Détection des fraudes : de l'image à la sémantique du contenu : application à la vérification des informations extraites d'un corpus de tickets de caisse

Chloé Artaud

UNIVERSITÉ DE LA ROCHELLE

École doctorale Euclide

Laboratoire Informatique, Image, Interaction (L3i)

THÈSE présentée par :

Chloé ARTAUD

soutenue le : **6 février 2019**
pour obtenir le grade de : **Docteur de l'université de La Rochelle**
Discipline : **Informatique**

Détection des fraudes : de l'image à la sémantique du contenu

Application à la vérification des informations extraites d'un corpus de tickets de caisse

---

JURY :

Patrice Bellot	Professeur, Aix-Marseille Université, Rapporteur
Nicole Vincent	Professeure, Université Paris Descartes, Rapportrice
Vincent Claveau	Chargé de recherche, IRISA, CNRS, Examinateur
Béatrice Daille	Professeure, Université de Nantes, Examinatrice
Petra Gomez-Krämer	Maîtresse de Conférence, Université de La Rochelle, Examinatrice
Antoine Doucet	Professeur, Université de La Rochelle, Co-Directeur de thèse
Jean-Marc Ogier	Professeur, Université de La Rochelle, Co-Directeur de thèse
Véronique Serfaty	Tutrice, Direction Générale de l'Armement, Invitée

Remerciements

Tout d'abord, je tiens à remercier mes directeurs, Antoine Doucet et Jean-Marc Ogier, pour avoir proposé ce sujet, à la fois riche et complexe, et pour m'avoir encadrée et encouragée pendant ces trois années et quelques mois. Leur confiance et leur expérience ont été de véritables atouts pour moi, et j'espère avoir l'occasion de collaborer à nouveau avec eux par la suite.

Je remercie les membres de mon jury d'avoir accepté d'évaluer ces travaux de thèse : Nicole Vincent et Patrice Bellot en tant que rapporteurs·rices, Vincent Claveau, Béatrice Daille et Petra Gomez-Krämer en tant qu'examinateurs·rices, ainsi que Véronique Serfaty en tant qu'invitée, en sa qualité de tutrice de la Direction Générale de l'Armement.

Je tiens également à remercier mes financeurs, sans qui cette thèse n'aurait pas été possible, et qui, par leurs demandes de rapports réguliers m'ont bien souvent permis de faire le point sur ma thèse et d'avancer : la Direction Générale de l'Armement ainsi que l'ex-région Poitou-Charentes, intégrée dans la Nouvelle Aquitaine.

Je me dois également de remercier mon laboratoire, le L3i, pour m'avoir offert tout le confort et le soutien matériel nécessaire à une thèse en informatique, ainsi que le département d'informatique de la Faculté des Sciences, l'IAE et l'IUT Technique de Commercialisation pour m'avoir permis d'enseigner en leurs murs et de découvrir et d'exercer cette belle facette du métier d'enseignant-chercheur.

A special thanks to the CASIA (Institute of Automation of the Chinese Academy of Sciences) to welcome me during two months in the NLPR (National Laboratory of Pattern Recognition). Thanks to Cheng-Lin Liu and his team for the organization of this stay and the very warm welcome they gave us.

Pour leur aide précieuse et salutaire, je tiens à exprimer ma profonde gratitude envers Vincent Poulain d'Andecy et Nicolas Sidère. Merci Vincent pour ton soutien et tes conseils réguliers qui m'ont permis d'avancer et de mieux mesurer les enjeux de cette thèse. Merci Nicolas pour avoir cru en mes tickets de caisse, m'avoir aidée à construire ce [magnifique] corpus, et puis bien sûr, pour tes encouragements et tes relectures !

Merci à toutes les personnes qui m'ont aidée durant cette thèse, que ce soit sur l'aspect technique, dans la construction de mon corpus ou encore pour l'évaluation de mon travail :

— Christophe, Joseph et Van pour avoir partagé leurs codes et m'avoir aidée dans l'utilisation des serveurs ; Mickaël pour avoir pris un peu de temps pour les réglages du dos numérique ;

— tou·te·s les donneurs·ses de tickets de caisse, anonymes et non-anonymes ;
— tou·te·s les correcteurs·rices d'OCR, dont la liste exhaustive des prénoms et surnoms se trouve ici : `http://receipts.univ-lr.fr/best-correctors/` ;
— les 25 fraudeurs·ses d'un jour, dont je tairai le nom pour des raisons évidentes de protection des sources ;-)
— mes 5 détectives humains préférés (par ordre alphabétique) : Florian, Jordan, Nicolas, Timothée et Valentin ;
— mes annotateurs les plus courageux : Damien et Guillaume S.

Merci à tou·te·s pour votre disponibilité et votre générosité.

Sur un plan plus personnel, je tiens à mettre ici à l'honneur toutes les personnes qui ont fait de ce doctorat une période enrichissante humainement et, bien que régulièrement éprouvante, finalement très épanouissante.

Je pense d'abord à l'ADocs, association des jeunes chercheurs de La Rochelle, qui, au-delà de la découverte du monde associatif, m'a permis de rencontrer des gens merveilleux, doctorants et jeunes chercheurs d'autres laboratoires et horizons, sans qui la thèse n'aurait clairement pas eu la même saveur !

Une pensée particulière pour Alice et nos longues discussions engagées ou sentimentales, pour nos accords et désaccords, pour avoir été là dans les moments compliqués comme dans les moments heureux : merci.

Aux anciens et aux nouveaux du labo, à ceux qui sont partis et ceux qui sont arrivés, à ceux qui étaient déjà là et qui resteront encore un peu, merci de m'avoir supportée et accordé une telle place au sein du L3i. Votre amitié me touche, et vos taquineries m'encouragent.

Un remerciement particulier pour Guillaume C., pour les longues soirées de discussions autour d'un verre ou d'un plat et pour son écoute si attentionnée : merci.

Des « special thanks » aux personnes ayant partagé mon bureau, anciennes ou nouvelles, qui m'ont supportée tout au long de ces années : Sovann, Imen, Bruno, Marcela, Damien, Jordan et Salah.

Un chaleureux merci également à ma famille : à mes parents pour leur présence et leurs relectures, et à mes sœurs, pour avoir placé la barre si haut ! J'envoie beaucoup d'amour à mon neveu Antoine, ainsi qu'à ma nièce Camille et à mon neveu-filleul Maxime, tous deux nés pendant ma thèse.

Merci également à mes « amis d'ailleurs », loin mais proches à la fois : mes amies de lycée *aka* « Les filles », mes amis de prépa et les quelques (trop rares) MIKH qui ont chaque fois été une très précieuse source de chaleur et une bouffée d'oxygène, et puis, Diane, évidemment, pour sa patience légendaire et son amitié indéfectible.

Enfin, ces derniers mois n'auraient pas eu la même intensité sans Guillaume S. : merci à toi de m'avoir soutenue au quotidien, sur place ou à 11 000 kilomètres de distance. A toi de jouer, maintenant !

Sommaire

Table des figures

Liste des tableaux

Introduction

« **Saint-Priest : elle escroque 73 500 euros à Pôle Emploi avec des faux bulletins de salaire**

Entre juillet 2012 et juin 2017, une comptable qui ne parvenait plus à joindre les deux bouts a perçu indûment 73 500 euros d'indemnités de chômage en falsifiant ses contrats de travail et ses bulletins de paie.

Des crédits, des dettes... puis l'escroquerie

Après des contrats à durée déterminée dans plusieurs sociétés, cette femme aujourd'hui âgée de 54 ans et habitant Saint-Priest, a décidé de se mettre à son compte. Mais l'expérience a échoué, elle a pris des crédits, a divorcé et, croulant sous les dettes, elle est passée à l'escroquerie.

Des faux bulletins de paie avec un salaire de 5 000 euros

Pour constituer un dossier auprès de Pôle Emploi, elle a utilisé ses anciens contrats de travail, dont elle a modifié les dates et allongé la durée. Puis son expérience de comptable lui a permis de créer des faux bulletins de paie, en s'octroyant par exemple un salaire de 5 000 euros là où elle n'en avait gagné que 1 800.

Mais des erreurs ont fini par attirer l'attention de Pôle Emploi, qui a déposé plainte. Entendue ce mercredi par la brigade de la délinquance astucieuse de la Sûreté (BDA), elle a reconnu l'escroquerie. Elle sera jugée ultérieurement. » [a]

a. Article du journal *Le Progrès*, publié en ligne le 27/09/2018 sur : `https://www.leprogres.fr/rhone-69-edition-lyon-metropole/2018/09/27/elle-percoit-73-500-euros-d-indemnites-de-chomage-en-falsifiant-des-documents`, consulté le 28/09/2018.

Contexte

La lutte contre la fraude est un enjeu économique très important, que ce soit pour les entreprises, pour les administrations ou pour les particuliers. Plusieurs études tendent à montrer que les entreprises cherchent de plus en plus à se doter d'outils performants de détection des fraudes, quel que soit leur domaine d'activité. C'est notamment le cas des études successives du cabinet d'étude *PriceWaterhouseCoopers* (Di Giovanni 2016, Lavion 2018) : en 2016, 36% des entreprises interrogées dans le monde entier déclaraient avoir été victimes de fraude au cours des 24 derniers mois, contre 49% en 2018 (Lavion 2018). Cette augmentation est, comme le souligne ce rapport, probablement plus due à une amélioration de la détection des fraudes qu'à une réelle augmentation de leur nombre durant cette période. Toujours selon cette étude, 42% des entreprises ayant répondu ont augmenté leurs dépenses pour la lutte contre la fraude au cours des 24 derniers mois et 44% veulent l'augmenter dans les 24 prochains. En France, en 2016, ce sont 68% des entreprises qui avaient détecté des fraudes au cours des 24 mois précédents (Di Giovanni 2016)).

Cette prise de conscience de la nécessité de lutter efficacement contre la fraude ressort également d'un rapport de l'entreprise SAS datant de 2014 [1] dans lequel on apprend qu'un peu plus d'un quart des compagnies d'assurance interrogées avaient un système automatisé de détection des fraudes en place ou en développement lors de l'enquête. Sur les compagnies d'assurance restantes, un peu moins de la moitié avaient pour projet d'en mettre en place. Il y a fort à parier que quatre ans plus tard, les technologies s'étant considérablement améliorées, de nombreuses compagnies d'assurances aient suivi le mouvement de la transformation numérique et travaillent à enrichir et exploiter leurs nombreuses données. C'est en tout cas ce que préconise pour toutes les entreprises le rapport PWC, qui voit dans l'intelligence artificielle et l'analyse de données le moyen de combattre la fraude efficacement.

En ce qui concerne les montants de la fraude, il est extrêmement difficile d'évaluer les pertes subies par les entreprises qui sont générées par des actes frauduleux, d'une part, parce que la fraude n'est pas encore bien détectée par les entreprises elles-mêmes, d'autre part parce que les entreprises ne font pas forcément état des fraudes subies, notamment lorsque cela pourrait avoir des répercussions sur la confiance de leurs clients à leur égard. Par exemple, une cyber-attaque sur un réseau social ou une plateforme de commerce en ligne pourrait, si la communication est mal gérée et si la fraude n'est pas détectée et réparée très rapidement, générer une perte d'utilisateurs qui n'auraient plus confiance dans le produit. Cela se traduirait nécessairement par des pertes financières dues à des ruptures de contrat avec d'éventuels partenaires, par exemple publicitaires, mais cela est très difficilement évaluable à court-terme.

L'Agence pour la Lutte contre la Fraude à l'Assurance (ALFA) a confié en exclusivité

1. Le rapport de cette étude est disponible sur le site de l'entreprise SAS : `https://www.sas.com/content/dam/SAS/fr_fr/doc/other1/etude-fraude-assurance.pdf`

à l'Argus de l'assurance un rapport [2] qui estime à 2,5 milliards d'euros le montant annuel de la seule fraude à l'assurance en France. En 2015, seuls 250 millions de ce montant ont été récupérés par les assurances, ce qui est un résultat cependant en forte augmentation par rapport aux années précédentes.

Dans un autre registre, les administrations publiques sont également victimes de fraudes ayant un lourd poids économique pour la société. La Direction Nationale de Lutte contre la Fraude (DNLF) répertorie trois grands types de fraudes [3] :

— la fraude fiscale (19,5 milliards d'euros en 2016, contre 21,2 milliards en 2015, année exceptionnelle)
— la fraude aux cotisations sociales (601 millions d'euros en 2016, contre 497 millions en 2015)
— la fraude aux prestations sociales (724 millions d'euros en 2016, contre 673 millions en 2015)

L'augmentation de la détection de ces fraudes est due à plusieurs facteurs : une politique renforcée de dialogue et d'échange d'informations au sein de l'OCDE et de l'Union Européenne, ainsi que la mise en place de procédures de *data mining* au sein des administrations. Pôle Emploi a ainsi mis en place une solution de *text mining* visant à détecter les offres d'emploi frauduleuses, ainsi qu'une solution de détection d'anomalies dans les données de ses bases.

Les institutions ne sont pas les seules à être victimes de fraudes : les particuliers eux-mêmes peuvent être victimes d'arnaques. Lors d'un achat de voiture, certains fraudeurs peuvent par exemple falsifier la carte grise en changeant la date de construction de la voiture. De même, lors de la vente d'une maison, il peut être tentant de modifier les diagnostiques énergétiques afin d'attirer plus d'acheteurs potentiels... Il n'est pas rare non plus de constituer un dossier avec de faux bulletins de salaires pour pouvoir accéder à un logement, dont le loyer doit être trois fois inférieur aux revenus, ce qui est une condition à laquelle un·e jeune employé·e parisien·ne peut difficilement répondre, même pour un studio.

En terme de nombre de procédures relatives à la fraude documentaire et/ou identitaire, ce sont 6039 procédures pour faux documents d'identité, 3922 procédures pour faux documents concernant la circulation des véhicules et 4800 procédures pour autres faux documents administratifs qui ont été enregistrées en 2016 par les services de police ou unités de gendarmerie, selon une note de l'Observatoire National de la Délinquance et des Réponses Pénales (ONDRP) [4]. La police aux frontières a quant à elle intercepté 6199 documents français impliqués dans des fraudes en 2015, dont plus de la moitié (3219)

2. La synthèse de ce rapport, publiée le 15 septembre 2016, est disponible sur : `https://www.argusdelassurance.com/institutions/fraude-a-l-assurance-une-facture-de-2-5-md-en-dommages-en-2014-alfa.96060`.

3. Tous les chiffres sont extraits du bilan « Lutte contre la fraude » publié par la DNLF pour l'année 2016, pp. 8, 74 et 82 : `https://www.economie.gouv.fr/files/files/directions_services/dnlf/bilan2016-dnlf.pdf`.

4. Note de l'Observatoire National de la Délinquance et des Réponses Pénales (ONDRP), « Éléments de connaissance sur la fraude aux documents et à l'identité en 2016 », Septembre 2017, disponible sur : `https://inhesj.fr/sites/default/files/ondrp_files/publications/pdf/note_16.pdf`.

Les différents types de documents français interceptés selon la nature de la fraude en 2015 en France

	Ensemble des documents	Titres de séjour(2)	Visas	Cartes d'identité	Passeports	Permis de conduire	Actes d'état civil(3)	Composteurs et timbres	Divers(4)
Total faux documents français(1)	6 199	788	148	853	495	268	413	15	3 219
Contrefaçons	2 258	153	41	128	14	90	188	14	1 630
Falsifications	954	68	14	32	77	15	47	1	700
Usages frauduleux	1 441	164	10	413	193	14	85	0	562
Obtentions frauduleuses	1 513	403	83	279	208	148	93	0	299
Volés vierges	33	0	0	1	3	1	0	0	28

Source : DGPN, DCPAF

(1) Tous types

(2) Carte résident, carte séjour CEE, carte séjour temporaire, certificat OFPRA, certificat résident Algérien, récépissé carte séjour, récépissé statut réfugié

(3) Certificats de nationalité française, actes de naissance (certificats), actes de naissance de l'Outre-mer, actes de naissance du SCEC de Nantes

(4) Attestation d'accueil, attestation d'assurance, certificat d'immatriculation, certificat médical, justificatif de domicile, etc.

FIGURE 1 – Extrait du rapport annuel 2016 de l'ONDRP (Tableau 4, page 10).

sont des documents divers, « servant à la vie de tous les jours » [5]. La figure 1 montre les différents types de documents français interceptés par la police aux frontières et les types de fraudes dont ils font l'objet. Ces divers types de fraudes et de documents sont définis ci-après.

Définitions et typologies

Fraudes et faux documents

Nous voyons dans la figure 1 une distinction entre 5 types de fraudes concernant les documents. Ces types de fraudes sont expliquées comme suit dans le rapport annuel 2016 de l'ONDRP :

- La contrefaçon : production intégrale par imitation d'un document d'identité.
- La falsification : modification d'un ou plusieurs éléments d'un document authentique. La falsification peut porter sur la date de validité, sur les mentions d'identité ou encore sur la photographie.

5. Rapport annuel 2016 de l'ONDRP, « Éléments de connaissance sur la fraude aux documents et à l'identité en 2015 », Janvier 2017, disponible sur : `https://inhesj.fr/sites/default/files/ondrp_files/publications/rapports-annuels/2016/2016_RA_fraude_doc.pdf`.

— Les volés vierges : documents authentiques ayant été volés avant leur personnalisation et qui seront ensuite complétés par le voleur, le receleur ou le faussaire devenant ainsi des falsifications.
— L'usage frauduleux : usurpation d'identité ou utilisation du document authentique appartenant à un tiers.
— L'obtention frauduleuse : document authentique délivré sur la base de faux documents (actes de naissance, justificatifs de domicile, déclarations de perte, etc.) pouvant être contrefaits, falsifiés, usurpés ou obtenus indûment.

Cette typologie des fraudes, qui relève de la Direction centrale de la Police aux frontières (DCPAF), permet d'éclaircir le terme obscur de « faux document ». Cependant, comme le souligne un rapport du Défenseur des Droits [6], il n'existe pas dans la loi de notion de « fraude » à proprement parler, permettant ainsi à l'administration de définir la fraude d'une manière très large, incluant les erreurs et les oublis, arguant du fait que les déclarations et/ou les documents fournis comme justificatifs sont faux, dans le sens défini par la loi. En effet, le code pénal définit la notion de faux de la façon suivante :

> « Constitue un faux toute altération frauduleuse de la vérité, de nature à causer un préjudice et accomplie par quelque moyen que ce soit, dans un écrit ou tout autre support d'expression de la pensée qui a pour objet ou qui peut avoir pour effet d'établir la preuve d'un droit ou d'un fait ayant des conséquences juridiques. » [7]

La fausse déclaration est également définie comme suit : « le fait de fournir sciemment une fausse déclaration ou une déclaration incomplète en vue d'obtenir ou de tenter d'obtenir, de faire obtenir ou de tenter de faire obtenir d'une personne publique, d'un organisme de protection sociale ou d'un organisme chargé d'une mission de service public une allocation, une prestation, un paiement ou un avantage indu » [8].

Ainsi, le délit de faux comprend le fait de fabriquer un document entièrement faux (ce sera donc une contrefaçon) ou de modifier frauduleusement un document (augmenter son salaire sur son bulletin de salaire, augmenter le nombre de jours d'arrêt maladie sur un certificat médical...). Imiter une signature est également un cas de faux [9].

La DNLF estime quant à elle que la « fraude documentaire » sert de « fraude support » [10] aux autres fraudes, en fournissant des faux documents justificatifs à l'administration. C'est, dans la typologie de l'ONDRP, le cas de « l'obtention frauduleuse ». Par exemple, pour une fraude fiscale, on pourrait fournir une fausse attestation de revenus ou une fausse déclaration de patrimoine. Dans le cas d'une fraude aux cotisations sociales, une entreprise peut mentir sur le nombre de ses employé·e·s. Dans le cas de

6. Rapport du Défenseur des Droits, « Lutte contre la fraude aux prestations sociales : à quel prix pour les droits des usagers ? », Septembre 2017, disponible sur : `https://www.defenseurdesdroits.fr/sites/default/files/atoms/files/rapportfraudessociales-v6-06.09.17_0.pdf`.
7. Article 441-1 du Code Pénal.
8. Article 441-6 du Code Pénal.
9. `https://www.service-public.fr/particuliers/vosdroits/F31612`.
10. Termes extraits du site de la DNLF : `https://www.economie.gouv.fr/dnlf/lutte-contre-fraude-documentaire`.

fraudes aux prestations sociales, on pourrait fournir un arrêt maladie comportant des dates inexactes...

L'Office européen de Lutte Anti Fraude (OLAF) de la Commission européenne, dans son guide sur la détection des faux documents OLAF (2014), émet également une définition très large du « faux document », mais intègre cependant une dichotomie intéressante entre le fond et la forme du document, ou entre son aspect physique et son contenu intellectuel :

> « Un faux document est un document dont le caractère authentique a été altéré : le document n'est donc plus conforme à la réalité. L'altération peut être :
> — physique : un document peut être modifié physiquement (suppression d'éléments ou de références, ajout manuscrit d'informations altérant le document, par exemple) ;
> — intellectuelle : le contenu du document n'est plus conforme à la réalité (description inexacte des services rendus, contenu erroné d'un rapport, apposition de fausses signatures sur la liste de présence, par exemple) ».

Par ailleurs, ce guide rappelle que seul un juge ou une juridiction peut être en mesure de qualifier un document de contrefaçon ou de falsification, et donc établir qu'il y a une fraude. C'est donc consciemment que nous utiliserons parfois le terme « faux document » de façon erronée aux yeux de la loi car, même si nos documents ne sont destinés qu'à des fins de recherche, et en aucun cas à des fins de fraudes, la visée applicative de notre travail est bien de fournir un outil de détection des fraudes via l'utilisation de documents falsifiés ou contrefaits.

Documents

Les fraudes documentaires aux yeux de la loi concernent donc des documents à destination des administrations publiques, qui sont définis par le Code des Relations entre le Public et l'Administration, qui remplace la loi du 17 juillet 1978, de la façon suivante :

> « Sont considérés comme documents administratifs, au sens des titres Ier, III et IV du présent livre, quels que soient leur date, leur lieu de conservation, leur forme et leur support, les documents produits ou reçus, dans le cadre de leur mission de service public, par l'Etat, les collectivités territoriales ainsi que par les autres personnes de droit public ou les personnes de droit privé chargées d'une telle mission. Constituent de tels documents notamment les dossiers, rapports, études, comptes rendus, procès-verbaux, statistiques, instructions, circulaires, notes et réponses ministérielles, correspondances, avis, prévisions et décisions. » [11]

L'information présente sur ces documents peut être très variée, puisqu'il s'agit de tout document émis ou destiné à l'administration. Cette définition ne prend donc pas en compte tous les documents que nous voulons exploiter dans notre travail de recherche :

11. Article L300-2 du Code des Relations entre le Public et l'Administration.

en effet, les documents échangés entre les entreprises (comme les factures, les devis) ou les entreprises et les particuliers (bulletins de salaire, bons de livraison...) ne sont pas pris en compte. De même, cette définition prend en compte des documents qui ne nous intéressent pas vraiment car ils ne font pas l'objet d'échanges et ne sont donc pas sujets à falsification ou fraude, comme les actes, les circulaires, les notes ministérielles... Ce n'est donc pas à proprement parler de documents administratifs dont nous traiterons durant notre recherche mais plutôt d'un mélange de documents juridiques, commerciaux, financiers, administratifs et autres pièces justificatives pouvant être réclamés lors d'un échange entre des administrations, des entreprises ou/et des particuliers.

Nous distinguons donc les documents non sur leur dénomination mais sur leur parcours, sur leur valeur communicationnelle. Deux autres façons de distinguer les documents sont décrites dans van Renesse (1997) : la première repose sur la sécurité intrinsèque du document et l'autre sur le type de valeur du document. La typologie des valeurs du document distingue cinq types : la valeur directe (valeur inconditionnelle et immédiate), la valeur indirecte (supporte un droit ou une transaction), la valeur conditionnelle (nécessite une inspection), la valeur informative et la valeur fictive. En parallèle, les documents sont distingués selon trois niveaux de sécurité : haute sécurité (passeport), moyenne sécurité (tickets d'entrée) et basse sécurité (facture). Selon l'auteur, ces deux approches doivent être combinées pour adapter la recherche de la fraude.

Les sécurités présentes dans certains documents, comme les passeports ou les cartes d'identité, rendent plus difficile la fraude, et plus facile la détection des documents falsifiés. Il existe d'ailleurs de nombreux travaux de recherche et de nombreuses applications de la détection des faux papiers d'identité, notamment dans le cadre de la surveillance aux frontières, comme nous l'explique le guide du bureau de Nations Unies (*United Nations Office on Drugs and Crime*) (UNODC 2010). Nous nous concentrerons donc plutôt sur les documents qui ne présentent pas de protection, puisque l'on peut considérer que le problème est traité en amont dans les documents d'identité et de voyage.

La définition légale du document administratif présentée précédemment prend en compte tout type de supports de l'information, qu'il soit sous forme de papier A4, cartonné, plastifié, fichier texte électronique, image, vidéo, bande-son... Or nous ne traiterons que de documents contenant du texte, et parfois des images, d'origine numérique et suivant une certaine structure, et non de documents multimédias, de manuscrits ou de textes littéraires par exemple.

Nous pouvons faire une analogie entre les documents administratifs et la typologie différenciant les documents informatiques structurés, semi-structurés et non-structurés. Ces types de documents que nous traiterons pendant cette thèse seraient des documents non-structurés : ils ne contiennent que du langage naturel ou des images sans éléments structurels qui expliquent que telle partie est un texte ou que telle partie du texte concerne tel type d'information. Les documents non-structurés s'opposent ainsi aux documents structurés que sont les bases de données ou les tableaux, où toutes les informations, ou données, sont réduites à leur forme la plus atomique possible et où c'est la structure, figée, qui porte le sens des données. Entre les documents structurés et les documents non-structurés, se trouve toutes sortes de documents semi-structurés (Madani et al. 2013).

