stage OCR Solution for Documents with Complex Layouts*. 3 fév. 2022. arXiv : 2202.01414 [cs]. URL : http://arxiv.org/abs/2202.01414 (visité le 21/06/2022).

[ZJS19] Tongtao ZHANG, Heng JI et Avirup SIL. « Joint Entity and Event Extraction with Generative Adversarial Imitation Learning ». In : *Data Intelligence* 1.2 (mai 2019), p. 99-120. ISSN : 2641-435X. DOI : 10.1162/dint_a_00014.

[ZK20] Brett ZONGKER et Leah KNOBEL. *Library of Congress Completes Digitization of 23 Early Presidential Collections | Library of Congress*. Library of Congress Completes Digitization of 23 Early Presidential Collections. 17 déc. 2020. URL : https://www.loc.gov/item/prn-20-085/library-of-congress-completes-digitization-of-23-early-presidential-collections/2020-12-17/ (visité le 30/05/2022).

[ZXZ21] Chengqing ZONG, Rui XIA et Jiajun ZHANG. « Topic Detection and Tracking ». In : *Text Data Mining*. Singapore, 2021, p. 201-225. ISBN : 9789811601002. DOI : 10.1007/978-981-16-0100-2_9.

Détection et suivi d'événements dans des documents historiques

Résumé : Les campagnes actuelles de numérisation de documents historiques issus de fonds documentaires du monde entier ouvrent de nouvelles voies aux historiens, historiennes et spécialistes des sciences sociales. La compréhension des événements du passé se renouvelle par l'analyse de ces grands volumes de données historiques : découdre le fil des événements, tracer de fausses informations sont, entre autres, des possibilités offertes par les sciences du numérique.
Cette thèse s'intéresse à ces articles de presse historique et propose, à travers deux stratégies que tout oppose, deux processus d'analyse répondant à la problématique de suivi des événements dans la presse. Un cas d'utilisation simple est celui d'une équipe de recherche en humanités numériques qui s'intéresse à un événement particulier du passé. Ses membres cherchent à découvrir tous les documents de presse qui s'y rapportent. L'analyse manuelle des articles est irréalisable dans un temps contraint. En publiant à la fois algorithmes, jeux de données et analyses, cette thèse est un premier jalon vers la publication d'outils plus sophistiqués. Nous permettons à tout individu de fouiller les fonds de presse ancienne à la recherche d'événements, et pourquoi pas, renouveler certaines de nos connaissances historiques.

Mots clés : Événements en traitement du langage · Suivi d'événements · Documents de presse historique

Detection and Tracking of Events in Historical Press Documents

Abstract: Current campaigns to digitise historical documents from all over the world are opening up new avenues for historians and social science researchers. The understanding of past events is renewed by the analysis of these large volumes of historical data : unravelling the thread of events, tracing false information are, among other things, possibilities offered by the digital sciences.
This thesis focuses on these historical press articles and suggests, through two opposing strategies, two analysis processes that address the problem of tracking events in the press. A simple use case is for instance a digital humanities researcher or an amateur historian who is interested in an event of the past and seeks to discover all the press documents related to it. Manual analysis of articles is not feasible in a limited time. By publishing algorithms, datasets and analyses, this thesis is a first step towards the publication of more sophisticated tools allowing any individual to search old press collections for events, and why not, renew some of our historical knowledge.

Keywords: Events in natural language processing · Event tracking · Historical press documents

Laboratoire Informatique, Image, Interaction
Institut LUDI - La Rochelle Université
Avenue Michel Crépeau

17 042 LA ROCHELLE CEDEX 1