Dire d'un document qu'il contient des données semi-structurées signifie qu'il contient des données textuelles hétérogènes et que l'organisation de ces données est souvent implicite et relativement flexible, selon Abiteboul (1997). Ainsi, les documents XML ou JSON contenant des phrases sont des documents semi-structurés, de même qu'un texte annoté avec des étiquettes ou des labels. Les documents qui nous intéressent ne sont pas parfaitement structurés, dans le sens où ils ne contiennent pas des données parfaitement formatées et décrites. Cependant, ces documents possèdent pour la plupart une forme de structure, bien qu'elle ne soit pas explicite comme dans un document XML. Nous pourrions donc dire que ce sont des documents quasi-semi-structurés, dans le sens où la structure, la mise en page des éléments est significative, mais généralement implicite.

Cet intitulé de type de document est encore imprécis, puisqu'il peut être matériel ou immatériel. Dans le premier cas, le document est palpable, préhensible, physique. C'est du papier, du papier cartonné (ancien permis de conduire, carte d'électeur), du plastique (carte vitale, carte d'identité, nouveau permis de conduire)... Dans le deuxième cas, le document est numérique, c'est-à-dire qu'il est accessible depuis un ordinateur ou un smartphone, par exemple. Dans le cadre d'un travail sur la détection des fraudes automatisée, il faut que les documents soient sous forme numérique. Pour cela, nous pouvons numériser les documents physiques avec un scanner par exemple, mais aussi avec d'autres moyens de capture d'image comme un appareil photographique, une caméra ou un smartphone. Certains documents administratifs existent sous les deux formes : les avis d'imposition peuvent par exemple être consultables sur Internet et être reçus sous format papier, ou bien avoir été imprimés, ou numérisés, pour des raisons d'usage. On dit alors que ce sont des documents hybrides. Les administrations doivent en effet s'adapter à la transition numérique et proposer leurs services en ligne. Cela présente de nombreux avantages pour les administrés ainsi que pour les administrations : un document numérique sur une plateforme sécurisée peut être accessible à tout moment et de partout et le traitement des données récoltées numériquement se fait automatiquement, alors que les documents complétés à la main reçus par les administrations nécessitent dans la plupart des cas une intervention humaine. Mise à l'échelle d'un pays, la transformation numérique devrait donc représenter des gains de temps et d'argent non négligeables.

Numériser un document permet de lui appliquer des traitements automatisés, et notamment d'en extraire le contenu, c'est-à-dire les éléments qui le composent : la structure, les images et le texte. Le contenu est le niveau le plus informatif du document : il détaille toutes les informations nécessaires pour le traitement d'un dossier. Le contenu peut être composite : un document peut contenir des images, du texte, des tableaux de chiffres... Chaque élément d'un document peut contenir des informations sémantiques, c'est-à-dire du sens. Ces informations sémantiques peuvent parfois être redondantes. Par exemple, le logo d'une entreprise est généralement accompagné du nom de cette entreprise ainsi que de ses coordonnées. Ces trois éléments font sens, chacun séparément, mais également ensemble : le logo correspond à une entreprise qui est implantée à une adresse, qui a un numéro de téléphone, un numéro de fax... Les documents ont donc un contenu composé d'éléments textuels et graphiques structurés logiquement et agencés dans l'espace du document. La forme de ces éléments peut également porter un sens intrinsèque puisqu'elle

correspond généralement au sens de son contenu. Par exemple, le tableau d'un bulletin de salaire ou d'un avis d'imposition signifie qu'il correspond à un document financier qui comporte des chiffres, des taux, des sommes et des totaux. Il s'agit alors de comprendre le sens de chaque élément ainsi que de ce qui les lie entre eux. Dans le travail de détection des faux documents, il faudrait donc pouvoir modéliser toutes les informations, à la fois sémantiques et graphiques, concernant les éléments et la structure, afin de pouvoir les comparer à d'autres informations, réputées vraies, et détecter ainsi les fausses informations ou les incohérences entre les informations.

Problématique

Les documents dont nous traiterons dans cette thèse sont donc des documents composés d'une structure implicite et d'éléments textuels et graphiques, numériques ou numérisés, dont la propriété principale est qu'ils servent de preuves ou de justificatifs dans le cadre d'échanges d'informations entre des individus, des entreprises et des administrations. Ce sont en effet ces documents qui peuvent être falsifiés par l'une des parties de l'échange. A ce titre, il est fréquent que certains documents soient modifiés ou imités, et donc faux, afin d'obtenir des avantages (gagner plus d'argent ou de temps, obtenir un emploi ou une allocation) ou d'éviter des inconvénients (payer des taxes, faire des démarches pour renouveler des certificats ou des ordonnances, être expulsé d'un pays...) Nous pouvons également noter que certaines modifications de documents interviennent justement pour corriger des erreurs (médecin se trompant dans les dates de certificats ou mauvaise orthographe de noms propres par exemple) ou pour se mettre en conformité, sur le papier, avec la loi (obligation de vie commune pour des partenaires liés par le PACS ou pour des époux, par exemple). Ces modifications sont tout de même considérées comme des fraudes, et ces documents comme des faux.

La falsification de documents communs, tels que les factures, les certificats, les quittances et autres attestations, si elle était réservée aux faussaires il y a quelques décennies, est devenue un jeu d'enfant avec les nouvelles technologies. En effet, il est facile et peu onéreux de scanner un document papier de format A4, de l'ouvrir avec n'importe quel éditeur d'image, de changer quelques chiffres ou d'effacer quelques lignes, et de réimprimer le papier, dans le même format. Seuls les documents d'identité et de voyage ont actuellement des systèmes de sécurité intrinsèques difficilement imitables.

Chaque document doit donc être contrôlé par l'entité (entreprise, administration ou particulier) qui les reçoit afin de détecter les éventuels faux documents. Cette vérification ne peut pas être faite sur les millions de documents reçus au quotidien par les entreprises ou par les administrations par des personnes expertes, d'autant que Schetinger et al. (2017) montrent que les humains ne détectent que peu efficacement les images falsifiées. Une solution pour aider les différents acteurs d'échanges de documents à la détection des fraudes est donc, comme le souligne le rapport de la DNLF, celui de l'entreprise SAS et l'étude de PwC, l'utilisation des méthodes analytiques et d'intelligence artificielle.

Le sujet de notre travail s'inscrit donc dans cette problématique délicate qu'est la lutte contre les fraudes : il s'agit de détecter les faux documents, qu'ils soient falsifiés ou

contrefaits. Pour cela, nous avons choisi de proposer une approche qui se concentre sur les informations textuelles contenues dans le document : nous cherchons à vérifier automatiquement que les informations d'un document sont cohérentes entre elles et qu'elles sont vraisemblables.

Cette problématique doit faire face à de nombreux verrous technologiques et scientifiques dus à la complexité de l'objet d'étude (le document, ou plutôt la multitude de documents et de types de documents), à la variété des informations à vérifier et au peu de travaux ayant été effectués sur ce sujet.

Organisation du manuscrit

La volonté de se protéger de la circulation de faux documents n'est pas récente. En effet, dès les années 1960, des brevets ont été déposés, à la fois pour sécuriser les documents et à la fois pour les détecter (van Renesse 1997). Dès lors, les progrès techniques et technologiques ont fortement fait avancer la lutte contre la fraude, notamment en matière de sécurité intrinsèque aux documents. Les travaux en vision par ordinateur et en traitement de l'image ont également permis d'améliorer la détection de modifications d'images, et l'arrivée des techniques d'apprentissage profond mettent largement en péril les fraudeurs. Dans un autre contexte, depuis quelques années, nous voyons également l'émergence du *fact checking*, c'est-à-dire la vérification des faits énoncés dans l'actualité par une investigation automatisée ou non. En effet, face à l'abondance de l'information disponible et à la possibilité pour tout un chacun de devenir créateur et éditeur de contenu, sur le Web notamment, il est important de pouvoir offrir des outils de détection de fausses informations (« *Fake news* »), de rumeurs ou de canulars (« *hoax* »). Nous procèderons dans le **chapitre 1** à une revue de ces travaux et nous verrons qu'ils ne peuvent répondre que partiellement à notre problématique.

Les corpus existants utilisés dans les recherches sur la détection de faux documents sont généralement créés pour les besoins spécifiques de la tâche à effectuer. Ainsi, plusieurs corpus de faux documents ont été créés automatiquement en générant des structures prédéfinies et en insérant des données aléatoires. Du bruit, aléatoire ou non, ou des modifications manuelles sont ensuite apportées au document. Si ces corpus se révèlent utiles pour des tâches de traitement de l'image, ils nous sont inutiles pour tester notre approche. Nous verrons donc dans le **chapitre 2** comment nous avons affronté ce problème en créant un corpus de documents originaux – des tickets de caisse, anonymes et dont les gens acceptent volontiers de se débarrasser – que nous avons collectés puis numérisés. Afin de travailler sur les informations textuelles que contiennent ce document, nous avons extrait le texte de ces images en utilisant un logiciel de reconnaissance de caractères (« OCR » pour « Optical Character Recognition »). Nous avons ensuite corrigé automatiquement puis manuellement les textes issus de l'OCR, notamment grâce à une interface mise en place pour la correction collaborative de ces textes. Enfin, nous avons fait modifier environ 12% de notre corpus (images et textes correspondants) à des non-experts afin d'obtenir un corpus de faux documents réalistes. Nous en analyserons

les limites et proposerons des perspectives d'améliorations en conclusion de ce deuxième chapitre.

Ce corpus de textes exploitables, vrais et faux, nous permet de tester et valider notre approche qui consiste à vérifier les informations. Pour réaliser cette approche, il nous faut tout d'abord extraire et modéliser les informations du document et donc en avoir une analyse fine. Après avoir décrit les informations contenues dans les documents de notre corpus, ainsi que les informations supplémentaires que l'on peut trouver de façon commune dans les autres documents, nous expliquerons dans le **chapitre 3** le choix que nous avons fait d'utiliser une ontologie pour modéliser les informations du document au vu de l'état de l'art de l'extraction d'information et de l'ingénierie des connaissances. Nous présenterons le modèle ontologique que nous avons construit pour représenter les informations des documents et mettre en évidence les liens sémantiques entre elles. Nous détaillerons ensuite le processus qui nous permet d'extraire chaque information du document et d'en peupler l'ontologie. Nous évaluerons enfin ce processus d'extraction d'information, de façon globale et en fonction de chacune des informations extraites.

L'un des problèmes principaux auxquels nous nous sommes confrontés dans nos documents, est que les tickets de caisse, comme les autres documents d'ailleurs, contiennent de nombreuses abréviations, du fait de la structure étroite du document et des pratiques des éditeurs de tickets de caisse. Or, pour rechercher des informations dans des bases de données ou de connaissances, ou encore sur Internet, il faut pouvoir fournir des requêtes compréhensibles. Le **chapitre 4** explique donc, après un rapide état de l'art de la gestion des abréviations et une typologie détaillée des abréviations présentes dans le corpus, comment nous avons tenté d'associer des abréviations issues de notre corpus à des possibles expansions venant d'une source externe. Etant donné l'absence de ressource linguistique sur les abréviations, nous avons construit nous même un corpus et une vérité terrain de noms de produits tels qu'inscrits sur les tickets de caisse et de noms de produits détaillés et complets tels qu'on peut les trouver sur le web. Nous détaillerons les algorithmes utilisés et nous évaluerons cette approche.

La résolution des abréviations et l'organisation des informations dans une ontologie nous permettent ensuite de passer à la dernière étape dans la résolution de notre problème initial : la vérification des informations. Afin d'obtenir un premier benchmark sur la détection des faux documents, nous présenterons dans le **chapitre 5** la compétition que nous avons organisée dans le cadre de la conférence ICPR (*International Conference on Pattern Recognition*) et les résultats que les candidats ont obtenus. Nous proposerons ensuite plusieurs approches pour détecter les faux documents :

— La détection des incohérences au sein du document (par exemple, le fait que le montant payé ne soit pas égal au montant total, ou alors que le montant total n'est pas égal à la somme des prix...)
— La détection des incohérences au sein du corpus (il serait anormal par exemple qu'un produit qui apparait sur plusieurs tickets de caisse soit subitement trois fois plus cher...)
— La détection des invraisemblances par comparaison avec des informations externes

(si le site internet d'un restaurant indique un menu à 15€, il serait étonnant qu'il soit à 30€ sur la facture. De même, si un magasin est introuvable sur Internet, il se peut qu'il n'existe pas...)

Si la dernière approche reste à l'état de proposition, les deux autres approches seront expliquées, illustrées et évaluées. Nous chercherons ensuite à combiner les différents indices extraits afin d'obtenir les meilleurs résultats de détection des faux documents. Nous discuterons enfin nos résultats par rapport à ceux obtenus lors de la compétition et nous proposerons des perspectives d'améliorations de notre approche.

Tout ce travail sera enfin repris dans la conclusion, qui rappellera chacune des contributions apportées au cours de cette thèse et proposera des perspectives pour de futurs travaux.

Chapitre 1

La détection des fausses images et des fausses informations

Sommaire

Le guide édité par les Nations Unies (UNODC 2010) annonce une volonté de renforcer les capacités des États et des laboratoires de sciences légales dans la détection des faux documents d'identité, des documents contenant des sécurités et d'autres types de documents qui n'ont pas de sécurité. Ce guide, en plus de lister les différentes compétences que doivent avoir les examinateurs de documents, les différentes bases de données pour vérifier les informations et les différents outils utiles à la détection des fraudes, recense les caractéristiques physiques des documents qui doivent être analysées, comparées et évaluées. Ces caractéristiques sont :

— Les caractéristiques du support (papier/polymère)
— Les caractéristiques liées à l'encre
— Les indices liés aux processus d'impression
— Les indices de sécurité
— Les caractéristiques physiques du document (assemblage et production)
— Les techniques de personnalisation/bio-données
— Les moyens électroniques (puces, bandes magnétiques...)

Ces caractéristiques ne sont pas détectables par un système purement informatique. En effet, l'analyse de l'encre et du support sont des activités de chimie ou de physique et la lecture de certains indices de sécurité nécessite des équipements spéciaux (lumière à rayon ultra-violet, microscopes, lecteurs de bandes magnétiques, compte-fils...). Cependant, si les techniques de détection de superposition d'écritures ou d'impressions étaient analysées physiquement jusque dans les années 2000, les outils informatiques prennent une place de plus en plus importantes dans la détection de faux, que ce soit sur l'analyse du support ou format du document ou sur l'analyse de son contenu, pour reprendre la distinction effectuée dans OLAF (2014).

La fraude ou l'acte criminel, quand elle est réalisée dans un cadre numérique, est étudié par une discipline appelée « l'informatique légale » ou la « criminalistique numérique » (*computational/digital forensics* en anglais). Ce champ est assez large et recouvre notamment tout ce qui peut être produit devant une cour de justice, c'est-à-dire les preuves de toutes sortes de délits, de crimes et d'alibis, commis à l'aide d'un outil informatique, comme par exemple l'identification de pédophiles regardant des vidéos de pédopornographie, la localisation de la source d'une cyber-attaque, l'horodatage d'une connexion pouvant servir d'alibi, etc...

Böhme et al. (2009), voyant la croissance tout azimut de ces domaines de recherche autour des *Forensics* a cherché à en clarifier les contours et à définir ses sous-domaines. Ils établissent ainsi une première dichotomie entre les *Digital Forensics* (ce qui a trait à la recherche de preuves numériques de fraudes) et les *Analog Forensics* (ce qui a trait à la recherche de preuves physiques de fraudes). Le domaine des *Digital Forensics* est à son tour divisé en deux parts moins évidentes à différencier : les *Multimedia Forensics* et les *Computer Forensics*. Là où les deux sous-domaines cherchent à prouver des fraudes via l'outil informatique, le deuxième s'intéresse aux données brutes produites par l'ordinateur (température, enregistrement d'activités, traces de stockage de fichiers...) tandis que le premier porte sur des médias produits — et potentiellement modifiés — par l'homme. Ces recherches sur les preuves numériques de fraudes, ou sur l'implication d'outils infor-

matiques dans un processus de fraude, ont émané de plusieurs milieux différents, dans la recherche comme dans le milieu judiciaire. Le manuel de Ho & Li (2015) souhaite ainsi rapprocher les communautés de l'informatique et de traitement du signal, ainsi que les méthodes utilisées dans les laboratoires de sciences légales et leurs recherches en cours dans le monde académique.

Plusieurs survols de l'état de l'art plus spécialisé sur l'investigation des fraudes multimédia ont été réalisés au début des années 2010, cherchant à faire le point sur les différentes directions prises pour authentifier les différents types de médias. Si Böhme et al. (2009) nomment ce domaine de recherche *Multimedia Forensics*, Poisel & Tjoa (2011) parlent eux de *Digital Forensics* et Stamm et al. (2013) évoquent quant à eux le domaine des *Information Forensics*.

Étonnamment, dans les différents papiers cités précédemment, et de manière générale dans la littérature, le domaine des « multimédias » recouvrent trois médias : l'image, le son et la vidéo. Ces formats sont effectivement ceux traditionnellement couverts par le champ du traitement du signal, premier pourvoyeur des travaux en détection des fraudes. Cependant, selon le Trésor de la Langue Française informatisé (TLFi) [1], « media » est l'abréviation de « mass-media » qui signifie : « Ensemble des moyens de diffusion de masse de l'information, de la publicité et de la culture, c'est-à-dire des techniques et des instruments audiovisuels et graphiques, capables de transmettre rapidement le même message à destination d'un public très nombreux. » La 9^{e} édition du dictionnaire de l'Académie Française [2] quant à elle donne à « média » la définition suivante : « Tout moyen de communication servant à transmettre et à diffuser des informations, des œuvres ». Le média est donc, en accord avec son étymologie, le moyen, le support, le format, qui permet de transmettre de l'information, voire un message. En accord avec cette définition, Cox et al. (2008) définissent le *medium* comme le support (le CD sur lequel est enregistrée une chanson, ou le papier sur lequel est imprimée une image) et la « création » (*Work*) comme étant les parties (texte, image, chanson...) du contenu porté par le medium, en accord avec la dénomination des *United States copyright laws*. Stamm et al. (2013) définissent les *Information Forensics* comme visant à « déterminer l'authenticité, l'historique du traitement et l'origine du contenu multimédia numérique sans recourir, ou peu, à des canaux secondaires autres que le contenu numérique lui-même ». Les « informations » sont donc celles que l'on peut extraire d'une analyse du document numérique, c'est-à-dire de l'image, du son ou de la vidéo.

A ce titre, un vecteur d'information essentiel, et souvent oublié, est le texte : que ce soit dans la presse papier, dans la presse en ligne, sur les réseaux sociaux, ou sur des documents, c'est la plupart du temps le texte qui porte l'information, quand l'image, le son ou la vidéo ne sont que l'illustration, la mise en forme ou le support du texte. On notera d'ailleurs que le texte a cette ambiguïté qu'il est à la fois support de l'information et qu'il est souvent porté lui-même par un autre « support » (papier, image de texte, enregistrement sonore de la parole...). On a donc souvent besoin de traiter un média pour en extraire le texte sous sa forme numérique.

1. `http://atilf.atilf.fr/tlf.htm`
2. `http://www.cnrtl.fr/definition/academie9/m%C3%A9dia`

Nous verrons ainsi dans un premier temps de ce chapitre comment les documents falsifiés peuvent être détectés par des techniques de traitement de l'image au sens large, puis, dans un deuxième temps, par des techniques d'analyse d'images de documents, avec les particularités graphiques qui leurs sont propres. Cependant, « l'altération intellectuelle » (OLAF 2014) ne peut être détectée si l'on ne s'intéresse pas au contenu sémantique du document. C'est pourquoi nous nous intéresserons dans un troisième temps aux recherches sur les fausses informations, qui s'accroissent depuis quelques années en réponse à l'émergence de *Fake News* ainsi qu'aux scandales successifs des affaires liées aux lancements d'alertes (Panama Papers, Swiss Leaks, Lux Leaks...).

1.1 Détection des fausses images de document

L'échange de documents a toujours été sujet à suspicion sur l'authenticité du document reçu, sur l'identification de l'auteur ou de l'éditeur, ou sur son intégrité physique. De nombreuses recherches ont été effectuées pour sécuriser l'échange et authentifier les images envoyées ou reçues, qu'ils soient sous format papier ou format numérique. Nous verrons dans cette section que les méthodes de traitement de l'image pour la sécurisation des images échangées et la détection des fausses images peuvent s'appliquer au images de documents.

1.1.1 Des *Image Forensics* aux *Document Forensics*

Sur les trois types de contenus multimédias présentés précédemment, nous ne nous intéresserons qu'aux images, excluant tous les documents qui ne peuvent pas exister sur support papier.

Les images présentent de très nombreuses caractéristiques et chacune d'elles peut servir d'indice pour la détection d'une falsification, chacune d'elles peut être une trace d'une manipulation frauduleuse. Piva (2013) et Birajdar & Mankar (2013) ont simultanément présenté une synthèse des travaux réalisés sur les *Image Forensics*. Les deux articles commencent par faire une distinction des travaux sur la sécurité des contenus entre les approches « actives » et « passives ». Les approches « actives » concernent l'authentification des documents par la vérification d'une signature apposée en amont, tandis que les approches « passives » (ou « aveugle » pour Birajdar & Mankar (2013)) englobent toutes les méthodes de détection des faux documents basée sur l'analyse du document final sans données *a priori*. Stamm et al. (2013) considèrent que ces approches portent respectivement sur les « empreintes extrinsèques », c'est-à-dire les indices apportés au média pour le sécuriser, et « empreintes intrinsèques », indices contenus de fait dans le média de par sa conception ou sa manipulation.

La figure 1.1, issue de Birajdar & Mankar (2013), représente une classification des travaux effectués dans le domaine des *Image Forensics*. Plus récemment, Korus (2017) a fourni un état de l'art très complet du domaine, en explorant tous les champs de l'authentification d'images, de l'analyse de la source à la vérification de l'intégrité de l'image en passant par l'analyse de sa cohérence.

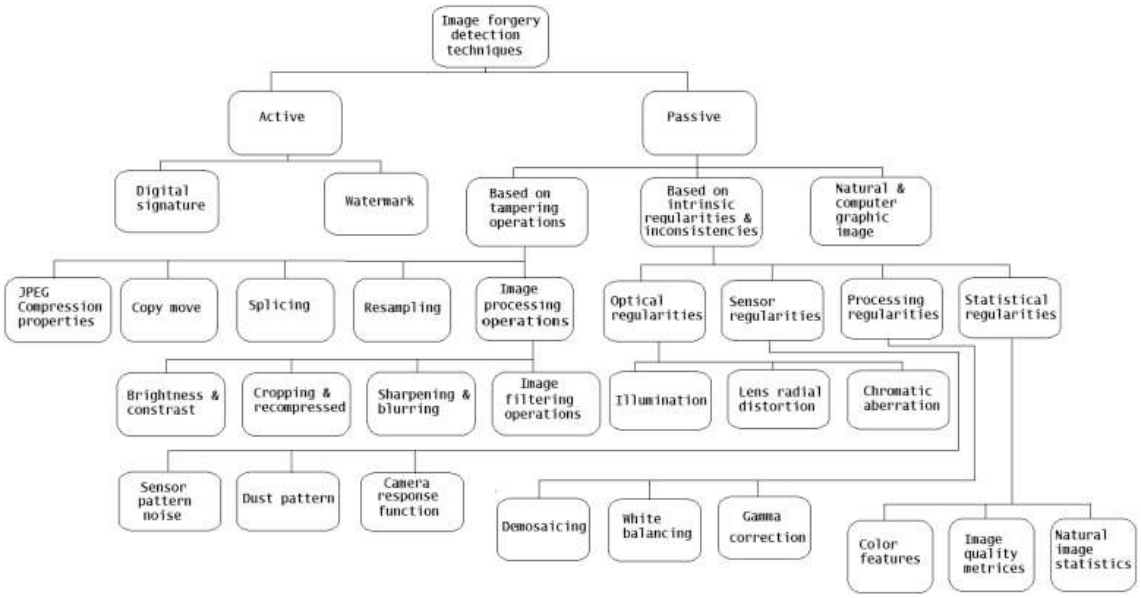


FIGURE 1.1 – Classification de Birajdar & Mankar (2013) des recherches menées en *Image Forensics*.

Les travaux sur les images de documents, images qui n'ont pas les mêmes caractéristiques de par leur parcours et leur utilisation, ne portent pas exactement sur les mêmes centres d'intérêt. En effet, le document est un type d'image particulier. Dans leur état de l'art sur la classification de document, Chen & Blostein (2007) utilisent trois catégories d'indices pour analyser le document : les indices de l'image, les indices de la structure (liens entre les différents éléments) et les indices textuels. Ces trois éléments ne sont pas des éléments que l'on peut retrouver dans une image de scène naturelle comme une photographie par exemple.

L'étude des liens entre les différents éléments d'un document correspond à ce que Chen & Blostein (2007) nomment l'analyse de la structure, qui est issue de l'analyse de l'agencement physique des éléments et de l'organisation logique. La structure physique du document correspond à la façon dont les contenus sont agencés les uns avec les autres et dans l'espace du document. Elle diffère de « l'image globale » par l'importance qu'elle accorde aux liens entre les éléments. Dans le cas de la fraude, si le faussaire change un élément sans modifier ceux qui lui sont liés sémantiquement, l'ensemble deviendra incohérent.

Les indices de l'image correspondent, toujours selon Chen & Blostein (2007), aux indices liés au traitement de l'image, c'est-à-dire à tout ce qui concerne les pixels, la reconnaissance des formes, la segmentation... La seule étude de l'image du document suffit parfois à le décrire et à retrouver le type de document auquel il appartient : les chèques par exemple, au-delà du support identique (papier, taille allongée), ont toujours le même agencement, quelle que soit la banque émettrice : une case à droite pour le montant en chiffres, suivie de deux lignes pour la date et le lieu, suivies d'un espace pour

la signature et de lignes à gauche pour le montant en lettres et l'ordre, les coordonnées de l'émetteur (« tireur ») et de sa banque (« tiré », dans une case labellisée « payable en France »), deux barres obliques... L'image dite globale du document de type chèque est donc facilement reconnaissable et l'absence d'un de ces éléments pourrait indiquer une fraude. Il s'agit donc de comparer les documents avec des modèles de documents. Le rapport de l'UNODC (2010) préconise donc l'utilisation de bases de données d'exemples de documents officiels, comme la base *Edison Travel Documents*[3].

Le troisième type d'indices qu'explorent Chen & Blostein (2007) pour la classification des documents concerne les indices textuels. Dans les travaux que les auteurs relèvent, ces indices se situent au niveau des mots (ou du lexique) avec ou sans OCR. Ce sont donc des images de mots quand il n'y a pas eu de reconnaissance de caractères. Nous nous démarquerons des auteurs en considérant que les images de mots avec des caractéristiques liées à l'image ne sont pas des indices textuels, et que les indices textuels peuvent être de formes plus variées que des simples listes de mots. Cependant, la distinction entre des images de mots ou de caractères et des images graphiques est pertinente dans la mesure où plusieurs travaux sur la détection des faux documents s'intéressent justement à la cohérence des caractères du document, comme nous le verrons dans la section 1.2.1.

1.1.2 Approches actives, ou protection par empreintes extrinsèques

Comme nous l'avons vu dans l'introduction générale, il est de plus en plus facile de créer des faux documents, soit par falsification en utilisant des logiciels de retouches d'image par exemple, soit par contrefaçon, en créant de toute pièce une fausse attestation ou un faux certificat médical avec un logiciel de traitement de texte. Des recherches sont donc menées pour lutter contre la fraude avant même que celle-ci n'ait lieu : il s'agit de protéger le document en y insérant des éléments de sécurité qui permettront d'authentifier plus sûrement et simplement le document. Ces approches se divisent en deux groupes : les approches des « données cachées » et celles des « signatures numériques ».

Les données cachées

La première approche pour sécuriser un document est de cacher des informations en son sein afin que le destinataire, averti, puisse les récupérer et vérifier l'authenticité du document.

Le filigrane est traditionnellement une technique utilisée dans la création de papier qui fait apparaitre un dessin ou un texte quand on regarde le papier d'une certaine façon, principalement par transparence. Le papier peut être plus fin, plus épais, ou formé de fibres différentes par exemple, suivant un motif qui identifiera le papetier ou qui permettra d'authentifier le document. Ainsi de nombreux billets de banques comportent des filigranes, qui sont très difficilement falsifiables (Cox et al. 2008). Bas et al. (2016) appellent « filigrane numérique » (*digital watermarking*) une insertion d'éléments qui

3. Base d'images de documents d'identité et de voyage mise à disposition par la Police centrale des Pays-Bas sur le site : `http://www.edisontd.net/`

doivent être invisibles, robustes (insensibles aux perturbations que pourrait subir le document) et sécurisés (difficiles à extraire sans une clé ou un algorithme) dans un média. Tkachenko (2015) ajoute à ces trois éléments attendus d'un filigrane deux autres éléments à prendre en compte : la capacité de stockage du filigrane (sa taille, par exemple) et sa complexité algorithmique. Certains filigranes peuvent toutefois être visibles (*overt*), par exemple pour mettre en évidence les crédits d'une photographie sur une image qui va circuler sur les réseaux sociaux (Dhiman & Singh 2016), quand d'autres peuvent être invisibles pour l'humain (*covert*) et nécessite donc un « détecteur », c'est-à-dire un logiciel (Cox et al. 2008).

Le filigrane se démarque de la stéganographie en ce qu'il ne cherche pas à faire passer un message, caché dans un média qui n'a rien à voir, mais plutôt qu'il cherche à modifier un contenu de façon imperceptible pour l'humain afin de porter des informations sur ce contenu. Néanmoins, comme la stéganographie et le filigrane utilisent les mêmes méthodes pour être embarqués dans un média et pour être détectés, ils sont souvent traités ensemble.

D'autres techniques se rapprochant du filigrane et de la stéganographie existent pour embarquer des éléments cachés de sécurité dans les documents, comme l'insertion pendant l'impression d'informations qui sont généralement la date et l'heure de la production du document (Mikkilineni et al. 2004)...

Zaidan et al. (2017) présentent un *benchmark* récent des méthodes de filigranes et de données cachées.

Signature numérique

Une autre façon de sécuriser le document, ou de permettre d'éviter les altérations entre la production du document et la réception finale, est de lui associer un *hash*, c'est-à-dire un code qui représente le contenu du document. Le document et le code associé sont alors transmis, ensemble ou séparément et le décryptage du code permet de vérifier l'authenticité du document.

Plusieurs fonctions de hachage existent, que ce soit pour les images ou pour le texte, résistant ou non au processus d'impression et de numérisation (Villán et al. 2007). Deux catégories de fonctions de hachages coexistent (Eskenazi 2017) : le hachage cryptographique et le hachage perceptuel ou flou. Dans le premier, un simple changement dans l'image ou dans le texte change complètement le code qui est calculé, alors que le deuxième permet d'être plus souple en n'affectant pas le code calculé pour quelques changements non perceptibles, ce qui est utile notamment pour authentifier des documents ayant subi des impressions et numérisations, des changements d'encodages ou de compression, des erreurs d'OCR, bref, « ayant une vie » de document hybride, c'est-à-dire une vie de document utilisé à la fois sous sa forme papier et sa forme numérique (Wu et al. 2009, Lei et al. 2011)...

Le code peut se présenter sous différentes formes (Tkachenko 2015) : chaîne de caractères, lien URL, code barre, code en deux dimensions (Quick Response code (QRcode), Data Matrix ou Aztec code par exemple), code en trois dimensions (rajout de la couleur) ou en quatre dimensions (ajout du temps) (Langlotz & Bimber 2007). Les codes en deux

dimensions présentent plusieurs avantages : ils peuvent contenir beaucoup plus d'informations que les chaines de caractères ou les codes barres classiques et ils sont beaucoup plus robustes à l'impression ou à la numérisation que les codes en couleur. L'enjeu des recherches sur les fonctions de hachage de documents est donc de rendre ces codes plus robustes aux transformations subies pendant le cycle de vie du document hybride, tout en le rendant le plus sécurisé possible : il ne doit pas pouvoir être recalculé par un éventuel fraudeur pour correspondre au nouveau contenu du document, par exemple.

Ces techniques de signature de document permettent l'authentification dans le cas d'un échange de ces documents entre des personnes averties et équipées des moyens de chiffrement et de déchiffrement de ces codes et/ou du code original, ce qui n'est pas le cas dans la plupart des échanges de documents.

1.1.3 Approches passives, ou détection d'empreintes intrinsèques

Contrairement aux approches dites actives, les approches passives de détection de fausses images peuvent s'appliquer à n'importe quel type d'images, sécurisées ou non, de scènes naturelles ou de documents, photographies ou dessins... La détection de faux documents parmi les originaux est une tâche de classification pour laquelle il s'agit de trouver les meilleurs indices, ou caractéristiques, de l'image et de sélectionner et entrainer le meilleur classifieur pour différencier les documents authentiques et les faux documents (Birajdar & Mankar 2013). Ainsi de très nombreux éléments constitutifs de l'image sont analysés dans la littérature et soumis à des tests de cohérence pour vérifier l'intégrité des images.

Sept manipulations de fraude sont identifiées par Poisel & Tjoa (2011) :

— le *Copy-Move Forgery (CMF)* : la duplication d'une partie de l'image dans la même image ;
— la retouche d'image ;
— le filtrage d'une partie non-désirée de l'image ;
— la suppression partielle d'un objet de l'image ;
— le *splicing* : la combinaison de plusieurs images ;
— la manipulation sur la luminance, la couleur et le contraste ;
— la manipulation sur la géométrie de l'image.

À ces catégories s'ajoutent et se superposent des éléments à analyser, qui peuvent être indices de modifications, répertoriés par Mahdian & Saic (2010) : les images de synthèse, la compression JPEG, les bruits locaux, les aberrations chromatiques, les flous et contrastes, l'identification de la source...

Si ces différents critères sont traités sous forme de liste dans Poisel & Tjoa (2011) et Mahdian & Saic (2010), ils apparaissent triés et classés dans Piva (2013) et Korus (2017). En effet, Piva (2013) dresse un panorama complet des travaux sur la détection d'images falsifiées en 3 parties relatives au cycle de la vie de l'image : empreintes basées sur l'acquisition de l'image (moment de la prise de vue), traces basées sur l'encodage (compression JPEG...) et indices basés sur l'édition de l'image. Korus (2017) reprend cette idée de classement des empreintes selon le cycle de la vie de l'image mais divise ce parcours en 9 étapes, comme le montre la figure 1.2.

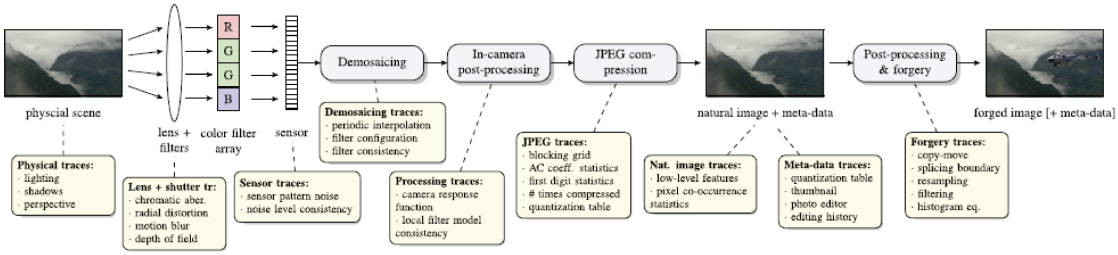


FIGURE 1.2 – Indices de détection de fausses images en fonction du cycle de vie de l'image (Korus 2017).

Nous ne nous intéresserons ici qu'à la détection de quelques manipulations qui nous semblent judicieuses dans le cas de modification d'images de documents. En effet, si le copier-déplacer, le mélange d'images et le re-dimensionnement ou la rotation d'une partie de l'image nous semblent être des éléments pertinents pour la détection de faux documents, les indices liés au contenu des images de scènes naturelles, comme la direction de la lumière et des ombres par exemple, ne le sont pas.

Détection des copier-déplacer ou duplication d'une région de l'image

L'une des techniques les plus répandues de falsification d'image est le copier-déplacer (*Copy-Move Forgery Detection (CMFD)*) au sein d'une image, c'est-à-dire la duplication d'une ou plusieurs parties de l'image. Cette technique est en effet très simple à mettre en œuvre avec des logiciels de dessin basiques : il suffit de sélectionner une zone, de la copier et de la coller ailleurs dans l'image, éventuellement en lui faisant subir quelques transformations (agrandissement, inclinaison, inversion...). Cette modification de l'image permet de doubler une information (par exemple, faire apparaître deux soleils au lieu d'un seul, ou un montant à deux caractères au lieu d'un seul), de cacher une information (faire disparaître le nuage d'un ciel en copiant des zones de bleu du reste de l'image, ou faire disparaitre un achat compromettant en copiant une zone de blanc du reste de l'image) ou bien de modifier une information en remplaçant une information par une autre (par exemple, coller le nuage par dessus le soleil, ce qui fera disparaitre le soleil, ou coller des chiffres sur d'autres, pour modifier une date).

Christlein et al. (2012) réalisent une évaluation des techniques de détection et de localisation des duplications dans les images en comparant 15 indices fréquemment utilisés et en proposant un nouveau corpus d'images pour l'évaluation. Nous pouvons citer, parmi les travaux qui ont marqué ce champ de recherche ceux de Fridrich et al. (2003), Popescu & Farid (2004), et plus récemment ceux de Kaur & Richa (2013), Amerini et al. (2014), Cozzolino et al. (2015).

Si ces travaux s'appliquent à des images de scènes naturelles, la détection des copier-coller peut également s'appliquer à des documents contenant du texte. En effet, il est aisé de copier un caractère pour le dupliquer à un endroit où il pourra, par exemple,

augmenter le prix d'une facture à rembourser. C'est ce que cherche à détecter Abramova (2016) sur des images de documents scannés.

La combinaison d'images

Le *splicing* consiste à combiner ou mélanger plusieurs images, c'est-à-dire à prendre un morceau d'une ou de plusieurs images pour l'insérer dans une autre. Cette technique n'est pas très différente du copier-déplacer vu précédemment mais est plus difficile à détecter étant donné que le morceau d'image n'apparait qu'une seule fois dans l'image modifiée et que l'image d'origine n'est pas toujours disponible pour comparer. Les méthodes de comparaison blocs par blocs ou par ensembles de points, avec des descripteurs comme SIFT (*Scale Invariant Feature Transform*) (Lowe 2004), ne peuvent donc pas fonctionner.

Ce sont donc plutôt des indices liés aux incohérences qui sont recherchés dans ce cas (Ng & Chang 2004), les images ayant probablement été capturées avec des appareils différents (appareil photo, smartphones, scanner...) (Fang et al. 2009) ou ayant subi des compressions (JPEG par exemple (Farid 2009)) et déformations diverses. Des travaux plus récents prennent en compte le flou ajouté aux contours des morceaux pour masquer les incohérences, comme ceux de Chen et al. (2017).

Le ré-échantillonnage

Le *resampling*, ou ré-échantillonnage, consiste à re-dimensionner une image ou à la tourner. Ces manipulations effectuées sur l'image peuvent être innocentes, mais les traces qu'elles laissent peuvent également être la marque d'une modification frauduleuse de l'image, notamment lorsque ces transformations sont détectées sur une partie de l'image seulement et non sur sa totalité. Les travaux sur la détection de ré-échantillonnage portent essentiellement sur l'analyse de la corrélation statistique entre les différents échantillons de l'image (Popescu & Farid 2005, Kirchner 2008, Mahdian & Stanislav 2008). La qualité des résultats de ces travaux diminue cependant fortement avec une forte compression JPEG, ou quand les images sont floutées pour dissimuler la fraude (Birajdar & Mankar 2013). Des travaux plus récents, prenant en compte ces problématiques, donnent de bons résultats (Bunk et al. 2017).

Mélange des indices et *deep learning*

L'arrivée du *deep learning* dans le domaine des *Image Forensics* n'a eu lieu que récemment, mais chamboule complètement les approches et les résultats. Il ne s'agit en effet plus seulement de détecter les fausses images selon un critère, mais de détecter et localiser toutes sortes de modifications de l'image, que ce soit la duplication, la superposition, le changement de taille ou toute autre altération... Plusieurs travaux récents montrent les performances de ces approches faisant intervenir des réseaux de neurones : Bayar & Stamm (2016), Rao & Ni (2016), Bappy et al. (2017), Rota et al. (2017), Bondi et al. (2017).

Si nous avons sélectionné des approches de traitement de l'image pouvant s'appliquer à des images de documents dans la section 1.1, nous nous intéressons dans la section 1.2 aux approches qui extraient des caractéristiques bien particulières aux documents qui nous intéressent.

1.2 La détection de faux documents (*Document Forensics*)

Comme nous l'avons vu, le domaine de l'investigation sur les images s'intéresse à tout type d'images. Cependant, les documents possèdent de nombreuses caractéristiques qui en font des images spéciales. En effet, les documents contiennent souvent peu de couleurs, peu de textures différentes, mais possèdent une structure, du texte, et parfois des images. Les documents sont également, de par leur cycle de vie, souvent confrontés à des impressions et des numérisations, qui apportent un certain bruit dans leur texture.

1.2.1 Détections des multiples impressions et numérisations (*Print&scan process*)

Plusieurs travaux dans le domaine des *document forensics* cherchent à déterminer l'origine du document, c'est-à-dire la façon dont il a été créé, avec quel type d'imprimante il a été imprimé, avec quel scanner il a été numérisé... En effet dans le cas d'un *splicing* (voir section 1.1.3), il est possible qu'un document contienne des morceaux qui n'ont pas été imprimés ou numérisés par le même appareil, ce qui rend suspect le document.

Lampert et al. (2006) classent automatiquement les caractères d'un document selon le type d'imprimante utilisé (imprimante laser ou imprimante jet d'encre) grâce à des *Support Vector Machines* (SVM). Les auteurs utilisent pour cela des indices liés aux contours des caractères qui sont beaucoup plus nets avec une imprimante laser qu'avec une imprimante jet d'encre, ce qui fait que, d'une part les contours sont plus rugueux (en haute résolution) avec des imprimantes jets d'encre, et d'autre part, plus de pixels en niveaux de gris apparaissent autour du caractère avec des imprimantes jet d'encre. Cette différence peut être mise en évidence en modifiant le seuil de binarisation afin de calculer la différence d'aire des caractères imprimés binarisés selon les seuils. Les auteurs calculent également un coefficient de corrélation et définissent un descripteur de texture basé sur trois transformations (filtre gaussien, filtre en ondelettes, et cartes binaires locales). D'autres travaux traitent de l'analyse de la texture des documents, comme ceux de Cruz et al. (2017), qui utilisent des modèles binaires locaux (*Local Binary Pattern* (LBP)) afin de détecter les régions qui contiennent des textures anormales au sein d'un document. Berenguel et al. (2017*b*) utilisent plusieurs descripteurs de texture, dont ils évaluent le temps et les performances sur un dataset de documents d'identité et de billets de banque. Les mêmes auteurs présentent également une nouvelle plateforme « *e-counterfeit* » qui permet de détecter des originaux et des copies de documents d'identité grâce à un smartphone (Berenguel et al. 2017*a*).

Schulze et al. (2009) cherchent également à différencier les impressions par une imprimante laser ou jet d'encre, mais veulent également distinguer les photocopies. En

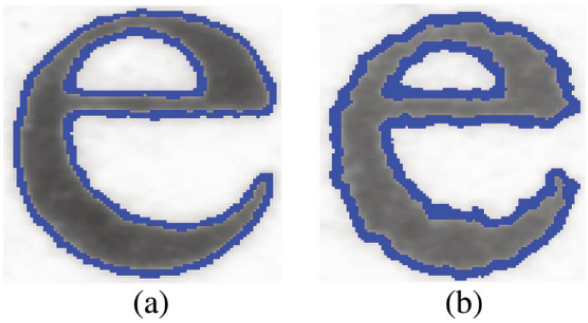


FIGURE 1.3 – Différence des contours (en bleu) entre une impression laser (a) et une impression jet d'encre (b) (Shang et al. 2014).

effet, le processus de numérisation induit une légère perte d'informations et un lissage des contours des caractères due à la diffusion de la lumière qui peut être perçue par la méthode qu'ils proposent. L'approche est cependant différente de Lampert et al. (2006) puisqu'elle cherche non pas à détecter plusieurs impressions différentes dans un document, mais à classer les documents entièrement imprimés par un seul appareil selon le type de cet appareil. Les indices utilisés ne concernent donc pas directement le caractère mais l'image entière. Les documents imprimés sont ensuite scannés avec différentes résolutions afin d'évaluer la performance de l'approche selon la qualité de la numérisation. Les travaux utilisent une approche assez similaire à celle de Tchan (2004), mais se distinguent de ceux-ci en ce qu'ils s'appliquent à des caractères, objets plus complexes que des simples carrés et ronds.

L'approche de Elkasrawi & Shafait (2014) consiste également à classer des documents textuels imprimés chacun par un seul appareil, mais, à la différence de Schulze et al. (2009), il ne s'agit pas de différencier le type d'imprimante mais bien l'imprimante directement. En effet, les imprimantes, qu'elles soient laser ou jet d'encre, n'émettent pas le même bruit et les auteurs réussissent à distinguer les 20 imprimantes utilisées à 77% d'*accuracy* (94% pour les seules imprimantes jet d'encre, plus faciles à différencier car produisant plus de bruit), en utilisant 15 indices liés au bruit dans la zone de texte du document.

Shang et al. (2014) cherchent également à distinguer les impressions lasers et jet d'encre et les photocopies, mais en se basant, comme Lampert et al. (2006), sur des indices liés aux caractères, qui sont : l'énergie du bruit dans la région du texte du caractère, l'énergie du bruit dans la région du contour du caractère, la rugosité du contour du caractère et le gradient moyen dans la région du contour du caractère. La figure 1.3, issue de Shang et al. (2014) illustre la différence entre les régions de contours (en bleu) d'un caractère `e` imprimé avec une imprimante laser et une imprimante jet d'encre.

Les documents textuels peuvent également être discriminés selon les scanners utilisés pour les numériser, comme dans les travaux de Khanna & Delp (2009), améliorés dans Khanna & Delp (2010), mais, à notre connaissance, il n'y a pas de travaux pour distinguer des parties d'un même document textuel provenant de différentes sources. Cela est sans

doute dû au fait que la plupart des indices utilisés pour détecter les différents scanners dans la littérature se base sur des régions non saturées des images, et donc s'applique bien à des photographies ou images de scènes naturelles, mais pas à des documents textuels.

1.2.2 Analyse des caractères et de la structure du document

Tout comme les travaux cités précédemment sur la détection des différentes impressions et numérisations, Bertrand et al. (2013, 2015) se concentrent sur les caractères pour détecter si un document est suspect ou pas. Dans (Bertrand et al. 2013), les auteurs cherchent à révéler deux types de fraudes : le copier-déplacer, en cherchant des caractères exactement identiques, et l'imitation, c'est-à-dire le fait de créer une zone de texte qui ressemble au texte du document, en cherchant des caractères qui seraient vraiment différents des autres en terme de forme, d'inclinaison ou d'alignement. En effet, il est très peu probable que deux caractères aient exactement la même forme dans un document ayant été imprimé puis scanné. Pour ce qui concerne l'imitation, les caractères modifiés par le fraudeur peuvent, dans la précipitation, être légèrement décalés par rapport aux autres, avoir une taille différente par rapport à une autre occurrence de la même lettre, ou bien être un peu penché par rapport aux autres caractères d'un même mot, dans le cas où il aurait été copié d'un autre document scanné (les documents scannés sont rarement parfaitement droits). Bertrand et al. (2015) s'intéressent également à la police des caractères du document : il s'agit de détecter des mots écrits avec des caractères de différentes polices. En effet, dans un document, si l'on peut trouver différentes polices pour les différentes parties du document, un même mot doit cependant être écrit d'une même et seule police.

Des travaux existent également, non plus à l'échelle des caractères, mais à l'échelle des lignes. van Beusekom et al. (2009) étudient leur inclinaison dans les zones de texte des documents pour détecter si des lignes ne sont pas bien alignées entre elles, ce qui rendrait le document suspect. Dans van Beusekom et al. (2010), les auteurs s'intéressent cette fois à l'alignement vertical des lignes, pour vérifier qu'aucune ligne du document ne dépasse dans les marges de droite et de gauche, ce qui pourrait être le cas si le fraudeur faisait un copier-coller d'un paragraphe, par exemple. Ces travaux s'intéressent donc à ce qu'on pourrait appeler la structure du document, c'est-à-dire l'organisation des éléments dans le document, leur alignement, leur positionnement. Ahmed & Shafait (2014) s'intéressent à cet aspect du document. Ils cherchent à déterminer si un document a subi des distorsions (notamment au cours d'une impression-numérisation), auquel cas il ne présenterait plus un alignement cohérent de sa structure et des différents blocs qui le composent avec un modèle donné.

1.2.3 Analyse des composantes graphiques des documents

Certains documents contiennent des images, censées prouver l'authenticité du document, comme des signatures, des tampons ou des logos. En ce sens, Micenková et al. (2015) considèrent que ce sont des éléments de sécurité extrinsèques, ajoutés au document. Cependant, les techniques utilisées pour vérifier si ces éléments sont réellement

originaux dans le document où s'ils ont été rajoutés ou modifiés *a posteriori* utilisent d'avantage des indices intrinsèques.

Ainsi, Micenková et al. (2015) cherchent à vérifier qu'un tampon est bien original et qu'il n'a pas déjà été imprimé et copié. Plusieurs indices permettent d'identifier un tampon original : l'absence d'une texture (contrairement au tampon scanné puis imprimé), la netteté et la rugosité des contours. Des travaux sont également menés sur la détection et la reconnaissance des logos dans les documents (Alaei & Delalandre 2014), et sur l'utilisation des logos dans la création d'une signature électronique (Eskenazi 2017), mais, à notre connaissance, aucun travail ne porte sur leur authentification. Les signatures manuscrites font aussi l'objet de nombreuses recherches, pour déterminer si un document a bien été signé par la bonne personne (Maergner et al. 2017, Okawa 2018).

Les techniques de traitement du signal et du traitement de l'image, ainsi que l'analyse plus fine des documents, présentées respectivement dans les sections 1.1 et 1.2, sont pertinentes dans la détection des faux documents et nous apportent de nombreux indices à exploiter. Nous pensons cependant que ces approches se concentrent essentiellement sur des fraudes de types falsification de documents. En effet, elles s'intéressent essentiellement à détecter des anomalies dues à la manipulation de l'image, à la combinaison de plusieurs images ou à la duplication de certains morceaux de l'image.

Nous constatons ainsi que les contrefaçons, c'est-à-dire les créations de documents contenant de fausses informations, imitant un modèle ou satisfaisant à une structure ordinaire, ne peuvent pas être détectées par ces approches. C'est pourquoi nous nous intéresserons maintenant, dans la section 1.3 à l'aspect sémantique du contenu des documents, en nous intéressant à ce qui se rapproche le plus de la vérification sémantique de contenus que nous proposons, à savoir la détection de fausses nouvelles et de rumeurs et la vérification automatique des faits, notamment dans les domaines journalistiques, scientifiques et sur les réseaux sociaux.

1.3 Détecter les fausses informations

Dans un contexte où le Sénat français a rejeté un projet de loi contre la circulation des *fake news*, craignant de porter atteinte à la liberté d'expression [4], il est toutefois important de remarquer que c'est un enjeu d'importance pour des démocraties dans lesquelles 54% des citoyens se sentent préoccupés par ce qui est réel et ce qui est faux dans les informations en ligne (62% en France) selon Newman et al. (2018). En effet, la manipulation de l'opinion publique et la propagande n'ont jamais été si difficiles à mettre en évidence que depuis qu'elles utilisent les nouveaux médias (les réseaux sociaux, les blogs, les micro-blogs, les sites d'information de qualités diverses...). Il est parfois très difficile de distinguer les vraies informations des fausses informations [5], ce qui peut

4. Article de LaCroix.fr du 27 juillet 2018 disponible sur : `https://www.la-croix.com/France/Politique/Le-Senat-rejette-propositions-loi-contre-fake-news-2018-07-27-1200958151`

5. Il suffit par exemple de regarder les commentaires sous certains articles du site parodique `http://www.legorafi.fr/` pour s'en apercevoir.

permettre de discréditer ou de promouvoir rapidement un candidat, un parti ou une idéologie.

Ce phénomène, qui n'est pourtant pas nouveau, mais qui était sans doute moins exacerbé quand l'information n'était diffusée que par les journalistes, est intéressant par sa complexité et est révélateur d'un nouveau monde, rendu possible par les nouvelles technologies, où tout un chacun peut être créateur de contenu, sans qu'il n'y ait de relecture ou de processus d'édition, et par conséquent, où la diffusion de l'information n'est plus verticale, mais horizontale.

Les *fake news*, ou « infox », contraction des mots « information » et « intoxication », comme le préconise maintenant la Commission d'enrichissement de la langue française [6], sont donc un nouveau terrain d'investigation pour la recherche. En effet, des travaux commencent à être publiés dans de nombreux champs disciplinaires, que ce soit en information et communication, en sociologie, en linguistique, en sciences politiques ou en économie, pour comprendre les enjeux de la diffusion de fausses informations, ses tenants et ses aboutissants, et son impact sur nos sociétés, comme nous le présenterons dans la section 1.3.1. Nous verrons dans la section 1.3.2 que les journalistes, professionnels de l'information, se sont emparé du phénomène et se proposent de n'être plus seulement des fournisseurs d'information, mais des vérificateurs d'information. Nous étudierons enfin dans la section 1.3.3 quelles réponses l'informatique peut apporter à cette problématique de détection des fausses informations...

1.3.1 Infox, désinformation et rumeurs

Le phénomène des infox (plus connues sous le terme *fake news*) est très intéressant à étudier d'un point de vue sociologique, politique et historique. En effet, l'émergence et l'importance du phénomène soulèvent de nombreuses questions dont les sciences sociales s'emparent : quels impacts les *fake news* ont-elles sur la démocratie ? Les fausses actualités et les rumeurs n'existaient-elles pas avant les réseaux sociaux et l'internet ? Que révèle la circulation des fausses informations sur les gens qui y croient, et les gens qui n'y croient pas ? Quels sont les milieux sociaux, les partis politiques, les aires géographiques qui partagent le plus de fausses nouvelles ? Quelles sont les motivations politiques, sociales et économiques des créateurs et des diffuseurs de rumeurs ? Quelles conséquences sur la crédibilité des médias et sur leur consommation ? La propagation des *fake news* est-elle un symptôme d'une transformation de la société, des processus de socialisation et de communication, et de la distribution des pouvoirs et des savoirs ?

Tucker et al. (2018) dressent une revue très complète des travaux en sciences politiques et sociales sur les relations entre les réseaux sociaux, la polarisation politique et la désinformation, terme qu'ils définissent comme étant un large spectre allant des *fake news* aux informations hyper-partisanes (c'est-à-dire diffusées par les sympathisants des partis d'extrêmes droite ou gauche) en passant par les rumeurs, les informations délibérément

6. Recommandation publiée dans le *Journal officiel* le 4 octobre 2018 : `https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000037460897&dateTexte=&categorieLien=id`, page consultée le 19 octobre 2018.

fausses, les informations erronées ou incomplètes ou les informations orientées politiquement. La revue est divisée en six parties qui concernent les conversations politiques en ligne, les conséquences de l'exposition à la désinformation et à la propagande en ligne, les producteurs de la désinformation, les stratégies et les tactiques de la diffusion de la désinformation sur les plateformes, les contenus du web et la polarisation politique et enfin l'impact de la « mésinformation » et de la polarisation sur la démocratie américaine.

Sur ces dernier termes, Born & Edgington (2017) font la différence entre la désinformation, la mésinformation et la propagande. Pour les auteurs, la désinformation est l'information qui est intentionnellement fausse ou incomplète diffusée dans l'objectif de diminuer la confiance du public. La « mésinformation » est une information inexacte ou incomplète qui est diffusée de façon non-intentionnelle. La propagande est une information qui peut être vraie ou fausse mais qui a pour but de présenter les points de vue divergents sous un mauvais jour et de nuire à la compréhension et à la diversité politique.

Allcott & Gentzkow (2017) cherchent à dresser le paysage politique de la circulation des *fake news* et à évaluer l'impact de la circulation de ces articles sur les résultats de l'élection présidentielle américaine de 2016. Selon les auteurs, les créateurs de fausses informations ont plusieurs motivations : l'argent (créer un site satirique comme Le Gorafi [7] peut permettre d'avoir une grande audience, à en juger par les 1 325 000 abonnés à la page Facebook du site et le nombre de commentaires et partages de chaque article [8]) ou l'idéologie (on peut par exemple chercher à faire circuler des fausses informations pour favoriser un candidat, comme le site endingthefed.com en faveur de Donald Trump [9]).

D'un point de vue éthique, Borden & Tew (2007) rappellent les fondements du journalisme et différencient ainsi les journalistes et les commentateurs, qui eux, peuvent s'affranchir de l'éthique journalistique qui veut que chaque information publiée soit vérifiée, tout en gardant les codes, notamment linguistiques, du journalisme.

1.3.2 Lutte contre les fausses informations et *fact checking*

Selon 75% des personnes interrogées dans l'étude de Newman et al. (2018), ce sont les journalistes et les entreprises de presse qui devraient avoir la plus grande responsabilité dans la séparation entre le vrai et le faux, suivis des entreprises technologiques (Google, Facebook...) qui sont le relais des médias en ligne (71%). En ce qui concerne l'intervention gouvernementale sur la régulation, la disparité des réponses selon les pays et selon la maîtrise des actualités est plus forte : en moyenne 61% des personnes interrogées en France pensent que les gouvernements devraient améliorer la séparation entre les fausses actualités et les vraies, contre 73% en Corée du Sud, et 41% aux États-Unis, où l'intervention du gouvernement serait une atteinte forte à la liberté d'expression et au premier

7. `http://www.legorafi.fr/`

8. On peut par exemple citer « Toulouse : il se fait abattre de 46 balles dans le corps pour avoir demandé un « pain au chocolat » », article commenté 2272 fois sur le site et partagé plus de 3000 fois sur Facebook depuis sa publication le 20 mars 2013.

9. Le site `endingthefed.com` est responsable de 4 des 10 *fake news* ayant eu le plus de succès pendant les trois mois avant l'élection présidentielle américaine. Ce site a été créé par un supporteur de Donald Trump pour partager des nouvelles en sa faveur, selon l'interview de Tess Townsend : `https://www.inc.com/tess-townsend/ending-fed-trump-facebook.html`

amendement. De même, les personnes avec une forte connaissance des actualités ne sont que 51% à penser que les gouvernements devraient intervenir.

Dans ce sens, l'ancienne ministre des droits des femmes, Laurence Rossignol, a fait adopter une loi en février 2017 contre la désinformation et l'entrave à l'interruption légale de grossesse [10], pénalisant les sites se faisant passer pour des sites d'information sur l'IVG et proposant des numéros gratuits pour accompagner les femmes enceintes, mais dont le véritable but était de les dissuader d'avorter, en leur faisant peur ou en les culpabilisant [11].

Dans un autre contexte, les hébergeurs de contenus et les responsables de réseaux sociaux cherchent à prévenir, ou à guérir, leur site de la désinformation. Ainsi, la plateforme américaine Reddit [12] a banni en avril 2018 près d'un millier de comptes suspectés d'avoir participé activement à la propagande russe visant à dénigrer Hilary Clinton au profit de Donald Trump pendant la campagne présidentielle américaine de 2016, notamment en faisant circuler des fausses informations [13]. Il aura donc fallu deux ans à la plateforme pour identifier et révéler des comptes suspects.

En France, il existe des plateformes pour que tout un chacun puisse vérifier les informations qui circulent sur le Web et dans la presse. Ces plateformes sont principalement proposées par des journaux en ligne. Par exemple, le Décodex est un outil mis en place par LeMonde.fr, permettant de vérifier si une source est plutôt fiable ou non, et qui existe sous la forme d'une page Internet, de plugins pour les navigateurs Firefox ou Chrome et de bot Messenger Facebook [14]. De même, le journal Libération.fr a mis en ligne deux rubriques sur son site web : Desintox [15], lancé en 2008, qui a pour vocation de mettre en lumière tout ce que les journalistes de Libération ont pu trouver de faux dans la presse (« bobards, intox, exagérations, etc ») [16], et Checknews [17], lancé en 2017, qui a pour ambition de répondre à toutes les questions factuelles que peuvent poser les internautes sur l'actualité après une réelle investigation journalistique.

A l'international, les médias s'organisent et coopèrent pour vérifier les informations qui circulent et mettre à jour des réseaux d'influence, et contrer l'opacité des systèmes politiques et financiers. L'*International Fact-Checking Network* (IFCN) est un réseau d'organismes de presse qui s'engagent à suivre les éléments d'une charte sur les nouvelles qu'ils diffusent et qu'ils vérifient [18]. Cette charte a actuellement cinquante signataires,

10. Articles L2223-1 et L2223-2 du Code de la santé publique.

11. Ces sites, à l'instar de `www.ivg.net`, étaient très bien classés par les moteurs de recherche, augmentant ainsi la confiance dans les informations de ces sites.

12. `https://www.reddit.com/`

13. Article du Monde.fr du 11 avril 2018 disponible sur : `https://www.lemonde.fr/pixels/article/2018/04/11/reddit-a-supprime-994-comptes-suspectes-de-propagande-pour-le-compte-de-la-russie_5284114_4408996.html`

14. `https://www.lemonde.fr/verification/`.

15. `http://www.liberation.fr/desintox,99721`

16. Citation extraite de l'explication de la différence entre les deux outils de *Fact Checking* de Libération.fr, disponible sur : `http://www.liberation.fr/checknews/2017/11/22/bonjour-quelle-est-la-difference-entre-desintox-et-checknews_1652789`

17. `http://www.liberation.fr/checknews,100893`

18. Charte et informations disponibles sur : `https://ifcncodeofprinciples.poynter.org/`

dont cinq organismes français. Un autre exemple de coopération internationale journalistique est l'*International Consortium of Investigative Journalists* (ICIJ) qui a notamment enquêté pendant des mois sur les *Panama Papers* afin d'en extraire et diffuser des informations vérifiées.

Brandtzaeg et al. (2015) étudient les pratiques des journalistes vis-à-vis des réseaux sociaux et constatent que ces derniers sont la première source d'informations. Cependant l'étude révèle que les journalistes ont parfois du mal à distinguer les fausses informations, et notamment les fausses images. Les besoins en matière d'outils – simples mais puissants – pour gérer ces données semblent importants au vu des entretiens présentés.

Justement dans l'optique d'aider les journalistes à vérifier et à traiter le flux continu d'informations, plusieurs projets et outils se mettent en place. Ainsi le projet de recherche européen Social Sensor, présenté dans (Schifferes et al. 2014), a pour objectif de proposer un outil de détection des actualités sur les réseaux sociaux, de repérage des tendances et de vérification des informations. De même, l'outil FactMinder (Goasdoué et al. 2013) propose une solution semi-automatique pour assister l'utilisateur dans sa démarche de remise en contexte de l'information et d'estimation de sa véracité. Un plug-in pour navigateur existe également pour aider les utilisateurs à vérifier les vidéos et est notamment utilisé par les journalistes de l'Agence France Presse (AFP) (Teyssou et al. 2017).

Chen et al. (2015) argumentent également la nécessité d'avoir des outils automatisés de détection de fausses informations sur les réseaux sociaux afin d'aider, non seulement les journalistes à faire leur travail, mais également les consommateurs, et plus particulièrement, à travers leurs éducateurs, les jeunes internautes, pour leur permettre d'acquérir un esprit critique et leur apprendre à vérifier leurs sources. Des programmes institutionnels existent d'ailleurs partout dans le monde pour faire de la prévention sur ces questions auprès des jeunes (entre autres), comme le *Better Internet for Kids* de la Commission Européenne [19], connu en France comme le *Safer Internet Program* [20].

1.3.3 Détection automatique des *fake news*

Il est important d'aider les journalistes dans leur travail de vérification des informations, mais il est également intéressant de créer des outils capables de faire la différence de façon entièrement automatisée entre de faux articles et vrais articles, entre des informations suspectes et des informations authentiques.

Rubin et al. (2015) dressent des typologies des *fake news*, des corpus que les chercheurs utilisent ou aimeraient utiliser pour les détecter et des critères attendus d'un bon corpus. Plusieurs corpus existent et permettent de comparer les méthodes, parmi lesquels Zubiaga et al. (2015) pour les rumeurs, ou Wang (2017), corpus annoté de *fake news*. Les auteurs séparent ainsi les fausses nouvelles en trois classes : les « fabrications sérieuses », c'est-à-dire les articles écrits par de vrais journalistes dans des vrais journaux ou tabloïds qui font état de faits faux ou non vérifiés, les « canulars à grande échelle », qui sont des fausses informations ou des rumeurs reprises par les médias traditionnels, et les « faux

19. `https://ec.europa.eu/digital-single-market/en/policies/better-internet-kids`
20. `http://www.internetsanscrainte.fr/le-projet/safer-internet-program`

humoristiques » (satire, parodie...). Ces trois types de fausses actualités requièrent selon les auteurs, des types de traitements différents, comme ils l'expliquent dans un autre article (Conroy et al. 2015).

Pour Shu et al. (2017) en revanche, les *fake news* ne sont que les articles qui sont intentionnellement faux et vérifiables. Cela permet de se concentrer sur les techniques de détection de fausses informations que les auteurs classent en deux parties : les méthodes de détection basées sur le contenu (analyse linguistique des messages, indices visuels), et celles basées sur le contexte (analyse du parcours de l'information, de sa source, des utilisateurs qui la partagent, de sa propagation sur les réseaux, etc.). Nous pouvons cependant constater que la plupart des approches de détection s'appuie à la fois sur des indices linguistiques et sur des indices contextuels.

L'une des approches de détection des fausses informations sur les réseaux sociaux s'appuie sur l'analyse des utilisateurs qui les propagent, même si la propagation en cascade n'est pas suffisamment discriminante (Conti et al. 2017). La propagation peut se mesurer et s'analyser à l'aide des « likes » sur le réseau social Facebook, qui peuvent être un très bon révélateur de *hoax*, comme le montrent Tacchini et al. (2017). L'analyse temporelle de la diffusion d'un message sur un réseau peut également être un bon indice pour détecter une rumeur, surtout quand elle est combinée à l'analyse du texte (Lukasik et al. 2015, Ma et al. 2015).

L'authentification de l'auteur de l'information est un bon moyen pour authentifier ou suspecter une information. Rajapaksha et al. (2017) cherchent par exemple à retrouver l'origine des informations diffusées sur Twitter par l'analyse du contenu du tweet, tout en ajoutant une dimension temporelle issue des données du tweet afin de prendre en compte la variabilité du style des utilisateurs. L'analyse linguistique qu'ils utilisent est basée sur un indice simple mais efficace, très utilisé dans la stylométrie (Stamatatos 2009) : la fréquence des n-grams de mots et de caractères, c'est-à-dire le nombre de fois que des suites de n mots (ou caractères) apparaissent dans un texte. Ahmed et al. (2017), par exemple, utilisent la fréquence des n-gram des mots (*Term Frequency (TF) et Term Frequency-Inverse Document Frequency (TF-IDF)*) et comparent les résultats obtenus avec six classifieurs différents, obtenant la meilleure distinction entre vrais et faux avec un *Linear Support Vector Machine* sur les unigrammes.

La stylométrie est également utile, non pour retrouver l'auteur d'un message, mais pour faire la différence entre des vraies informations et des rumeurs ou des fausses informations. Il s'agit d'utiliser des critères linguistiques pour faire de la classification des articles/messages/avis/documents à partir de leur contenu. Potthast et al. (2017) proposent ainsi une approche basée sur le style pour différencier des articles hyper-partisans (de gauche ou de droite) d'articles neutres et parviennent à distinguer des articles satiriques de vrais articles ; cependant leur approche ne parvient pas à détecter de fausses informations. Feng et al. (2012) utilisent des critères syntaxiques (*Part-of-Speech tags (POS), Probabilistic Context Free Grammar (PCFG)*) pour différencier des faux avis sur des hôtels parus sur des sites de réservation en ligne de vrais avis, avec de très bons résultats.

De nombreux indices linguistiques sont testés pour catégoriser des articles entre satire,

faux et vrais par Horne & Adali (2017). Les auteurs ont en effet comparé trois sortes d'indices linguistiques sur les titres et le contenu des articles des corpus utilisés afin de relever les indices les plus pertinents. Ces types d'indices sont :

— les indices de style : nombres de mots par phrase, nombre de chaque nature de mots (POS), nombre de chaque pronom, temps des verbes...
— les indices psychologiques : nombre de mots positifs, négatifs, appartenant au champ sémantique de l'analyse, du pouvoir, du risque, de la causalité...
— les indices de complexité : profondeur des arbres syntaxiques, profondeur des arbres des groupes nominaux ou verbaux, indices de lisibilité des textes, longueurs des mots...

Les résultats de cette étude montrent que les articles de *fake news* sont plus courts, utilisent moins de ponctuation, de citations, et de mots techniques ou analytiques, mais beaucoup plus de pronoms personnels. Les titres sont également très différents entre les articles de *fake news* et les articles réels, en ce qu'ils contiennent beaucoup plus de noms propres afin de créer des associations mentales entre des entités et des affirmations. Pour détecter les articles satiriques, Rubin et al. (2016) définissent quatre vecteurs que sont l'humour, l'absurdité, la grammaire et la combinaison de la ponctuation et des termes négatifs à partir de différents indices linguistiques, lexicaux ou syntaxiques.

Un autre type d'approche est proposé par Misra et al. (2008) : la vérification de la cohérence interne d'un document textuel. Cette approche part de l'hypothèse qu'un article, ou n'importe quel texte, contient plus de sujets différents lorsqu'il est généré automatiquement, par fusion de contenus extraits du web par exemple. Elle s'attache donc à détecter les sujets des documents, ce qui permet, sur le corpus utilisé, d'obtenir des résultats intéressants.

Le contenu des articles n'étant pas seulement textuel, des travaux se sont également intéressés à la vérification des images utilisées dans les articles. Outre les travaux d'analyse de détection d'images falsifiées vus précédemment, nous pouvons citer les travaux de Elkasrawi et al. (2016), qui cherchent à détecter si les images contenues dans les articles de presse ont été reprises d'un autre article ou d'ailleurs sur le Web, ou si l'image est originale. Pour cela, les auteurs utilisent le moteur de recherche inversé de Google Image [21] pour retrouver des images similaires, qu'elles aient été reprises telles quelles ou qu'elles aient été modifiées. Le travail de Amerini et al. (2017), s'il ne s'inscrit pas directement dans la détection de fausses informations, peut tout de même y participer : il s'agit d'identifier par l'analyse de l'image si elle a été téléchargée d'un réseau social ou si elle provient directement d'un appareil photographique.

Comme Elkasrawi et al. (2016), Goasdoué et al. (2013) utilisent des connaissances externes, issues de bases de données et de connaissances pour comparer les informations dans leur outil semi-automatique FactMinder. Enfin, le travail de Magdy & Wanas (2010) est très intéressant pour notre approche dans la mesure où il permet d'acquérir des connaissances externes à partir des informations de documents. En effet, après avoir extrait des « faits » (ou informations) des textes qu'ils étudient, les auteurs cherchent à les vérifier un par un en les transformant en requêtes qu'ils envoient sur un moteur

21. `https://www.google.fr/imghp`

de recherche. Ils analysent ensuite le contenu des pages des dix premiers résultats du moteur de recherche Bing [22] afin d'en extraire les faits importants. Ils comparent ensuite les faits du document et les faits extraits du web pour en sortir une valeur support. Une valeur unique pour le document est ensuite calculée en pondérant l'addition des valeurs de chaque fait. Si ces premiers résultats ne sont pas très satisfaisants, cette approche peut certainement être améliorée avec les pistes que donnent les auteurs. Cette approche cependant ne peut pas s'appliquer sur nos documents dans la mesure où les « faits » dont parlent les auteurs proviennent d'un texte en langage naturel et sont extraits après un *POS tagging* permettant de repérer les syntagmes de type nom-*to*-nom ou *to* est une chaîne de caractère (souvent un verbe) qui porte une relation sémantique. Les faits sont hiérarchisés en fonction de leur poids sémantique calculé par rapport à leur fréquence, leur ordre d'apparition dans le texte et leur appartenance à une relation sémantique définie dans WordNet. Les auteurs créent enfin pour chaque fait les plus importants du document un lot de quatre requêtes constituées respectivement du premier nom seul, du deuxième nom seul, des deux noms, et du fait complet. Si l'approche globale ressemble à ce que nous imaginons pour obtenir des connaissances externes, le type de nos documents ne nous permet pas d'obtenir des informations aussi idéales pour la génération de requête. Nous reviendrons sur cette approche dans le chapitre 5.

1.4 Conclusion

Les travaux présentés dans cette section montrent la diversité des études pour distinguer le vrai du faux, que ce soit dans le domaine de la protection des documents, de la détection des fausses images ou dans celle des fausses informations. Cependant, si toutes ces approches répondent à une problématique générale commune – la quête de la vérité et de l'authenticité –, elles ne peuvent que partiellement répondre à notre problématique de détection des faux documents ordinaires, altérés ou contrefaits. Les techniques de détection présentées dans la section 1.1 et 1.2 ne s'intéressent qu'aux images qui ont subi des modifications, et celles présentées dans la section 1.3 s'appliquent essentiellement à des documents contenant du texte en langage naturel, avec des phrases et une grammaire permettant d'utiliser les méthodes de traitement automatique des langues.

Nous proposons donc dans cette thèse de contribuer à la détection de documents quasi-semi-structurés contenant de fausses informations, par la vérification de chacun des ses éléments textuels. Cette approche permettra à la fois de répondre aux lacunes des approches liées au traitement de l'image, qui ne peuvent pas détecter des documents contrefaits ou falsifiés puis réimprimés ou recompressés et à la fois de proposer une méthode de vérification des informations totalement automatisée, généralisable à tout type de document. La méthode que nous allons proposer est donc innovante de tout point de vue, traitant les documents sous un angle sémantique, et analysant du texte qui n'est pas du langage naturel.

22. `https://www.bing.com/`

Chapitre 2

Création d'un corpus d'images et de textes de documents

Sommaire

La première difficulté d'un travail sur la détection de faux documents est d'obtenir des faux documents réels. En effet, afin de pouvoir construire un processus capable de distinguer des faux documents, il faut pouvoir en analyser des exemples, afin de comprendre les mécanismes de fraudes, les méthodes des fraudeurs ou les informations fréquemment modifiées. Nous pensons effectivement qu'il est plus réaliste de construire un modèle basé sur des observations que de créer un outil à partir d'hypothèses et le confronter ensuite à la réalité. Cette approche, inspirée de la linguistique de corpus, nécessite l'utilisation d'un corpus représentatif de documents fraudés.

Nous verrons dans la section 2.1 que les différents corpus utilisés pour la détection de faux documents ne peuvent pas répondre à notre problématique. En effet, les corpus existant sont généralement des corpus synthétiques, générés à partir d'informations aléatoires, convenant parfaitement à des besoins spécifiques au traitement de l'image, mais ne pouvant répondre à des attentes de réalisme pour une analyse sémantique. Nous avons donc construit un nouveau corpus de documents, répondant à de nombreuses contraintes scientifiques. Nous avons ainsi collecté de nombreux documents physiques authentiques que nous avons ensuite numérisés afin d'obtenir un corpus d'images de documents (section 2.2). Nous avons ensuite extrait le texte de ces images grâce à un logiciel de reconnaissance de caractères, puis corrigé automatiquement et manuellement ces textes afin d'obtenir une transcription la plus fidèle possible de chaque document (section 2.3). Nous avons enfin fait modifier ces documents authentiques, images et textes, afin d'obtenir différents types de fraudes, que ce soit en termes de manipulation d'image ou de changement d'information (section 2.4).

2.1 Corpus de faux documents existant

Les différents travaux présentés dans la section 1.2 utilisent des corpus de faux documents appropriés à leurs approches basées sur l'analyse des images de ces documents. Ces corpus sont souvent construits pour une tâche bien précise, afin de valider une approche particulière de détection de faux documents, ce qui en fait des corpus très spécifiques. C'est le cas par exemple du corpus créé par Bertrand et al. (2013) pour détecter les caractères falsifiés dans les documents : après avoir généré des documents automatiquement, les auteurs ont déplacé des caractères, changé leur taille ou les ont inclinés légèrement. Ils ont également copié certains caractères pour les replacer sur d'autres et ont rajouté du bruit de Kanungo et al. (1993) pour simuler le bruit d'une impression-numérisation. Les datasets sont composés de 20 000 caractères numériques d'une même police de caractères dont 5% ont été fraudés de manière aléatoire, pour chaque type de fraude. Maîtriser les conditions de création des documents et des fraudes est un réel avantage pour évaluer son approche et éliminer toutes sortes d'éléments imprévus ou de difficultés liées à la complexité de la réalité. Cependant, le fait que ces corpus ne peuvent justement pas être considérés comme représentatifs de la réalité est un inconvénient pour notre approche dans la mesure où ils se limitent à établir les conditions idéales pour un traitement bas-niveau, c'est-à-dire se concentrant sur les pixels formant les caractères : il n'y a qu'une seule police utilisée, des caractères uniquement numériques, des types de fraudes hypo-

thétiques et un taux de fraude fixé à 5% pour tous les jeux de documents. Ces critères, s'ils sont utiles, voire nécessaires, pour une analyse pixellaire, ne nous conviennent pas pour les approches que nous souhaitons mettre en place car ils simplifient la nature du document et de la fraude et ne représentent pas des cas réalistes de fraude.

Dans leurs travaux suivants, Bertrand et al. (2015) utilisent un corpus de documents contenant du texte aléatoire (*lorem ipsum*) généré avec des polices et des tailles de caractères différentes. La fraude, qui consiste en des copier-coller ou des imitations de caractères dans des mots, est appliquée aléatoirement sur 1% des mots de chaque document. Là encore, le corpus est construit en fonction de son application première qui est la détection des anomalies dans la police ou la taille des caractères au sein du mot, c'est-à-dire avec des approches se concentrant sur l'image, mais n'est pas utilisable pour une approche globale, utilisant de la sémantique entre autres, de détection de faux documents.

van Beusekom et al. (2009, 2010) utilisent également des documents générés pour leur application propre. Il s'agit de documents contenant le texte du traité proposé pour construire une Constitution Européenne et modifiés afin de répondre à la problématique de recherche sur l'orientation des lignes et le dépassement dans les marges de certaines lignes du texte. Si les documents contiennent cette fois du texte « réel », le cas d'usage de fraude de ces documents reste flou : pourquoi quelqu'un voudrait-il frauder un texte de loi qui, par ailleurs, était accessible en ligne ? Est-ce qu'il est fréquent de coller une phrase, ou un morceau de phrase, par dessus un texte ? Cette pratique parait complexe, *a fortiori* dans un document textuel en langage naturel, où toute modification apportée au texte doit se soumettre à une contrainte syntaxique due à la grammaire de la langue.

Les mêmes auteurs ont pu, par la suite, appliquer leurs approches sur un large corpus de 143 000 factures réelles scannées (van Beusekom et al. 2015). Cependant, s'ils savaient que certaines de ces factures étaient fausses, ils n'avaient pas la vérité terrain associée à ces factures et n'ont donc pas pu évaluer leur approche quantitativement. Afin d'évaluer au mieux leurs approches, l'un des auteurs, qui n'est pas un expert en faux document, a donc regardé lui-même les résultats afin de déterminer, grâce aux informations sémantiques contenues dans le document et en comparant avec les autres documents, la nature frauduleuse ou non des documents suspects. Les auteurs mettent en avant dans leur conclusion l'importance d'utiliser des données réelles, et non « ce à quoi les informaticiens pensent que les données réelles et les faux documents pourraient ressembler ». Par ailleurs, les factures qu'ils utilisent leur sont fournies par des clients et contiennent donc des données sensibles, ce qui ne leur permet pas de partager le dataset. Ceci est un véritable frein aux recherches dans le domaine des *document forensics*, car la construction artificielle de corpus de faux documents ne peut certainement pas refléter ni les proportions réelles de la fraude, ni la créativité des fraudeurs.

Plus récemment, un nouveau corpus de documents a été proposé pour la tâche de détection de fausses images de documents par Sidère et al. (2017). Il s'agit de bulletins de salaire créés artificiellement en complétant par des informations aléatoires les champs requis par les formats classiques de ce type de document. La génération automatique des informations a été motivée par le fait que les données personnelles contenues dans les

bulletins de salaire sont très sensibles (en France du moins) et ne peuvent donc pas être publiées. La structure de ces documents n'étant que peu impactée par les fraudes, elle est totalement identique d'un document à l'autre et seul le contenu change. Ce corpus étant construit pour des approches de détection et de localisation des fraudes basées sur l'image, ce n'est pas un problème d'avoir des informations aléatoires, telles que des personnes imaginaires, des codes postaux qui ne correspondent pas aux villes ou des salaires qui ne sont pas cohérents avec le poste de l'employé. Les documents sont générés en utilisant 5 polices et 4 tailles de caractères différentes. Les documents sont ensuite falsifiés par plusieurs non-experts à qui il est demandé de modifier l'image de document en copiant-collant des caractères au sein de la même image (CMF) ou d'une autre image (*splicing*), ou en créant du texte ressemblant à celui déjà présent dans l'image. La vérité terrain est ensuite créée automatiquement par une différence entre les images falsifiées et les images authentiques afin de retrouver les zones altérées. L'utilisation du logiciel de reconnaissance de caractères TesseractOCR permet de retrouver la localisation des caractères modifiés. Cette approche originale, faisant intervenir de nombreux fraudeurs différents, a permis de créer un premier corpus de faux documents réalistes. Cependant, *réaliste* ne veut pas dire *réel*, et la méthode de création de ces documents, avec une structure identique et un contenu textuel aléatoire, simplifie artificiellement le traitement de l'information et limite les approches possibles.

Ces corpus ont pour point commun d'être constitués en vue d'un traitement spécifique basé sur l'image. Nous cherchons dans notre thèse à détecter les faux documents par les informations textuelles qu'ils contiennent, afin de renforcer ces approches par de nouveaux indices. Outre le fait que ces corpus ne proposent pas de transcriptions textuelles, les informations inscrites dans les documents sont souvent générées aléatoirement, combinant des prénoms et des noms d'employés fictifs, insérant des prix au hasard... Ces informations ne peuvent donc pas être considérées comme vraies par défaut, et la vraisemblance des informations ne pourrait être évaluée. Cela rejoint le discours de Rastier (2005), qui dans un tout autre domaine, explique qu'un corpus ne peut être construit et utilisé que pour une gamme d'applications précises. S'il parle de corpus de textes, en linguistique de corpus, l'idée qu'un corpus ne peut pas être parfaitement représentatif de l'ensemble du problème, voire de la réalité, reste la même. Ainsi, l'auteur dit :

> « Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications.
>
> « Tout corpus suppose en effet une préconception des applications, fussent-elles simplement documentaires, en vue desquelles il est rassemblé : elle détermine le choix des textes, mais aussi leur mode de « nettoyage », leur codage, leur étiquetage ; enfin, la structuration même du corpus. Allons plus loin, un corpus doit « être aimé » : s'il ne correspond pas à un besoin voire un désir intellectuel ou scientifique, il se périme et devient obsolète.

> « Cette dépendance à l'égard d'une application ou d'une gamme d'applications permet de dédramatiser les problèmes récurrents de la représentativité et de l'homogénéité. Aucun corpus ne représente la langue : ni la langue fonctionnelle qui fait l'objet de la description linguistique, ni la langue historique, qui comprend l'ensemble des documents disponibles dans une langue. En revanche, un corpus est adéquat ou non à une tâche en fonction de laquelle on peut déterminer les critères de sa représentativité et de son homogénéité. La linguistique de corpus peut ainsi être objective, mais non objectiviste, puisque tout corpus dépend étroitement du point de vue qui a présidé à sa constitution. »

Nous avons donc choisi de construire notre propre corpus, constitué de documents authentiques contenant des informations réelles, qui puisse servir aux méthodes traditionnelles basées sur le traitement de l'image, mais surtout qui est adapté à notre approche innovante de vérification des informations. En effet, construire un corpus unique pour ces deux types d'approches permet d'offrir à la communauté *document forensics* un *benchmark* idéal pour évaluer et comparer les performances de leurs approches.

2.2 Du papier à l'image

Constituer un corpus de faux documents réalistes est une tâche complexe et chronophage. Il s'agit tout d'abord de restreindre le choix dans l'étendue des documents possibles répondant aux contours dessinés dans l'introduction générale. En effet, de très nombreux types de documents existent, échangés entre les entreprises, les administrations et les particuliers, contenant toutes sortes d'informations intéressantes et falsifiables. Cependant, si l'objectif de notre thèse est de proposer une approche générale applicable à tous ces types de documents, nous avons fait le choix de nous restreindre à un seul cas d'étude : le ticket de caisse.

Ce type de document a été choisi pour de nombreuses raisons :

— Les tickets de caisse sont, la plupart du temps, anonymes, et ne contiennent pas d'informations personnelles. L'anonymat des documents est une contrainte importante pour la publication des jeux de données, d'autant plus si l'on évoque le fait que ces documents seront amenés à être fraudés. Obtenir des documents déjà anonymes permet d'éviter de les anonymer nous-mêmes, ce qui serait très long et fastidieux d'une part, et ce qui provoquerait une modification de l'image et une perte d'information dommageable pour les systèmes de détection de falsifications.

— Les tickets de caisse ne comportent pas, ou rarement, de données sensibles ou à caractère confidentiel sur les entreprises qui les émettent. Toutes les informations inscrites concernant l'entreprise sont des informations publiques, comme les coordonnées, le numéro de SIRET, ou le numéro de TVA.

— Les tickets de caisse sont des documents très courants en France : en magasin, sur les marchés ou dans les restaurants, tout achat peut être l'occasion d'obtenir un ticket de caisse. Cela permet d'obtenir beaucoup de documents en peu de temps.

La collecte aurait été beaucoup plus longue si nous avions décidé de nous intéresser aux bulletins de salaire (mensuels) ou aux avis d'imposition (annuels).

— Les tickets de caisse présentent une diversité étonnante, tant dans la forme (ou l'image) que dans le fond (ou le texte), ce qui rend l'analyse riche et ne la limite pas à un cas simpliste.

— Les tickets de caisse ne sont pas toujours utiles pour les particuliers, ce qui leur permet de s'en débarrasser aisément, et par conséquent, ce qui nous permet d'en collecter facilement. En effet, les tickets de caisse permettent de prouver un achat en cas de contestation ou de demande de remboursement. Ainsi, pour de nombreuses dépenses de la vie quotidienne, comme la salade du déjeuner ou le paquet de gâteaux du goûter, les particuliers n'ont pas besoin de ces justificatifs et ne souhaitent pas les garder sur le long terme.

— Les tickets de caisse restent des documents qui peuvent servir de preuves et peuvent donc être falsifiés pour des raisons économiques ou judiciaires. Un ticket de péage ou de taxi peut par exemple être réclamé pour prouver un alibi d'un suspect. Ce sont alors les informations de dates et de lieux qui pourraient être sujettes à modification. On peut également imaginer qu'un·e employé·e en mission pour son entreprise veuille se faire rembourser un peu plus que ce qu'il ou elle n'a réellement dépensé et augmente ainsi légèrement la note du restaurant...

Toutes ces raisons ont fait des tickets de caisse des documents de choix pour la constitution d'un corpus riche et consistant pour la détection des faux documents, aussi bien pour des approches basées sur l'image que pour des approches basées sur le contenu.

2.2.1 Collecte de tickets de caisse

Afin de collecter de nombreux tickets de caisse venant de magasins différents, nous avons demandé à nos collègues du laboratoire de garder les tickets de leurs achats et de les déposer dans une boîte prévue à cet effet dans la salle de repos du laboratoire. Tout un chacun pouvait donc déposer les tickets qui encombrent les porte-feuilles de manière anonyme et à tout moment. Nous avons également demandé à nos amis et notre famille de nous fournir les tickets de caisse dont ils voulaient se débarrasser.

Inciter les gens à fournir des documents est une tâche de longue haleine : il s'agit de le leur rappeler régulièrement et de faire en sorte qu'ils s'en souviennent, à la fois au moment de prendre le ticket à la caisse du magasin (beaucoup de personnes refusent les tickets, ou les prennent et les chiffonnent immédiatement pour les jeter un peu plus loin), et à la fois quand ils ont la possibilité de les déposer dans la boîte ou de me les donner directement (plusieurs collègues ont gardé des tickets chez eux ou dans leur bureau pendant des mois avant de penser à me les rapporter). Nous avons donc, entre décembre 2016 et juin 2017, envoyé sept mails à l'ensemble du laboratoire incitant à la consommation en diverses occasions (Noël, soldes d'hiver, Saint Valentin, ménage de printemps, Ramadan...) afin de marquer les esprits. Il s'agit également de faire preuve de pédagogie afin d'expliquer la différence entre un ticket de caisse et un reçu de carte bancaire, souvent fournis en même temps par le commerçant, et parfois imprimés à la suite l'un de l'autre. Le reçu de carte bancaire (parfois appelé « facturette ») contient des informations sur le paiement effectué

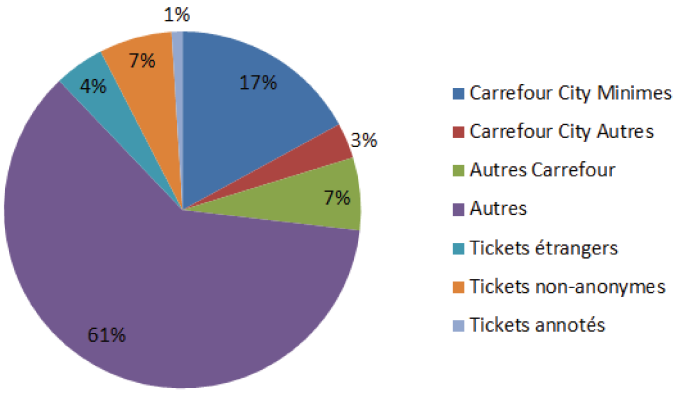


FIGURE 2.1 – Proportion des tickets collectés selon leur type ou leur provenance.

alors que le ticket de caisse porte sur les produits ou les services achetés. Nous ne nous sommes intéressée qu'aux tickets de caisse, les reçus de cartes bancaires ne contenant pas suffisamment d'informations vérifiables.

Nous avons ainsi pu récolter environ 2250 tickets en six mois, provenant de nombreux magasins, restaurants, bars, stands de marché, péages, restaurants d'entreprise, etc. essentiellement de la région rochelaise, mais également de partout en France et dans le monde. La pause déjeuner des enseignants-chercheurs n'étant pas toujours très longue, une part non négligeable (environ 17%) des tickets récoltés proviennent d'un Carrefour City situé à proximité du laboratoire, et concerne l'achat de sandwichs, de salades, de plats préparés, de fruits et de petits gâteaux. Le graphique de la figure 2.1 montre comment nous avons trié les tickets, séparant ceux provenant du Carrefour City des Minimes, d'autres Carrefour City, des autres types d'enseignes Carrefour (market, express, contact, bio, hypermarché Carrefour), et les « autres », regroupant tous les autres tickets.

Ce graphique met également en évidence la proportion de tickets non-anonymes, souvent issus de magasins de vêtements ou de chaussures soucieux de personnaliser leurs tickets de caisse quand les client·e·s possèdent une carte de fidélité. Cette technique commerciale permet d'améliorer la relation client en offrant un service « sur-mesure ». Nous avons écarté ces tickets du corpus afin de préserver l'anonymat de nos contributeurs. Les tickets « annotés » du graphique correspondent aux tickets contenant des inscriptions manuscrites, telles que des calculs, des croix face à certains produits, des prix entourés... Les tickets servent en effet pour certaines personnes à faire leurs comptes, et elles n'hésitent pas à griffonner ces documents pour mieux les comprendre. Nous avons également retiré ces tickets afin de ne pas dégrader les résultats de l'OCR ou de l'extraction d'information.

Les tickets collectés ont des tailles très variables selon leur provenance : là où les tickets du marché sont étroits et contiennent un minimum d'informations (comme sur la figure 2.2), les tickets de certaines enseignes de grande distribution peuvent atteindre les 30 ou 40 centimètres de long pour un seul produit acheté et contiennent des publicités, des

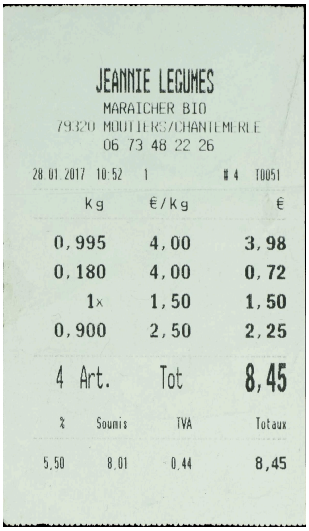

JEANNIE LEGUMES
MARAICHER BIO
79320 MOUTIERS/CHANTEMERLE
06 73 48 22 26
28.01.2017 10:52 1 # 4 T0051
Kg €/Kg €
0,995 4,00 3,98
0,180 4,00 0,72
1× 1,50 1,50
0,900 2,50 2,25
4 Art. Tot 8,45
% Soumis TVA Totaux
5,50 8,01 0,44 8,45

FIGURE 2.2 – Ticket de caisse de marché, sans noms de produits.

informations sur les promotions temporaires, sur les cartes de fidélité ou sur le magasin (comme dans la figure 2.3)... Certains tickets, issus de courses familiales par exemple, comportent de très nombreux produits, et sont alors très longs, pouvant atteindre 80 cm, ce qui pose des problèmes pour la numérisation.

Enfin, nous avons choisi de ne pas traiter les tickets étrangers afin de simplifier le traitement de l'information du ticket. En effet, notre approche n'est pas multilingue et nos algorithmes, en l'état, ne peuvent s'appliquer qu'à des documents français. Cependant, nous gardons soigneusement ces tickets étrangers, écrits dans divers systèmes d'écriture (japonnais, arabe, alphabet cyrillique...), car il serait très intéressant de les prendre en compte dans une prochaine approche.

2.2.2 Numérisation

Nous avons numérisé 1969 tickets de caisse. Pour obtenir les meilleures images possibles, nous avons pris les tickets en photo avec un appareil photographique numérique relié à un ordinateur et à un logiciel de capture d'image grâce à son dos numérique (Leaf Credo Digital Back sur Mamiya 645DF+ System, Schneider). L'appareil était fixé verticalement à une colonne graduée sur laquelle il pouvait monter ou descendre, tandis que les tickets étaient posés sur une table située sous des projecteurs dans une pièce sans autre source de lumière. Les tickets sont donc photographiés du dessus, face à l'appareil, évitant ainsi d'avoir à gérer des angles de profondeur des documents dans un espace en trois dimensions.

Les tickets de caisse sont généralement imprimés sur un papier très fin, qui peut se froisser très vite. Cela est d'autant plus vrai qu'ils sont souvent manipulés avec peu de soin, rangés rapidement dans des poches ou des porte-feuilles sans grande précaution. Nous avons donc essayé de défroisser et déplier au maximum les tickets avant de les

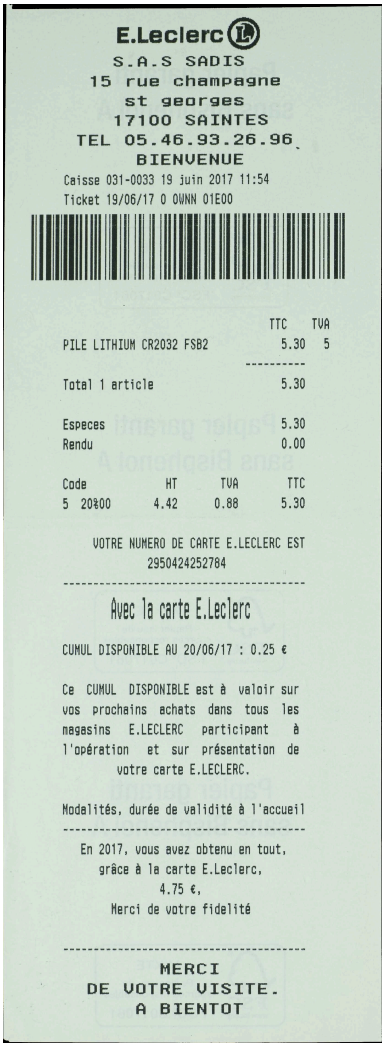


FIGURE 2.3 – Ticket long de grande surface pour un seul produit acheté.

installer sous une plaque de verre pour être aplatis pendant la prise de vue. Cette installation permet de vérifier que les tickets ne glissent pas quand on referme la plaque de verre sur eux, contrairement au scanner traditionnel où on ne peut pas voir ce qui va être réellement numérisé. Les projecteurs ont été réglés de façon à ne pas avoir de reflets sur la vitre. Chaque cliché contenait plusieurs tickets de caisse afin de gagner du temps. Pour bien pouvoir les séparer par la suite, nous avons fait en sorte de les disposer sur un fond coloré et de les espacer de quelques millimètres, créant ainsi des contours faciles à distinguer. La figure 2.4 montre cette installation.

Les plus grands tickets de caisse, dépassant les 45 centimètres, n'ont pas pu être numérisés, d'une part parce que la vitre n'était pas assez grande, d'autre part parce que cela posait des problèmes pour la mise au point, qui se faisait en mode automatique. Le

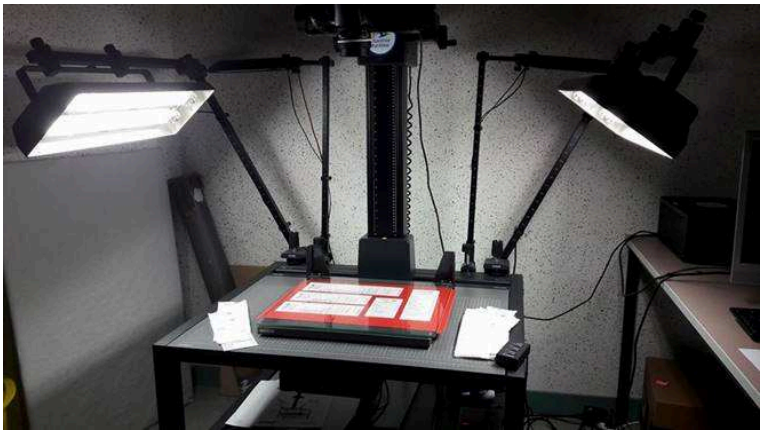


FIGURE 2.4 – Installation pour la numérisation des tickets de caisse.

ticket le plus long que nous ayons collecté mesure 76 centimètres. Seules des solutions de captures mobiles vidéo pourrait restituer des images de qualité suffisante pour le traitement que nous souhaitons effectuer. Bien que des travaux soient en cours sur ce domaine au laboratoire (Chazalon et al. 2017, Burie et al. 2015), nous avons préféré laisser ces documents de côté.

Cette méthode de numérisation avait pour objectif principal de faciliter la reconnaissance de caractères, en optimisant l'éclairage et en aplatissant les tickets afin de les rendre les plus lisibles possible, le traitement de l'image n'étant qu'une étape nécessaire dans la construction du corpus, mais n'étant pas notre tâche principale, focalisée sur l'aspect sémantique et textuel des tickets de caisse. Ce n'est que dans un deuxième temps que nous avons numérisé environ 2000 tickets de caisse supplémentaires selon trois procédures différentes afin de refléter des cas d'usages plus « réels ». En plus de numériser les tickets avec l'installation précédente, nous les avons photographiés, non aplatis, avec un appareil photo classique (Canon EOS 200D) et un smartphone Sony XPeria Z3 compact sous Android. Cette numérisation ayant été réalisée tardivement, ces tickets n'ont pas été traités dans notre thèse. Cependant, cette base d'images de documents constitue une perspective intéressante pour la combinaison d'indices graphiques et sémantiques de détection des fraudes.

2.2.3 Extraction des tickets

Les clichés obtenus contiennent entre 3 et 8 tickets de caisse, horizontaux ou verticaux, qu'il faut séparer et redresser. Nous nous sommes inspirée pour cela d'un algorithme de détection de cases de bande-dessinée, présenté dans Rigaud et al. (2013). En effet, la disposition des tickets de caisse, dont un exemple est visible sur la figure 2.5, est analogue à celle de cases bien formées de bande-dessinée. Il s'agit d'extraire des composantes connexes d'une certaine taille, plus petite que la totalité de l'image, mais plus grande que des zones de texte. La détection des contours se réalise assez simplement après

FIGURE 2.5 – Exemple de cliché pris contenant huit tickets à extraire et redresser.

FIGURE 2.6 – Un ticket horizontal incliné avec un masque noir.

binarisation, grâce à la bibliothèque openCV et la fonction `findContours`. Le rectangle englobant les coordonnées du contour est ensuite extrait avec la fonction `boundingRect`. Cette fonction permet d'obtenir un rectangle vertical contenant tous les points d'un ensemble. Ici, l'ensemble de points correspond aux images des tickets, qui ne sont pas toujours placés « droits » sur l'image, et ne sont donc pas verticaux ni parallèles entre eux. Ils sont parfois penchés ou horizontaux. Les zones autour des tickets qui sont comprises dans le rectangle englobant sont dans un premier temps recouvertes d'un masque noir, comme dans la figure 2.6. Une image intermédiaire est créée pour chaque ticket à partir de l'image non-binarisée, avec le masque noir autour des contours, afin de contrôler le processus.

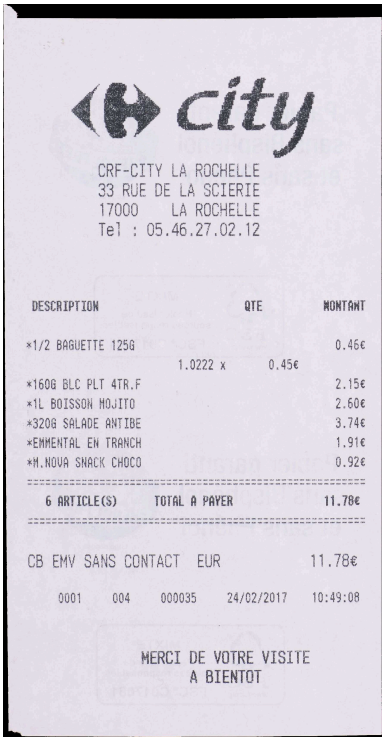


FIGURE 2.7 – Ticket pivoté et redressé.

Nous avons ensuite calculé l'angle d'inclinaison du ticket dans le rectangle afin de le redresser. Pour cela, nous avons utilisé la fonction `minAreaRect` qui retourne, entre autres, l'angle d'inclinaison d'une aire rectangulaire d'un objet détecté dans l'image. Le ticket est donc encadré au plus près, et son angle d'inclinaison par rapport à l'image est calculé, ainsi que le centre de l'image. On annule alors l'angle de rotation de l'aire rectangulaire avec la fonction `getRotationMatrix2D` en l'alignant verticalement. On redécoupe ensuite l'image afin de supprimer au maximum les bordures noires. On pivote ensuite manuellement les tickets de caisse, car il est difficile de trouver automatiquement dans quel sens le ticket est orienté. En effet, aucun indice ne permet de définir l'orientation du document, à part le texte, qui est lui-même difficilement reconnu par les logiciels de reconnaissance de caractères lorsqu'il est incliné. La figure 2.7 montre le ticket de la figure 2.6 après la réduction de l'inclinaison.

Bui et al. (2017) proposent plusieurs méthodes de pré-traitement pour améliorer la reconnaissance de caractères sur des tickets de caisse numérisés à la volée, c'est-à-dire avec un smartphone. Ces documents, à la différence des nôtres, ont des luminosités très variables, contiennent beaucoup de bruits car ils ne sont pas aplatis, et sont déformés ou flous, ce qui explique la nécessité d'appliquer un ensemble de pré-traitement impliquant de la binarisation, de la réduction de bruits et de l'augmentation des contrastes. Nous estimons que notre procédure d'acquisition des images est telle que ces pré-traitements ne sont pas nécessaires. Nos images sont donc prêtes pour l'extraction du texte.

2.3 De l'image au texte

Lorsqu'on évoque le mot *texte* dans des discussions au sein de la communauté « documents », les scientifiques du domaine pensent aussitôt à l'image du texte, aux pixels qui forment des caractères. Travailler sur du texte revient, pour eux, à travailler sur de la segmentation d'images de documents, sur de la détection de texte dans l'image, ou sur de la reconnaissance de caractères (OCR). Dans la communauté du traitement automatique des langues, le *texte* évoque du discours, de la sémantique et une syntaxe liée à une langue naturelle. Dans notre travail en revanche, le *texte* sur lequel nous travaillons ne correspond ni à une image de texte, ni à un texte en langage naturel, mais plutôt à une suite de caractères qui forme des informations plus ou moins compréhensibles pour l'humain. Afin de pouvoir traiter ces informations, il nous faut en amont extraire le texte de l'image, c'est-à-dire transformer des pixels en chaîne de caractères exploitables.

2.3.1 OCR

Pour extraire le texte des images, nous avons utilisé le moteur de reconnaissance optique de caractères d'ABBYY FineReader Engine 11 Sample (c) 2013 ABBYY Production LLC. Cet outil, très performant sur de nombreux types de documents et dans de nombreux systèmes d'écriture, permet de choisir des options pour améliorer la reconnaissance de caractères [1]. Nous avons sélectionné plusieurs de ces paramètres qui nous paraissaient intéressants à explorer :

- enableAggressiveTextExtraction (aeate) : détecte tout le texte dans l'image y compris le texte de mauvaise qualité ;
- detectTextOnPictures (adtop) : détecte le texte inclus dans les images ;
- dontDetectTables (adt) : les tableaux ne sont pas détectés ;
- singleLinePerCell (tslpc) : reconnait les tableaux avec une ligne de texte par cellule ;
- fastObjectsExtraction (aftda) : détecte tout le texte de l'image, y compris dans les images ;
- permitModelAnalysis (apma) : analyse différentes mises en page et sélectionne la meilleure variante, ce qui peut améliorer la qualité de la reconnaissance ;
- TextTypes Receipt et Matrix (rtt) : le type de texte correspond au type de texte labellisé « ticket de caisse » ou « matrice » ;
- altoDontWriteNondeskewedCoordinates (adwndc) : les coordonnées des caractères, mots et blocs sont celles de l'image utilisée pour la reconnaissance, qui peut donc avoir été pivotée.

Nous avons testé les différentes combinaisons de ces options sur plusieurs tickets afin de trouver celle qui reconnait le mieux le texte particulier des tickets de caisse. En effet, le texte des tickets est disposé dans l'image de façon non-linéaire comme pourrait l'être du texte en langage naturel. Comme nous pouvons le voir sur les figures 2.2, 2.3 ou 2.7, les éléments textuels sont souvent séparés par de grands espaces blancs, notamment entre

1. Toutes les options sont disponibles dans la documentation en ligne sur `https://www.ocr4linux.com/en:kb:documentation:start`, consultée le 11 septembre 2018.

les produits et les prix, ou par des lignes horizontales, ce qui nous a fait tester différentes configurations concernant la mise en page. De même, les éléments sont souvent disposés sous forme de tableaux et les logos des magasins, qui contiennent souvent du texte sont considérés par défaut comme des images.

La commande utilisée sur l'ensemble des tickets est finalement : `finereader -if 0.tif -rl french -adt -rtt Receipt -tet UTF8 -f ALTO -adwndc -of 0.xml`. Elle permet de prendre une image au format TIF, de spécifier la langue (français), le type de texte (ticket de caisse), de ne pas détecter les tableaux et d'afficher les coordonnées de l'image redressée dans le fichier XML ALTO de sortie.

Nous extrayons ensuite le texte du fichier XML de sortie OCR grâce à un script qui rétablit les lignes complètes de texte dans le bon ordre grâce aux coordonnées verticales et horizontales des mots. En effet, la disposition particulière du texte fait que la segmentation de l'image sépare des fragments de textes alignés mais séparés par un grand espace. Nous avons donc estimé que les éléments à moins de 20 pixels de différence verticalement étaient alignés. Pour information, dans nos images, un caractère fait en général entre 50 et 70 pixels de haut.

Nous obtenons ainsi un texte brut où tout espacement entre groupe de caractères d'une même ligne est réduit à une espace typographique [2] et dans lequel chaque ligne correspond à une ligne du texte. Nous faisons ainsi le choix de perdre l'information spatiale du document, qui pourrait cependant être utile pour l'extraction d'information. Néanmoins, cette opération s'avère nécessaire pour obtenir un texte exploitable et il aurait été complexe et lourd de signifier l'emplacement des éléments dans le texte.

2.3.2 Correction automatique

Les résultats de l'OCR ne sont pas parfaits, et le texte en l'état n'est pas exploitable tant le nombre d'erreurs est, pour certains tickets, élevé. Cela est certainement dû au manque de contraste de l'encre des tickets de caisse, souvent effacée du fait de leur cycle de vie, voire mal appliquée sur le document par des imprimantes fatiguées. Les pliures et les taches peuvent également affecter la segmentation et la reconnaissance. Certaines erreurs de reconnaissance de caractères sur les tickets de caisse proviennent également du lexique utilisé : les tickets de caisse contiennent de nombreuses abréviations, or la plupart des OCR utilisent des dictionnaires pour améliorer leur performance. Seulement, les abréviations présentes dans les tickets de caisse n'existent pas dans le dictionnaire, ce qui crée des erreurs de reconnaissance. Pour certains tickets cependant, le nombre d'erreurs reste difficile à expliquer, comme c'est le cas pour la sortie présentée dans la figure 2.8, issue de l'image de la figure 2.7.

Nous pouvons observer dans ce texte plusieurs erreurs :

— les signes de la première ligne qui remplacent le logo « city » du magasin,

— les signes € de fin de lignes qui sont transformés en `C` et en `f`,

2. En typographie, « l'espace » entre deux mots est féminine. Cela vient de la « petite lame de métal qu'on emploie lors de l'impression pour séparer les mots », selon le Trésor de la Langue Française informatisé.

```
1  *•;.
2  CRF-CITY LA ROCHELLE
3  33 RUE DE LA SCIERIE
4  LA ROCHELLE
5  05.46.27.02.12
6  DESCRIPTION QTE MONTANT
7  *1/2 BAGUETTE 125G 0.46C
8  1.0222 x 0.45€
9  *1608 BLC PLT 4TR.F 2.15€
10 4L BOISSON MOJITO 2,60€
11 *3208 SALADE ANTIBE 3.74€
12 ^EMMENTAL EN TRANCH 1.91€
13 *M.NOVA SNACK CHOCO 0.92f
14 •*# w,; a# p **# «K. WW MW
15 U#.' *w ** «n ww m+m m* *** n: tw #*v «*• -■ W L*' w «W ı
16 6 ARTICLE(S) TOTAL A PAYER 11.78€
17 om- u*t ** *& •au- cm# «a# #w* *n* w mw kü* f" w* «lai w
   #*a wi #t* **i n# /wa #* ma 'U, au, w &w# w *a*
18 CB EMV SANS CONTACT EUR 11.78€
19 000035 24/02/2017 10:49:08
20 MERCI DE VOTRE VISIT E
21 A BIENTOT
```

FIGURE 2.8 – Texte issu de l'OCR sans correction.

— les `G` signifiant « grammes » qui sont interprétés comme des chiffres (ici, deux fois `8`), étant inscrits dans la continuité de trois caractères numériques,
— une espace ajoutée à l'avant-dernière ligne,
— des chiffres oubliés dans les lignes 4 et 19,
— les étoiles de début de ligne transformées en accents circonflexes ou non reconnues,
— des points reconnus en virgules,
— et enfin, le plus visible, les lignes horizontales encadrant le total qui ont été reconnues comme des lignes de texte.

Ces erreurs, notamment la non-reconnaissance de certaines parties du document et la sur-reconnaissance de lignes graphiques en texte, n'apparaissent heureusement pas dans tous les tickets. Les autres erreurs cependant sont des erreurs fréquentes que nous avons cherché à corriger automatiquement.

Plusieurs travaux existent sur la correction post-OCR, utilisant des lexiques (Thompson et al. 2015), des statistiques sur les erreurs fréquentes (Afli et al. 2016), des méthodes utilisée en traduction automatique (Mokhtar et al. 2018), des statistiques sur les probabilités des mots selon le contexte (Mei et al. 2016), (Kissos & Dershowitz 2016), des grammaires locales (Sagot & Gábor 2014) ou encore des mesures de similarités basées sur le contexte et des distances d'édition (Jean-Caurant et al. 2017). Reynaert (2014) fait le constat qu'il est difficile de trouver des corpus avec leur vérité terrain, et de comparer

Tableau 2.1 – Moyennes du nombre de caractères et de corrections automatiques par types de tickets

	Lignes/ticket	Corrections/ticket	Pourcentage caractères corrigés
Carrefour City	16.8	6.7	1.70
Autres Carrefour	31.7	13.3	1.71
Autres tickets	30	6.6	0.96

les algorithmes utilisés dans les différents papiers. Chiron et al. (2017) ont répondu à ce problème en organisant une compétition qui avait pour tâches de détecter les erreurs dans un corpus de 12 millions de caractères OCRisés et de les corriger.

Si toutes ces techniques peuvent s'appliquer à des documents en langage naturel ou avec beaucoup de données, elles ne sont pas applicables dans notre cas. Nous avons donc décidé d'utiliser une méthode à base de règles permettant de modifier plusieurs éléments à l'aide d'expressions régulières.

Nous avons constaté que le symbole € se transformait en de nombreux caractères dans notre corpus : `4, 5, 6, C, c, e, E, F, f, K, s, S, T, (, # et *`. Quand ces caractères interviennent donc à la fin d'une ligne, suivant une séquence comprenant un chiffre, une virgule ou un point puis deux chiffres, ce qui correspond à un montant, ils sont donc remplacés par le symbole €.

De même, lorsque le quatrième caractère d'une séquence de quatre chiffres est un 0, un 6, un 8 ou un 9, et que cette séquence se situe sur une ligne se finissant par un prix (avec le symbole €), nous considérons qu'il s'agit d'une erreur de reconnaissance, et que ce quatrième caractère doit être transformé en `G`. Il est très peu probable que cette modification ne soit pas correcte, étant donné que la plupart des produits a un poids qui ne dépasse pas le kilogramme, et que rares sont les noms de produits contenant quatre chiffres à la suite (sauf les bières 1664, le chocolat Poulain 1848, les bouteilles de vin associées à une année et les piles électriques Energizer).

Les autres modifications que nous avons effectuées s'adressent en priorité aux tickets provenant de Carrefour city. Comme ils ont tous la même structure, il est en effet plus facile de corriger les erreurs fréquentes sans risquer de rajouter plus d'erreurs que l'on n'en corrige dans ce sous-corpus que dans les documents venant de divers magasins. Nous avons donc également rajouté des règles pour modifier les virgules des prix et des numéros de téléphone en points, les caractères non-alpha-numériques de début de ligne de produit en étoile et les mots « QTE », « ARTICLE(S) » et « PAYER » qui subissent de nombreuses altérations, notamment sur les lettres Q (O,0,8,G,D), R et Y (V, T). Ces erreurs sont typiquement dues à l'utilisation d'un OCR, qui rend plus probable la reconnaissance de « OTE » (du verbe « ôter ») que « QTE ».

Le tableau 2.1 montre quelques statistiques sur les corrections effectuées de manière automatique. Nous pouvons constater plusieurs choses : d'une part les tickets provenant de Carrefour city sont généralement très courts, comme nous l'expliquerons dans le chapitre 3, d'autre part le taux de correction est très faible pour les tickets provenant de

l'ensemble des magasins de la chaîne Carrefour, et encore plus faible pour les autres tickets, ce qui s'explique par le fait que ces derniers présentent des mises en page, du lexique et des présentations textuelles parfois très différents de ceux pour lesquels notre correction automatique supervisée a été construite. Par exemple, si certains tickets utilisent des symboles € après le montant des produits en fin de ligne, d'autres les utilisent avant le montant, d'autres encore placent les montant sur la ligne en dessous du nom du produit, d'autres enfin utilisent la lettre E, les lettres EUR, EURO ou EUROS, des chiffres pour renvoyer à différents taux de TVA ou encore aucun caractère du tout autour du montant. Il est donc très difficile de créer des expressions régulières dans ce contexte si irrégulier, tout comme il est quasiment impossible d'utiliser un dictionnaire à cause des abréviations, ou d'utiliser des méthodes syntaxiques à cause de l'absence de grammaire.

2.3.3 Correction manuelle participative

La correction automatique n'étant pas une solution pertinente pour notre corpus, nous avons décidé de corriger manuellement nos textes de tickets de caisse. Nous nous sommes vite rendu compte cependant que nos 1971 tickets représentaient 53166 lignes, et plus d'un million de caractères (1232353 pour être précis), ce qui ne pouvait être corrigé par une seule personne dans le temps imparti. Nous avons donc créé une interface web de correction manuelle collaborative de résultats OCR afin que la tâche de correction soit partagée par un grand nombre de correcteurs motivés.

Construire une plateforme de correction participative présente quelques contraintes : l'interface doit être simple d'utilisation, les consignes doivent être claires et les participants doivent être stimulés. Nous avons donc mis en place un simple site Wordpress hébergé par l'université de La Rochelle, qui permet à tout un chacun d'accéder à la page de corrections [3]. Cette page présente aléatoirement une image parmi les tickets non corrigés, et son texte correspondant pré-imprimé dans une zone de texte modifiable en vis-à-vis, comme illustré sur la figure 2.9.

Une autre zone de texte permet au participant d'entrer un nom, ou un pseudonyme. Cette identification est très importante car elle permet aux participant·e·s de s'impliquer dans la démarche, dans le sens où, dès lors qu'un nom, ou même un pseudonyme, est inscrit, il a une existence et est donc porté par un individu. Le ou la participant·e n'est plus anonyme, quand bien même il, ou en l'occurrence, elle, a choisi de se nommer par les noms de tous les personnages d'un dessin animé. L'identification permet également d'afficher un message personnalisé après la soumission, comptant le nombre de tickets corrigés par la même personne au cours de la session. Le ou la correcteur·rice peut alors choisir de corriger un autre ticket, ce qui le ramène sur la page de correction avec un nouveau ticket et son nom pré-enregistré, ou bien il peut aller consulter le classement des correcteurs·rices.

Ce classement, également disponible depuis le menu, permet de voir tou·te·s les participant·e·s classé·e·s par ordre décroissant de nombre de tickets corrigés. Nous avons ainsi

3. La page est toujours accessible et en service sur : `http://receipts.univ-lr.fr/correct-ocr-results/`, consultée le 15/09/2018.

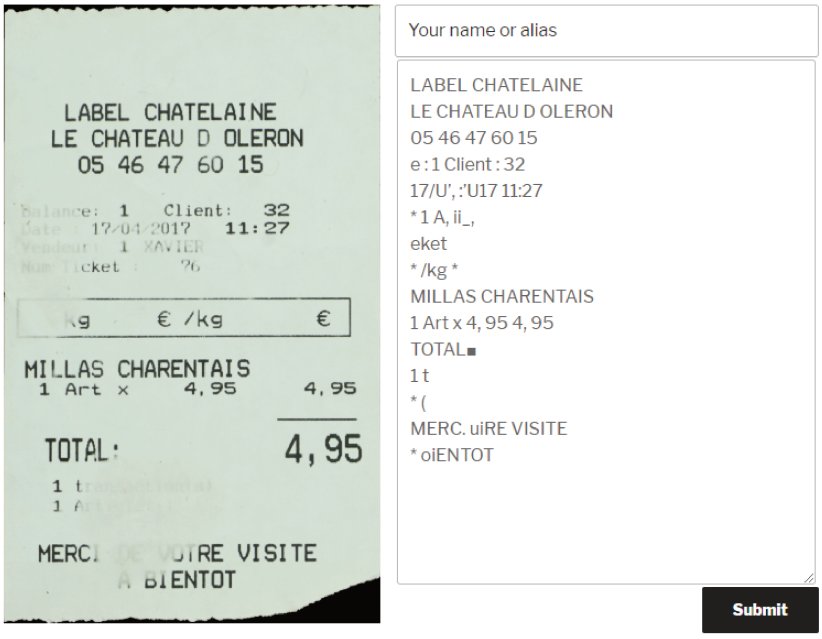


FIGURE 2.9 – Interface de correction participative.

plus de 90 noms différents, représentant plus d'une quarantaine de personnes physiques, pour 1706 tickets corrigés [4]. Ce classement a permis de stimuler les joueurs·ses en créant de la compétition entre elles·eux, ce qui a poussé les premier·e·s à corriger plus de 200 tickets chacun·e.

Ces premiers retours étant basés uniquement sur le nombre de tickets de caisse corrigés, nous avons par la suite décidé d'afficher d'autres statistiques. Concernant la longueur des corrections, nous affichons le nombre de caractères et de lignes traités par participant·e, ce qui permet de mettre en avant les joueur·se·s qui ne rafraichissent pas la page jusqu'à trouver des tickets courts. Par exemple, « Le Lion », bien qu'ayant corrigé 17 tickets de moins que « JeSappelleGroot », a traité 3590 caractères de plus, soit 187 lignes. Nous affichons également le nombre de caractères corrigés par personne, le pourcentage de corrections par caractères et la moyenne du nombre de corrections par ticket, ce qui permet de mettre en évidence les joueur·se·s ayant traité des tickets très mal reconnus et de constater que les taux de corrections sont parfois très différents d'un·e participant·e à un·e autre. Ainsi « Canard de l'espace », qui a corrigé 100 tickets, a un taux de corrections par ticket de 21,06, alors que « Le Lion », qui a corrigé 112 tickets, a un taux de corrections par ticket de 68,47. La moyenne, sur les 1111 premiers tickets corrigés, est de 50,64 corrections par ticket et de 9,94% de caractères ayant été corrigés.

Nous pouvons constater que ces mesures sont largement supérieures aux corrections automatiques que nous avions effectuées, ce qui prouve l'intérêt d'une telle démarche.

4. A la date du 15/09/2018. Le site étant encore fonctionnel, ce chiffre est en constante évolution.

L'autre intérêt de cette démarche de *crowd-sourcing* est, comme nous l'avons dit, le gain de temps dans le partage d'une tâche répétitive et fastidieuse. En effet, l'opération « correction manuelle » a été lancée le 23 novembre 2017 et environ 1500 tickets ont été corrigés en deux mois. Cependant, cette démarche apporte également son lot d'inconvénients : le temps gagné à ne pas corriger de tickets est réinvesti dans la maintenance de l'interface et dans la communication auprès des éventuel·le·s participant·e·s (mails, affiches, publications sur les réseaux sociaux) pour leur expliquer l'intérêt de la démarche et les inciter à participer, et auprès des volontaires pour leur expliquer les consignes et répondre à leurs questions. Un autre inconvénient de cette démarche est le nombre élevé de volontaires anonymes ayant corrigé un ou deux tickets dont nous ne pouvons pas être sûre qu'ils aient compris et appliqué les consignes sans vérifier chacun des tickets.

Andro & Saleh (2015), après avoir présenté plusieurs projets de correction OCR collaboratives par le jeu, établissent une distinction entre *crowdsourcing* explicite et *gamification*. En effet, même si le premier « fait souvent déjà largement appel aux ressorts du jeu puisqu'il classe les meilleurs contributeurs et les gratifie de récompenses symboliques ou réelles », ça n'est pas pour autant de la *gamification*, car « la manière dont les internautes sont invités à produire des données ne se fait pas sous la forme de jeux. » La *gamification*, si elle est beaucoup plus coûteuse à mettre en place, permet cependant de garder les joueur·se·s plus longtemps et de limiter la tricherie, et donc d'obtenir des bien meilleurs résultats que le simple *crowdsourcing*. Nous avons en effet pu constater que la mobilisation et la motivation de nos correcteur·rice·s se sont très vite affaiblies.

Cette correction collaborative améliore donc la qualité des textes et les rend d'avantage exploitables. Néanmoins, de nombreuses erreurs subsistent encore, et une deuxième passe sur le corpus est nécessaire pour en diminuer à nouveau le nombre et obtenir une correspondance quasi-parfaite entre l'image et le texte, afin de pouvoir extraire les informations correctement et vérifier que les documents sont bien authentiques.

2.4 Falsifications de documents

La dernière étape dans la préparation du corpus est la falsification des documents. Afin de créer un standard dans le domaine des *Document Forensics*, nous avons décidé de falsifier à la fois les images et les textes de nos tickets de caisse. Ce corpus pourra donc servir de référence pour l'évaluation des méthodes aussi bien basée sur l'image que sur le texte, comme nous le verrons dans le chapitre **??**, et de comparer notre méthode de vérification des informations textuelles à d'autres approches, utilisant en grande majorité des approches basées sur l'image.

2.4.1 Organisation de la fraude

Comme nous l'avons montré dans la partie 2.1, la plupart des travaux portant sur la détection des fraudes utilisent des corpus de documents qui ont été fraudés de manière aléatoire et/ou automatique. Si cela n'est pas gênant pour des méthodes basées sur

l'image, cette méthode risque d'apporter des biais que nous voulons éviter, comme le fait que les proportions de caractères modifiés ne soient pas représentatifs de la réalité, ou que les informations modifiées soient toujours sensiblement les mêmes. De plus, il est difficile de modifier des informations de façon aléatoire tout en gardant le type d'information au bon emplacement : il faudrait au préalable connaitre les parties de l'image qui correspondent à un type d'information (adresse, nom de produit, prix...) et les changer par des informations du même type, aussi bien dans l'image que dans le texte, au bon emplacement. Cela impliquerait d'avoir une base de connaissances dans laquelle piocher des informations, ainsi que la localisation parfaite du texte dans l'image, ce qui est difficile après les corrections automatiques et manuelles apportées sur les résultats de l'OCR.

Afin d'obtenir des fraudes variées, nous nous sommes donc inspirée du travail de Sidère et al. (2017) et avons demandé à de nombreuses personnes de modifier des documents, lors d'une journée où nous avons mis à disposition des ordinateurs équipés de plusieurs logiciels de retouche d'image (Paint, Paint 3D, GIMP, Photoshop, InkScape...), ainsi que de quoi grignoter (le fraudeur est gourmand). Vingt-cinq personnes, étudiant·e·s, doctorant·e·s, post-docs, enseignant·e·s-chercheur·se·s, personnels administratifs, informaticien·ne·s ou non, ont ainsi modifié environ dix tickets de caisse chacun·e, selon les consignes données, consultables dans l'annexe A. Les consignes sont volontairement floues afin de laisser toute la créativité voulue aux participants, même si nous avons donné quelques exemples de modifications possibles, à la fois en termes de manipulation de l'image et de choix de l'information.

Pour chaque ticket du lot de dix, le ou la fraudeur·se d'un jour avait donc pour consigne d'apporter une ou plusieurs modifications au ticket, en faisant en sorte de modifier le texte et l'image de la même façon. Ainsi, si le prix d'un produit est modifié sur l'image, par l'ajout d'un chiffre copié et déplacé au sein de l'image, l'information dans le texte doit être modifiée de façon à lire le nouveau montant dans le texte. En plus de cette double modification image-texte, nous avons demandé aux volontaires de remplir un fichier avec les méthodes de modification de l'image qu'ils avaient utilisées afin de pouvoir effectuer quelques statistiques.

2.4.2 Description des fraudes

Les manipulations de l'image se répartissent en cinq catégories, que nous avions par ailleurs proposées aux volontaires comme manipulations possibles. Les catégories utilisées sont signifiées par la même terminologie que celle utilisée par Sidère et al. (2017), à savoir :

— CPI : copier-coller à l'intérieur du document (ou *copy-move forgery*)
— CPO : copier-coller à l'extérieur du document (ou *splicing*)
— IMI : pour *imitation*, boite textuelle imitant la police de caractères du ticket
— CUT : suppression d'un ou plusieurs caractères
— Autres

La dernière catégorie (« Autres ») avait pour objectif de laisser libre cours à l'imagination des volontaires pour créer des fraudes auxquelles nous n'aurions pas pensé. Ces fraudes sont principalement des rajouts de traits dessinés à main (numérique) levée pour modifier une lettre ou un chiffre, se rapprochant alors de l'imitation. Deux tickets ont

Tableau 2.2 – Nombre de documents par combinaison de types de manipulations de l'image

Types de manipulations	Nombre de documents concernés
CPI	88
CPI CPO	10
CPI CPO CUT	3
CPI CPO IMI	2
CPI CUT	53
CPI CUT IMI	10
CPI IMI	13
CPI CPO CUT IMI	1
CPO	12
CPO CUT	4
CUT	22
CUT IMI	2
IMI	24
Autres	7
Total	251

Tableau 2.3 – Nombre de documents selon les types de manipulations de l'image

Types de manipulations	Nombre de documents concernés
CPI	180
CPO	32
CUT	95
IMI	52
Autres	7

été fraudés par ajout d'images de taches de café prise sur internet, particulièrement bien placées afin de cacher le montant des achats.

Le tableau 2.2 présente le nombre de documents en fonction des différentes combinaisons possibles de types de fraudes, classées par ordre alphabétique. Seulement 7 fraudes sortent des sentiers battus, les volontaires préférant largement les combinaisons comprenant des copier-déplacer, ou des suppressions, comme le montre le tableau 2.3, qui présente le nombre de documents comprenant au moins une occurrence de chacun des types de fraudes.

Cette répartition n'est pas étonnante : il est plus facile et plus rapide de sélectionner un caractère ou une série de caractères et de déplacer la sélection, que de faire une sélection sur un autre document, de la copier, puis de la coller au bon endroit dans le document à modifier. De même, la suppression d'éléments est une fraude moins visible à l'œil nu que l'imitation : là où il suffit de reprendre la couleur du fond du document pour l'appliquer sur la partie à effacer, il s'agit de trouver la bonne police avec la bonne taille, la bonne couleur et le bon alignement pour que l'imitation soit invisible. L'effort

fourni varie donc en fonction des types de manipulations de l'image, ce qui justifie la répartition présentée.

Si cette proportion de types de fraudes se justifie d'un point de vue méthodologique, il est difficile d'évaluer si elle est réaliste. En effet, nous n'avons aucune donnée concernant les types de manipulation de l'image dans la vie réelle, et nous ne pouvons que faire des suppositions. Dans un cas de fraudes où la personne doit effectuer la modification de l'image très rapidement, pour ne pas être surprise par un·e collègue par exemple, il y a fort à parier que le copier-déplacer (CPI) sera utilisé. Cependant, quelqu'un qui cherche à faire un faux indétectable et dispose de temps et des logiciels les plus perfectionnés préfèrera certainement copier des caractères d'autres documents ou créer des caractères imitant le style du document.

Concernant les informations modifiées, la créativité des volontaires s'est beaucoup plus révélée. En effet, si les informations concernant les montants dépensés sont les plus concernées par les modifications, ce qui était attendu vu la nature des documents, d'autres informations ont fait l'objet de modifications, d'ajout ou de suppression, comme le nom de la ville ou du magasin, l'adresse, le numéro de téléphone, la date, l'heure, le numéro de caisse, les poids, quantités et intitulés des produits, les devises, les informations sur la carte de fidélité, celles sur la TVA ou sur le numéro de SIRET, les conditions générales de vente ou encore les formules de politesse ou les slogans.

Ces modifications d'informations sont majoritairement des cas de fraudes réalistes : on peut vouloir changer le prix d'un produit pour percevoir un remboursement plus important, on peut changer une date de facture pour faire fonctionner la garantie, on peut modifier l'heure et/ou l'adresse pour avoir un alibi... Mais certaines ne le sont pas. En effet, la nature « réaliste » des fraudes a échappé à certain·e·s volontaires, et nous nous trouvons en présence de nombreux tickets modifiés, allant de l'exercice de style (modification quasiment aléatoire de caractères), au « troll » pur et simple. Tucker et al. (2018) définissent le « troll » sur internet, et plus précisément sur les réseaux sociaux, comme un « compte humain qui publie des contenus provocateurs, souvent en langage imagé et avec un contenu misogyne, soit par conviction politique soit simplement pour le « frisson » de le faire ». Sur internet, le « troll » peut être l'humain derrière le compte, ou bien, par métonymie, le contenu lui-même. Dans notre cas, il s'agit plutôt de blagues (parfois douteuses) et de modifications de contenu qui ont plus vocation à faire (sou)rire qu'à être utilisées pour frauder. On peut par exemple citer le ticket d'un magasin Biocoop qui s'est transformé en « Robocoop », ou le nom de la ville de La Rochelle transformé en « La Cholera ».

Ces cas, s'ils n'étaient pas attendus, se révèlent tout de même intéressants à exploiter, notamment dans l'idée où ils pourraient se retrouver sur internet, passant du « faux document » au « *fake* », c'est-à-dire à la fausse image qui circule sur internet dans un but, ici, humoristique. Le participant qui a modifié l'expression « ni échangé ni remboursé » par « ni échangé ni soumise » se rapprocherait alors, s'il était partagé sur les réseaux sociaux, de la définition de troll citée précédemment.

2.5 Conclusion

Comme nous l'avons vu, construire un corpus d'images et de textes de documents fraudés est une tâche longue et fastidieuse qui nécessite l'investissement de nombreuses personnes. En effet, que ce soit pour collecter des tickets de divers magasins et restaurants provenant de toute la France, pour corriger des milliers de textes issus de l'OCR ou pour modifier des centaines d'images et de textes correspondants, la participation de profils nombreux et variés était nécessaire. Cela nous a permis tout d'abord d'obtenir des tickets contenant toutes sortes de produits achetés, qu'il serait par ailleurs intéressant d'analyser des points de vue économique et sociologique, d'accélérer ensuite la création du corpus et enfin de diversifier les modifications graphiques et sémantiques.

Nous avons ainsi constitué un corpus d'environ 250 faux documents et 1250 documents réputés authentiques selon une méthodologie qui semble plus proche de la réalité de la fraude que tous les autres corpus utilisés dans le domaine des *Document Forensics*. Ce corpus est également exceptionnel dans le sens où il propose à la fois des images et leurs transcriptions, afin de proposer à la communauté un unique corpus de référence, permettant de comparer et/ou de combiner les approches de détection des fraudes par l'analyse de l'image ou de la sémantique du contenu. Nous avons présenté ce corpus dans Artaud et al. (2017).

Chapitre 3

Des données textuelles aux connaissances

Sommaire

Dans l'optique de vérifier les informations des documents pour les authentifier, ou au contraire, les détecter comme faux, nous devons d'abord extraire les informations du document, c'est-à-dire reconnaître le type d'information de chaque partie du texte. Dans le cas des tickets de caisse, il s'agira de repérer dans le texte ce qui est un nom de produit, ce qui est un prix, les adresses, les noms de magasins, les totaux, les numéros de SIRET et de nombreuses autres informations. Ces informations, que l'on peut assimiler à des entités nommées, doivent donc être extraites et sauvegardées de façon à conserver et à mettre en évidence les liens qu'elles ont entre elles. Ces liens nous permettront par la suite de créer des requêtes que nous enverrons sur les moteurs de recherche afin de comparer les informations du document et les informations du Web. Cette représentation nous permettra également d'enregistrer les informations de tous les tickets et de pouvoir les comparer entre elles.

Nous verrons dans un premier temps de ce chapitre les travaux qui existent sur les entités nommées, sur l'extraction d'informations et sur leur représentation. Dans un deuxième temps, nous présenterons le modèle ontologique que nous avons créé pour représenter les informations du document tout en gardant toute leur sémantique. Enfin, dans un troisième temps, nous analyserons le texte du document et nous décrirons notre approche pour extraire les informations et peupler l'ontologie.

3.1 Extraction et représentation des informations

De nombreux travaux existent en traitement automatique du langage naturel sur la recherche et l'extraction d'informations et la représentation des données textuelles, informations ou connaissances. Nous présenterons ici un rapide tour d'horizon de ces travaux, afin de définir les contours du domaine et de situer notre approche.

3.1.1 Données, informations, connaissances et entités nommées

Afin de mieux comprendre ce qui se trouve dans nos documents, il est important de réfléchir dans un premier temps aux différentes notions catégorisant le contenu sémantique des documents en général. Si nous avons, dans l'introduction, distingué les documents structurés, semi-structurés et non structurés en fonction de l'explicitation et de l'atomicité de leur contenu, nous cherchons ici à définir le type de contenu de nos documents. En effet, les documents qui nous intéressent ne contiennent pas de phrases, comme dans les documents analysés en linguistique de corpus, ni de méta-données ou d'annotations sur le contenu, qui permettraient de connaitre le type et le sens des données auxquelles nous avons affaire.

Cette définition en négatif n'est pourtant pas satisfaisante pour définir le contenu de nos documents. C'est pourquoi nous cherchons dans un premier temps de ce chapitre à comprendre la différence entre données, informations et connaissances, termes très discutés en sciences de l'information, afin de mieux appréhender le contenu de nos documents et de dégager nos besoins pour détecter les faux documents par leur contenu.

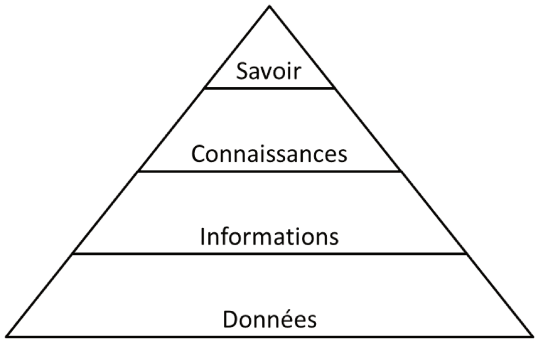


FIGURE 3.1 – Pyramide Data, Information, Knowledge, Wisdom (DIKW)

Les définitions d'Ackoff (1989) font date pour initier la réflexion sur ces différentes notions : les données seraient des symboles représentant les propriétés d'objets et d'événements ; l'information est issue de données traitées et est présentée de telle manière qu'elle peut répondre aux questions *qui*, *quoi*, *où* et *quand*, ce qui la rend plus utile que les simples données ; la connaissance passe une étape supplémentaire vers la compréhension du monde en permettant de répondre à la question *comment*. Aux trois termes précédemment cités (données, informations et connaissances) s'ajoute généralement un quatrième terme dans le domaine du traitement de l'information et de la connaissance : le savoir, défini par Ackoff (1989) comme l'interprétation des connaissances en fonction des valeurs et de la capacité cognitive de jugement.

Comme l'explique Rowley (2007), ces quatre termes sont souvent représentés de façon hiérarchique, comme dans la figure 3.1. Cette hiérarchie semble indiquer que le socle de toute représentation du monde – et nous entendons par là d'une partie du monde ou d'une vision du monde – est la donnée brute, le mot, le chiffre, le fait objectif et constaté, et qu'elle s'élève jusqu'au savoir à travers un traitement de l'information et des connaissances qui enrichit la représentation du monde par des liens sémantiques puis pragmatiques. La notion même de hiérarchie semble indiquer que le savoir est plus noble que la donnée, et cela peut peut-être s'expliquer par la difficulté d'acquisition et de traitement automatique : l'ordinateur peut tout à fait gérer des données, qui n'ont pas vraiment de signification en elles-mêmes, mais seul l'humain peut interpréter les connaissances pour atteindre ainsi le savoir, tout en haut de la pyramide, selon l'auteur.

Cette hiérarchie est parfois présentée non pas comme un empilement de classes séparées, mais comme un continuum allant du signal à la connaissance, comme le proposent Choo et al. (2000). Le signal provient de capteurs et présente une suite d'éléments qu'il convient de sélectionner et de structurer physiquement pour obtenir des données. L'humain ajoute ensuite de la signification à ces données, ainsi qu'une structure cognitive, afin de créer de l'information. L'ajout à cette information de croyances et de justifications, ainsi qu'une structure de ces opinions, permet d'accéder à la connaissance. Ce continuum se place ainsi à la fois sur une hiérarchie de structure du contenu, allant d'une structure physique à une structure de valeurs et en même temps sur une hiérarchie d'implication

de l'humain dans le processus de traitement, allant de la simple localisation des données dans le signal à l'apport de sa vision du monde.

A la lumière de ces définitions, nous pouvons considérer que le contenu de nos documents, sous la forme d'un texte brut, est un signal qu'il convient de traiter afin de séparer les éléments en unités structurées – des données – et de leur attacher leur signification pour qu'elles deviennent des informations. La frontière entre information et connaissance reste floue dans la littérature, comme le souligne Rowley (2007), même si l'on peut considérer que la contextualisation des informations et leur mise en relation entre elles participent plus de la gestion des connaissances que du traitement de l'information.

En traitement automatique des langues, c'est le terme « entité nommée » qui regroupe les données textuelles désignant des noms de choses qui existent, d'objets concrets ou abstraits. Ce terme est utilisé depuis la sixième Message Understanding Conference (MUC-6) (Grishman & Sundheim 1996) pour désigner des données textuelles à reconnaître, annoter, extraire de textes non-structurés pour former des informations structurées. Cette conférence avait pour but de partager des tâches d'extraction d'information, afin de développer des outils capables d'identifier les noms de personnes ou d'organisations et des noms de lieux géographiques, ainsi que des expressions numériques de type indications temporelles, monnaies ou pourcentages (Nadeau & Sekine 2007).

Les entités nommées sont définies par Sharnagat (2014) comme un mot ou un syntagme qui identifie clairement un élément d'un ensemble d'autres éléments qui ont des attributs similaires et qui ont un ou plusieurs « désignateurs rigides ». Ce terme est défini par Kripke (1980) comme des « entités fortement référentielles désignant directement un objet du monde ». Ces entités sont souvent des noms propres qui peuvent désigner des personnes, des organisations ou des lieux, mais peuvent également être des noms communs, dans des situations de coréférence par exemple. Les dates peuvent également être considérées comme des entités nommées, et sont appelées expressions temporelles ou expressions de localisation temporelles, comme les décrit, entre autres, Teissèdre (2012). Bick (2003), lui, prend en compte les noms d'objets et notamment de marque, de produit, de nourriture et de boisson dans les entités nommées qu'il veut extraire. Witten et al. (1999) intègrent dans leur définition « d'entités génériques » différentes coordonnées, comme les adresses, les adresse électroniques, les numéros de téléphone et de fax...

Les entités nommées sont donc des données textuelles auxquelles on ajoute une annotation, des données que l'on classe, afin d'en faire une information. La chaîne de caractères annotée peut ainsi répondre aux questions : qui, où, quand ou quoi.

3.1.2 Extraction d'informations

Extraire l'information, pour Serrano (2014), c'est généralement « construire une représentation structurée (bases de données, fiches, tableaux) à partir d'un ou plusieurs documents à l'origine non-structurés », c'est-à-dire compléter un modèle de données à partir d'éléments relevés dans des textes. Pour Muslea (1999), les systèmes d'extraction d'information s'appuient sur un ensemble de modèles d'extraction qu'ils utilisent pour

relever les informations pertinentes de chaque document analysé. Selon Hobbs & Riloff (2010), l'extraction d'information est le processus qui consiste à balayer le texte pour trouver de l'information pertinente pour certains sujets, y compris les entités, les relations et, ce qui est le plus difficile, les événements, c'est-à-dire qui a fait quoi, à qui, quand, et où. Cette deuxième définition se rapproche de la définition d'Ackoff (1989) sur les informations, et nécessite de relever bien plus que de simples données, ou des simples mots-clés. Ces trois éléments – les entités (nommées), les relations et les événements – constituent trois branches de l'extraction d'information. Ces trois types d'extraction d'information étaient souvent cantonnés à une langue, à un domaine ou à un genre de textes, dépendant principalement du ou des corpus utilisés (Poibeau 2003), mais les approches visent de plus en plus à être génériques et/ou multilingues (Poibeau et al. 2012).

La reconnaissance d'entités nommées est l'une des tâches principales de l'extraction d'information, et même du traitement automatique des langues, selon Sharnagat (2014). Si les méthodes de reconnaissance d'entités nommées consistaient principalement en des règles faites à la main jusque dans les années 1990, l'avènement de l'apprentissage automatique et de l'apprentissage profond ont permis d'assouplir et d'améliorer les résultats. Ainsi Nadeau & Sekine (2007), pour la période 1991 – 2006 puis Sharnagat (2014) pour les années suivantes, dressent un panorama des travaux réalisés et divisent les approches d'apprentissage en trois catégories classiques : supervisées (modèles de Markov cachés, modèles basés sur l'entropie maximale, machines à vecteurs support, champs aléatoires conditionnels, arbres de décision...), semi-supervisées (ré-échantillonnage) et non-supervisées (clustering, comparaison de contexte ou de corpus parallèles...). Les approches supervisées nécessitent des grands corpus annotés et de nombreux indices pour l'apprentissage, alors que les approches non-supervisées ne prennent en entrée que des textes et utilisent des ressources externes, comme WordNet par exemple. A partir de 2010, des approches utilisant des réseaux de neurones ont émergé pour reconnaitre les entités nommées, comme dans les travaux de Collobert et al. (2011) et Lample et al. (2016).

Les indices utilisés pour apprendre à reconnaitre les entités nommées se situent par exemple au niveau des caractères du mot ou de l'expression : la casse, la ponctuation, la présence ou non de chiffres, la présence de certaines suites de lettres marquant un préfixe, un radical ou un suffixe, la longueur des mots... L'utilisation de dictionnaires et de patrons est également fréquente pour détecter certaines entités nommées, ainsi que l'analyse syntaxique des phrases (fonctions des mots, appositions, incises), l'étude du contexte (les mots autour) et de manière plus large du corpus utilisé (co-référence, multiples occurrences, fréquences des termes...). Ces indices sont également utilisés dans les règles d'extraction réalisées à la main, sans apprentissage automatique. Certains outils, comme Unitex (Paumier 2003) ou Nooj (Silberztein 2007), utilisent des automates à états finis, combinés avec des informations sur les *tokens* (nature et fonction du mot dans la phrase notamment) afin de repérer des segments de phrases, des expressions ou des mots, comme des entités nommées par exemple. Les *tokens* sont des suites de caractères qui sont délimitées par divers séparateurs, généralement des espaces ou des retours à la ligne. Les *tokenizers*, ou analyseurs lexicaux, créent des listes de *tokens*,

généralement des mots, mais parfois de la ponctuation ou des chiffres, à partir d'une chaîne de caractères. Les phrases sont également séparées les unes des autres par ces deux outils afin de les analyser syntaxiquement et d'attribuer à chaque mot un label concernant sa nature (verbe, adjectif, préposition...).

L'extraction de relations a pour objectif d'extraire des faits, des actions, des relations entre deux entités. C'est, typiquement, le verbe dans la phrase, ou le prédicat en linguistique. Cela permettrait par exemple de répondre directement aux requêtes posées dans un moteur de recherche, et est nécessaire dans un système de questions-réponses ou de *chatbots*. Les approches et les indices utilisés pour l'extraction de relations, sont les mêmes que pour l'extraction d'entités nommées. L'outil KnowItAll, présenté par Etzioni et al. (2004), fournit en entrée de son application une ontologie et un petit nombre de règles d'extraction pour chaque classe et propriété de l'ontologie. Il envoie ensuite des requêtes sur différents moteurs de recherche pour obtenir un grand nombre de textes pour chaque classe et répartit ensuite à l'aide d'un classifieur bayésien naïf les différentes occurrences d'une même entité, afin de peupler l'ontologie. Banko et al. (2007) introduisent l'*Open Information Extraction*, l'extraction d'information ouverte, qui a pour ambition de se passer de l'aide humaine, sur l'extraction de relations sur des grands volumes de données. Ainsi, leur outil TextRunner permet d'extraire des relations non-prédéterminées là où l'outil KnowItAll (Etzioni et al. 2004) prenait en entrée des noms de relation.

Hogenboom et al. (2016) recensent les différents travaux et courants majeurs dans le domaine de l'extraction d'événements souvent utilisée dans les systèmes d'aide à la décision. De nombreuses applications peuvent avoir besoin de détecter quelque chose qui se passe sur une certaine période de temps, comme en épidémiologie (Lejeune et al. 2015), ou dans la surveillance anti-terroriste (Conlon et al. 2015). Hogenboom et al. (2016) séparent les approches conduites par les données, plus orientées statistiques et apprentissage automatique, et les approches conduites par les connaissances, plus orientées règles et modèles lexicaux, syntaxiques et sémantiques. Ces approches aujourd'hui se combinent et sont utilisées conjointement pour améliorer les systèmes.

3.1.3 Peuplement d'ontologies

La thèse de Fotsoh (2018) décrit de nombreuses façons de représenter les entités nommées. L'auteur réalise une première distinction entre entités nommées élémentaires, qui correspondent aux définitions données précédemment, et entités nommées complexes, qui sont des « entités composées de plusieurs propriétés », qui peuvent être d'autres EN, élémentaires ou complexes, ou du texte de manière générale. L'idée de cette distinction est de séparer les noms des entités, ou les syntagmes qui les représentent, de leur essence conceptuelle : une personne a un nom – son désignateur rigide, mais est aussi un concept, une entité, qui est attachée à d'autres informations, comme son numéro d'identité, sa date de naissance, son genre... Cette première distinction lui permet de distinguer les différents modèles de représentation des EN dont les modèles de types XML, comme ENAMEX (Grishman & Sundheim 1996) et les modèles ontologiques sont les principaux types. L'ontologie est selon Fotsoh (2018), l'une des ressources les plus utilisées pour l'extraction des EN.

L'ontologie est, selon Gruber (1993), la « spécification explicite d'une conceptualisation partagée pour un domaine de connaissance ». Cette définition met en évidence le principe d'application d'une ontologie : toute ontologie est construite pour un domaine, pour représenter une partie du monde. Chandrasekaran et al. (1999) donnent une définition plus précise en écrivant que « les ontologies sont des théories sur les types d'objets, les propriétés des objets et les relations entre les objets qui sont possibles dans un domaine de connaissance donné. Elles fournissent des termes potentiels pour décrire notre connaissance du domaine. »

Selon Lehmann & Völker (2014), il existe de nombreuses réalités derrière ce que les gens nomment « ontologie » : dictionnaire, taxonomie, thésaurus, ou formalisations de haut niveau. Il existe néanmoins des standards de notations d'ontologies dans le domaine de la représentation des connaissances et du web sémantique, comme OWL2 [1] (Web Ontology Language) qui s'écrit avec une syntaxe RDF/XML. OWL2 est un langage d'ontologie proposé par le W3C pour le web sémantique. Ce langage peut être représenté par un graphe RDF (Resource Description Framework), qui est un modèle proposé également par le W3C permettant l'interopérabilité des documents et le partage d'informations non-structurées.

L'ontologie est donc le moyen de structurer et documenter un domaine de connaissance et de développer une compréhension commune des concepts qui le forment (Lehmann & Völker 2014). De plus, le besoin et la difficulté d'accès structuré à l'ensemble des données du Web augmentent avec le nombre de données, et nécessitent des moyens automatiques pour acquérir et organiser la connaissance afin de la représenter. Maedche & Staab (2001) ont donc introduit le terme « ontology learning » ou « apprentissage d'ontologie », qui consiste à générer une ontologie à partir de textes, d'inférences logiques, de données hétérogènes ou de fusions d'ontologies.

Parfois, le modèle de l'ontologie est déjà construit, généralement par des experts du domaine, même si cela peut être très coûteux en termes de temps et d'argent, comme le calculent Simperl et al. (2012). Il s'agit alors de peupler automatiquement l'ontologie d'instances en utilisant des sources variées, d'enrichir ou d'adapter des ontologies existantes grâce à des méthodes statistiques ou de traitement automatique des langues à partir de documents non-structurés, ou des approches de *data mining* ou structurelles sur des documents semi-structurés, comme les décrivent Hazman et al. (2011).

Des outils permettent aux utilisateurs de créer et/ou peupler des ontologies à partir de textes. C'est le cas par exemple de Text2Onto (Cimiano & Völker 2005). Cet outil utilise des règles d'extraction JAPE (*Java Annotation Patterns Engine*), ainsi que la détection des phrases, la tokenisation et l'analyse en partie du discours, toutes intégrées à la plateforme GATE (*General Architecture for Text Engineering*) (Cunningham et al. 2002) de traitement automatique des langues. Il utilise également une approche statistique : un algorithme de POM (*Probability Ontology Model*), ainsi que les fréquences des termes dans les documents. Une similarité à partir de vecteurs est également calculée entre les différents termes et concepts pour établir les liens d'équivalence et pour attri-

1. Les spécifications de ce langage sont présentées sur : `https://www.w3.org/TR/owl2-overview/`, consulté le 30/11/2018

buer la bonne instance au bon concept. De même, toujours intégré à la plateforme GATE, un autre outil, plus récent, a été mis en place par Maynard et al. (2009) : SPRAT (*Semantic Pattern Recognition and Annotation Tool*). Cet outil utilise des règles proposées par Hearst (1992), d'autres règles lexicales et syntaxiques, l'outil d'extraction d'entités nommées ANNIE, ainsi que les mêmes composantes de GATE que Text2Onto, et plusieurs autres dont l'analyseur morphologique, le séparateur de syntagmes nominaux, des nomenclatures et des automates.

Dans le domaine de l'ingénierie des connaissances et particulièrement de l'apprentissage d'ontologie, on se sert donc, entre autres, du traitement automatique des langues et de l'extraction des entités nommées pour construire et peupler les ontologies. Dans le domaine de l'extraction d'information, on se sert de l'ontologie pour contraindre le domaine d'extraction, pour aider le système à spécifier les informations ou entités recherchées dans un texte non-structuré. Ainsi, le terme « OBIE », pour *Ontology-based information extraction*, est utilisé par Maynard et al. (2006) d'un point de vue centré sur le peuplement d'ontologie et son évaluation conceptuelle, et est repris plus tard par Wimalasuriya & Dou (2010) qui dressent un paysage des recherches sur l'OBIE d'un point de vue plus focalisé sur l'extraction d'informations et de données textuelles. Ces derniers définissent l'OBIE comme « un système d'extraction d'information basé sur l'ontologie : un système qui traite un texte en langage naturel non structuré ou semi-structuré au moyen d'un mécanisme guidé par des ontologies pour extraire certains types d'information et présenter les résultats en utilisant des ontologies ».

Pour finir, Shah & Jain (2014) placent l'extraction d'information guidée par l'ontologie à l'intersection de cinq grands domaines : l'extraction d'informations, bien sûr, avec l'extraction de termes et de relations, la recherche d'information, avec des méthodes de requêtes et de TF-IDF, le traitement automatique des langues, avec l'analyse de corpus et les diverses analyses lexicales et syntaxiques, l'apprentissage automatique, avec l'utilisation de classifieurs notamment, et enfin le web sémantique, avec les ontologies. Cette pluridisciplinarité montre la complexité d'une telle ambition, que de chercher à extraire le sens d'un texte, la signification des entités et les relations qu'elles entretiennent entre elles, et d'automatiquement les conceptualiser pour leur faire une place dans la modélisation de l'information, de la connaissance.

3.2 Un modèle ontologique pour mettre en évidence la sémantique du document

Nous avons choisi de modéliser les informations des documents dans une ontologie OWL2, exprimée en RDF/XML. Nous expliquerons tout d'abord ce choix, malgré le fait que nous ne pouvons pas utiliser les méthodes précédemment décrites pour apprendre ou peupler cette ontologie étant donné la nature de nos documents. Nous présenterons ensuite les concepts de notre ontologie ainsi que les propriétés qui les relient.

3.2.1 Choix de l'ontologie

Le contenu du document est un contenu qui contient une structure implicite : l'emplacement des composantes textuelles les unes par rapport aux autres, les formats de ces composantes textuelles (chiffres, toutes lettres, syntaxe) et leur sémantique en font des informations qui répondent à toutes sortes de questions sur nos documents (qui (ou quelle entreprise) les émet, où, quand, et pour quel motif ?), à condition que nous soyons humains. En effet, il est difficile pour une machine de voir spontanément le lien sémantique entre deux données textuelles juxtaposées, c'est-à-dire quelques caractères séparés de quelques centimètres sur l'image, ou deux suites de caractères concomitantes dans du texte sans grammaire de langue naturelle, et de désigner quelle partie du document correspond à quelle question.

Nous avons donc choisi de modéliser ces informations dans une ontologie afin de transposer ce que l'humain peut analyser facilement, au premier coup d'œil, par sa connaissance. En effet, l'humain sait comment les informations sont disposées dans le document, comment se présente une adresse ou comment lire un tableau... Il sait également faire le lien tout seul entre deux données textuelles : une adresse sous un nom d'entreprise correspond à l'adresse de l'entreprise. Il y a donc une double relation : l'entreprise e a pour adresse l'adresse a et l'adresse a est l'adresse de l'entreprise e. On peut également noter que l'entreprise peut avoir plusieurs adresses, et qu'une adresse peut être partagée par plusieurs entreprises, dans le cas d'un immeuble de bureaux par exemple.

Un autre avantage de l'ontologie est sa flexibilité : on peut facilement ajouter des concepts ou des propriétés, sans casser sa structure. Cela nous permet notamment de créer des concepts dynamiquement grâce à nos règles d'extraction d'information, que nous présenterons dans la partie 3.3.

Nous avons également choisi l'approche ontologique pour ses raisonneurs : depuis des données, que nous extrayons de nos documents, nous pouvons déduire automatiquement des inférences. Par exemple, si deux noms d'entreprise différents ont le même numéro de téléphone, alors que nous avons spécifié qu'un numéro de téléphone ne pouvait appartenir qu'à une seule entreprise, alors, il est probable qu'il n'y ait *en réalité* qu'une seule entreprise.

L'ontologie que nous allons présenter ici concerne les tickets de caisse uniquement, mais pourrait tout à fait être adaptée pour prendre en compte tous les documents. De fait, la plupart des informations que nous relevons ne sont pas propres aux tickets de caisse et pourraient se trouver sur d'autres documents émis par des entreprises ou des administrations.

3.2.2 Les concepts

L'ontologie que nous avons créée pour représenter les informations des tickets de caisse est assez simple. Comme toutes les ontologies, elle possède une classe `Thing`, mère de toutes les autres classes. Les concepts utilisés sont illustrés dans la figure 3.2 qui est une capture d'écran du logiciel Protégé (Musen 2015) que nous utilisons pour visualiser les ontologies. Cette figure montre l'arborescence des concepts de notre ontologie.

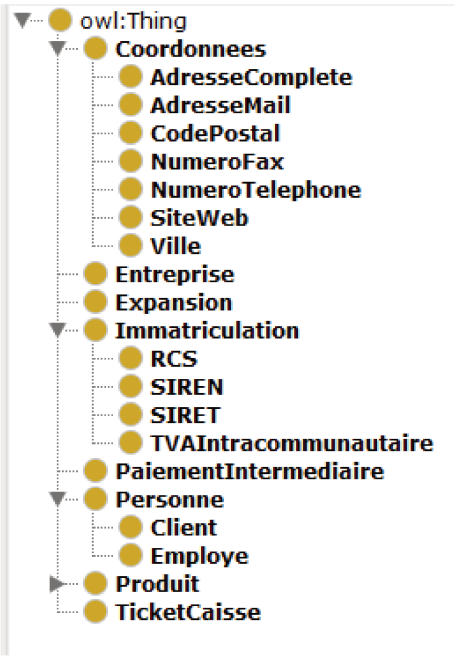


FIGURE 3.2 – Concepts de l'ontologie représentant le contenu des documents

Nous avons fait le choix de créer cette ontologie en français, étant donné qu'elle concerne des documents écrits en langue française et qu'elle contient des éléments spécifiques à la France, comme les informations sur les entreprises par exemple. Cependant, nous pourrions rajouter des éléments `rdfs:label` avec l'attribut `xml:lang` dans la définition des classes afin de la rendre lisible en plusieurs langues.

TicketCaisse

Le concept TicketCaisse représente le document en lui-même. Dans une ontologie qui prendrait en compte d'autres documents, nous aurions pu créer une classe `Document` qui aurait eu pour sous-classe `TicketCaisse`, mais aussi tous les autres types de documents, comme `Facture`, `BulletinSalaire`, `CertificatMedical`... Nous avons cependant fait le choix de restreindre ici notre ontologie à la seule représentation des tickets de caisse afin de ne pas la complexifier davantage. Cependant, ajouter d'autres documents à l'ontologie ne changerait rien à la démarche et à l'utilisation de l'ontologie.

Les instances de tickets de caisses sont nommées par le numéro qui leur est attribué depuis l'extraction de chaque image de ticket, et qui est le même que celui du fichier image et des fichiers texte au fur et à mesure de leurs corrections, cf. chapitre 2.

PaiementIntermediaire

Le concept `PaiementIntermediaire` représente un paiement inscrit sur un ticket de caisse. Cette classe est née de l'observation de nos tickets : le règlement d'un montant

peut se faire en plusieurs parties, ou en plusieurs moyens de paiement. Par exemple, vous pouvez effectuer un règlement avec deux tickets restaurants puis compléter en espèces. De même, le paiement en espèces nécessite souvent de rendre de la monnaie, ce qui fait que la ligne extraite pour le paiement en espèces n'est pas nécessairement le montant payé, après rendu de monnaie. Nous avons donc choisi de représenter cette information par un concept car elle contient plusieurs propriétés intrinsèques, ou caractéristiques, qui lui sont attachées comme nous le décrirons dans la section suivante. Les instances de ce concept sont simplement nommées par un préfixe `paiementintermediaire` suivi d'un numéro.

Produit

Le concept `Produit` regroupe tous les noms de produits qui sont inscrits sur les tickets de caisse, qui sont référencés dans des sous-concepts dont le nom est le nom d'un type de produit sans espace, sans chiffres et sans caractères spéciaux (par exemple `CREVETTESROSES`) qui peut avoir pour instances plusieurs occurrences sur un ou plusieurs tickets comme dans la figure 3.3.

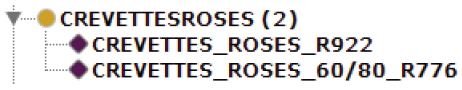


FIGURE 3.3 – Exemple d'un sous-concept de `Produit` avec ses instances.

Coordonnees

Le concept `Coordonnees` regroupe plusieurs sous-concepts qui relèvent toutes les coordonnées de l'entreprise ayant émis le ticket de caisse. Ainsi, si ces informations sont écrites sur le ticket, nous relevons l'adresse complète, formée à partir de l'adresse, du code postal et de la ville, le site web, l'adresse mail ainsi que les numéros de fax et de téléphone. Toutes ces informations sont des sous-concepts de `Coordonnees` car nous considérons qu'elles sont toutes des coordonnées. Nous avons séparé les éléments de l'adresse complète pour des raisons liées à l'extraction, qui ne nous permet pas toujours d'avoir tous ces éléments. Nous avons donc choisi d'être exhaustifs.

Entreprise

Le concept `Entreprise` représente l'entité émettrice du ticket de caisse. Dans le cas d'une ontologie prenant en compte plusieurs types de documents, il aurait certainement fallu créer une distinction entre émetteur et destinataire, ainsi qu'entre les différents types d'entités ou d'institutions pouvant émettre et recevoir les documents. Nous avons choisi de considérer que les tickets de caisse étaient émis par des entreprises (restaurants, magasins, commerces ambulants, etc.), même s'il peut arriver que certains de nos tickets soient émis par des associations. La terminologie est alors erronée.

Il aurait été intéressant de créer des sous-concepts pour les différents types d'entreprises, voire d'émetteurs de documents, mais nous n'avons que rarement les moyens d'extraire automatiquement ces informations des tickets de caisse, car elles sont rarement explicites. En effet, si l'humain peut déduire de la taille et de la police d'un ticket, ainsi que des produits qu'il mentionne, qu'il provient plutôt d'un achat sur le marché que d'une grande surface, notre algorithme en l'état ne le devine pas.

Immatriculation

Les entreprises possèdent toutes un numéro unique d'identification – le SIREN – qui est un numéro à 9 chiffres attribué par l'INSEE (Institut National de la Statistique et des Études Économiques).

Le numéro SIRET est composé du numéro SIREN et de 5 chiffres supplémentaires qui permettent d'identifier chaque établissement d'une même entreprise. Le numéro SIREN est permanent à la vie de l'entreprise, tandis que le SIRET change à chaque déménagement de l'établissement.

Certains tickets présentent également l'immatriculation au Registre du Commerce et des Sociétés (RCS), qui est composé des lettres « RCS » suivies du nom de la ville puis du SIREN.

Une autre mention que l'on peut également trouver sur les tickets est le numéro de TVA intracommunautaire. Ce numéro sert à identifier auprès de l'administration fiscale toute entreprise assujettie à la TVA dans l'Union Européenne. En France, cette immatriculation commence par les lettres « FR », puis une clé de deux chiffres calculée par l'algorithme de Luhn sur le SIREN et enfin, le SIREN justement.

Tous ces numéros sont des immatriculations, ce qui en fait des sous-classes du concept `Immatriculation`.

Personne

Le concept `Personne` a pour sous-concept dans notre ontologie `Employe` et `Client`. En effet, il est fréquent que le prénom des employés ayant servi ou encaissé le client soit inscrit sur le ticket de caisse, tout comme le nom du client quand une carte de fidélité est enregistrée chez le commerçant. Etant donné que nous avons retiré de notre corpus tous les tickets non anonymes, nous ne pouvons pas instancier le sous-concept `Client` dans notre travail pour des raisons de publications des données. Nous pouvons toutefois extraire les prénoms des employés, ou de manière plus générale, des gens qui travaillent pour l'émetteur du ticket de caisse. Là aussi, la terminologie est donc légèrement abusive, car les instances de la classe `Employe` pourraient tout à fait être le ou la chef d'entreprise ou un·e bénévole d'association, mais ce n'est généralement pas explicité.

Expansion

Comme nous le verrons plus en détail dans le chapitre 4, notre corpus contient beaucoup d'abréviations, notamment dans les noms de produits, ce qui est un obstacle à

la confrontation des informations extraites des documents aux informations supposées vraies que l'on pourrait trouver sur internet (*cf.* chapitre 5). Nous avons donc décidé de créer un concept `Expansion` qui prend pour instances des noms de produits non abrégés, étant de fait plus lisibles pour l'humain et plus adaptés à la recherche d'information sur le web.

Nous aurions pu également placer ces instances en propriété de type *string* des instances de noms de produits, mais cela aurait eu pour effet de répliquer l'information inutilement, puisque plusieurs noms de produits abrégés peuvent avoir la même expansion. De plus, une expansion n'est pas une simple propriété d'un nom abrégé, mais bien son équivalent en terme sémantique.

Tous les concepts de notre ontologie sont disjoints sauf les enfants de `Produit`. Cela signifie que les instances d'un concept ne peuvent pas être celles d'un autre concept, sauf si c'est un nom de produit. En effet, une entreprise ne peut pas être un produit vendu, du moins pas sur un simple ticket de caisse, alors qu'une banane bio pourrait également être une banane Cavendish. La méthode de peuplement de l'ontologie que nous utilisons et que nous détaillerons dans la section 3.3 ne nous permet pas d'implémenter ces relations sémantiques, mais cette représentation du monde nous paraissait plus proche de la réalité.

3.2.3 Les propriétés

Comme nous l'avons expliqué précédemment, l'un des intérêts de l'ontologie réside dans les propriétés, c'est-à-dire ce qui définit les concepts – et donc leurs instances – qui sont les *Object Properties* ou « propriétés de type objet » d'une part, et les *Data Properties*, ou « propriétés de type données », d'autre part. Les propriétés, tout comme les concepts, peuvent avoir des sous-propriétés et également des contraintes, de nombre de caractères par exemple ou de valeurs maximales ou minimales. Les relations sont les instances des propriétés entre les individus.

Propriétés d'objet

Les propriétés d'objet (`owl:ObjectProperty`) représentent des liens entre des concepts. Ces liens sont orientés et vont d'un domaine (`rdfs:domain`) vers une image (`rdfs:range`) [2]. Les propriétés inverses relient les deux mêmes concepts mais avec une inversion du `rdfs:domain` et du `rdfs:range`. Déclarer $P1$ `owl:InverseOf` $P2$ revient à dire que pour tous concepts x et y, $P1(x, y) \iff P2(y, x)$. De même les propriétés peuvent être :

— transitives : $P(x, y)$ et $P(y, z) \Rightarrow P(x, z)$
— symétriques : $P(x, y) \iff P(y, x)$
— fonctionnelles : $P(x, y)$ et $P(x, z) \Rightarrow y = z$

La figure 3.4 et les tableaux 3.2 et 3.1 illustrent l'organisation de l'ontologie et les propriétés qui existent entre les concepts.

2. Le terme « image », qui traduit *range*, est utilisé dans la version francophone du guide du langage d'ontologie Web OWL disponible à l'adresse `http://www.yoyodesign.org/doc/w3c/owl-guide-20040210/`, consultée le 16 novembre 2018.

Tableau 3.1 – Propriétés d'objet

Domaine	Propriété d'objet	Propriété inverse	Image
Ville	a_code_postal	est_code_postal_de	CodePostal
Entreprise	a_coordonnee	est_coordonnee_de	Coordonnee
Entreprise	a_adresse	est_adresse_de	AdresseComplete
Entreprise	a_adresse_mail	est_adresse_mail_de	AdresseMail
Entreprise	a_fax	est_fax_de	NumeroFax
Entreprise	a_site_web	est_site_web_de	SiteWeb
Entreprise	a_telephone	est_telephone_de	NumeroTelephone
Entreprise	a_emis	est_emis_par	TicketCaisse
Produit	a_expansion	est_expansion_de	Expansion
Entreprise	a_immatriculation	est_immatriculation_de	Immatriculation
TicketCaisse	a_paiement_intermediaire		PaiementIntermediaire
Client	achete	est_achete_par	Produit
TicketCaisse	concerne_achat		Produit
TicketCaisse	contient	est_inscrit_sur	Produit, Immatriculation, Coordonnees
SIREN	comporte	est_composante_de	SIRET, RCS, TVAIntracommunautaire
Entreprise	emploie	est_employe_par	Employe
Ville, CodePostal	est_partie_de		AdresseComplete
Entreprise	se_situe_a		Ville
Entreprise	vend	est_vendu_par	Produit

Les propriétés `a_adresse`, `a_adresse_mail`, `a_fax`, `a_site_web` et `a_telephone` sont des sous-propriétés de `a_coordonnees`, de même que leurs propriétés inverses sont des sous-propriétés de `est_coordonnee_de`.

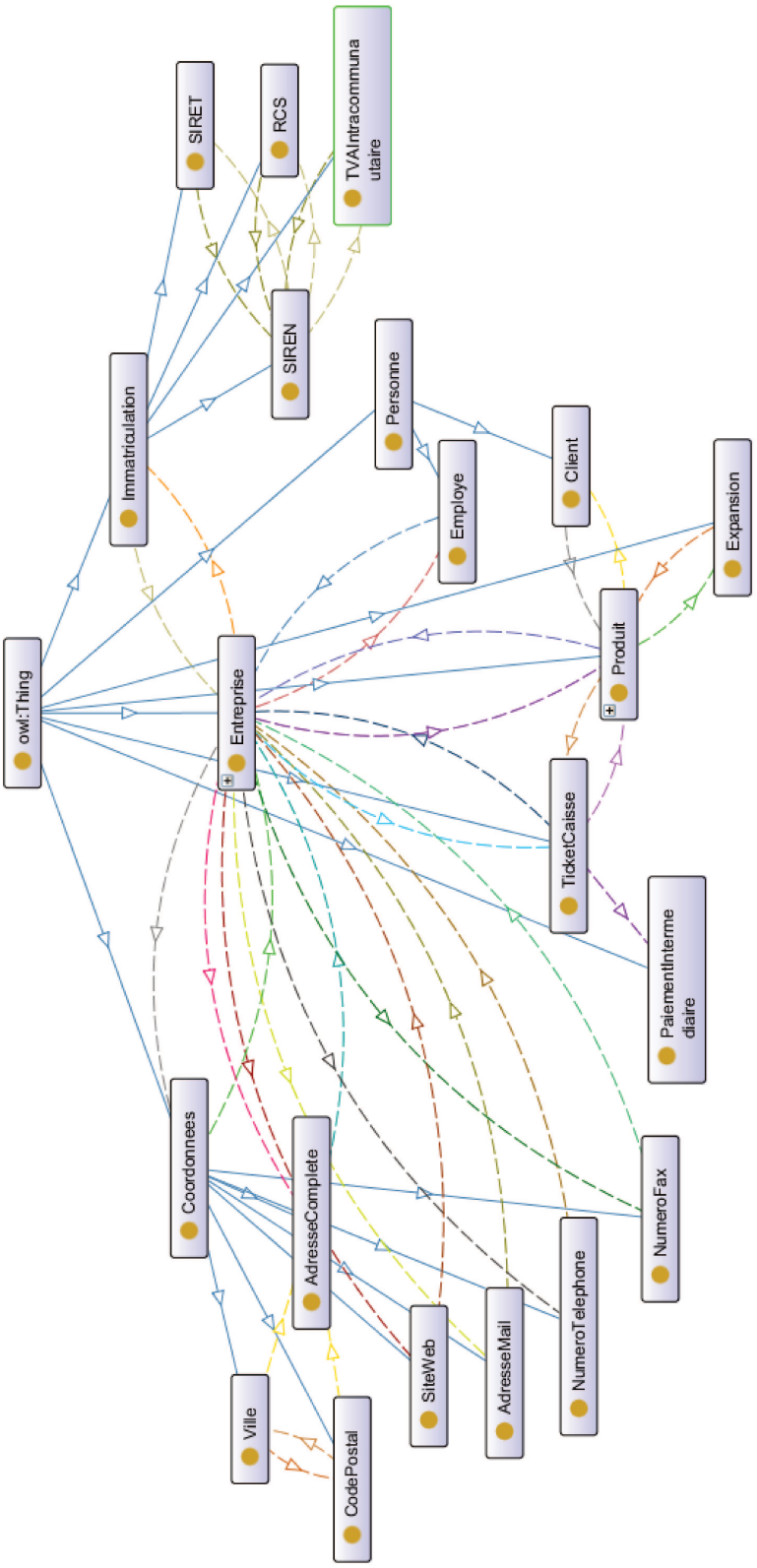
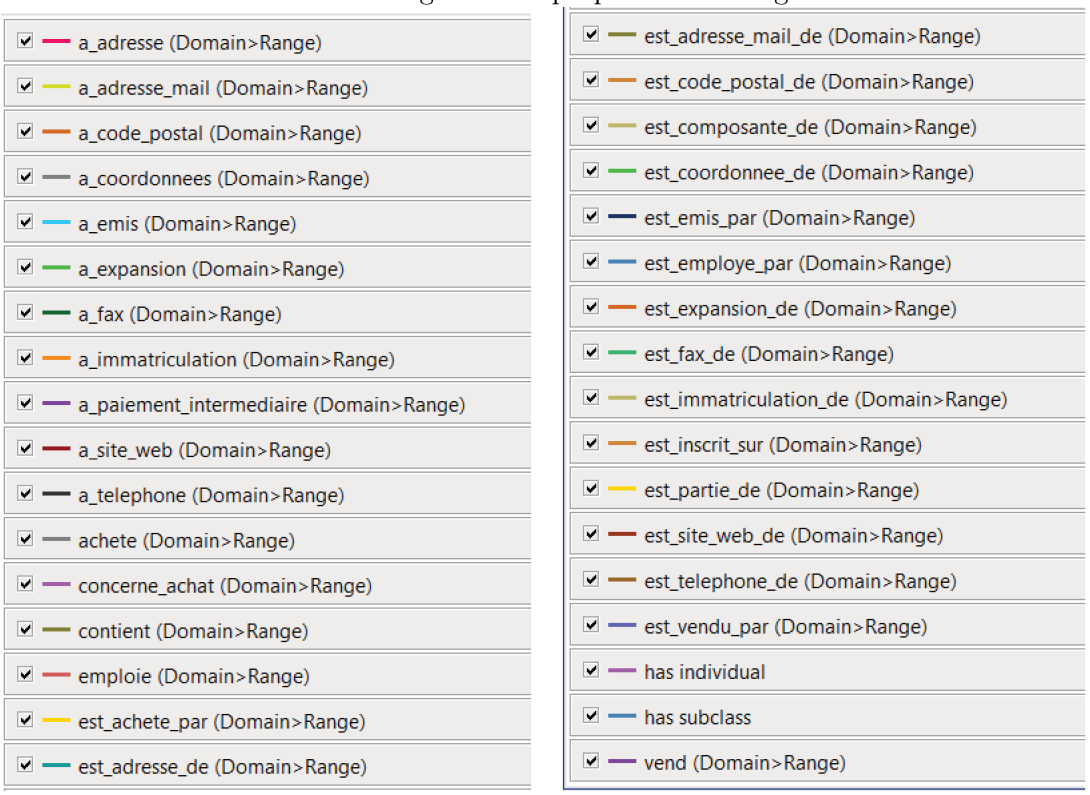


FIGURE 3.4 – Organisation de l'ontologie (visualisation par OntoGraph2.0.3)

Tableau 3.2 – Légende des propriétés de la figure 3.4



Propriétés de type données

Les propriétés de type données (`owl:DatatypeProperty`) sont des propriétés intrinsèques aux concepts. Nous en avons défini 13, qui s'appliquent à seulement trois concepts : `TicketCaisse`, `Produit` et `PaiementIntermediaire`.

Le ticket de caisse a été émis à une date et à une heure précises. Nous avons donc les attributs `a_date` et `a_heure`, respectivement de types `xsd:date` (aaaa-mm-jj) et `xsd:time` (hh :mm :ss). Le ticket de caisse présente également un montant total (`a_montant_total`) et un paiement total (`a_paiement_total`), qui sont tous deux de type `xsd:decimal`, c'est-à-dire des nombres décimaux. Enfin, le nombre d'articles vendus est parfois inscrit sur les tickets de caisse. Nous l'avons donc extrait dans une donnée de type `xsd:integer`, c'est-à-dire un nombre entier, dans l'optique de vérifier que ce chiffre correspond bien au nombre de produits relevés, sans quoi une fraude serait possible (à moins que ce ne soit une mauvaise extraction ou une erreur de l'OCR).

Les produits possèdent également plusieurs propriétés de type données : un poids et un prix au kilogramme, un prix unitaire et une quantité et un prix total. La propriété `a_prix_total`, de type `xsd:decimal` est, normalement, inscrit sur le ticket pour chaque produit. Les quatre autres propriétés dépendent du type de produit et ne sont donc pas

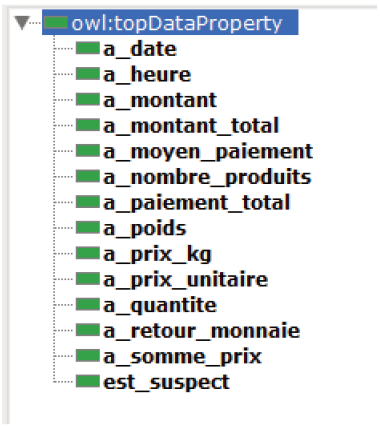


FIGURE 3.5 – Propriétés de type données

toujours présentes. Les propriétés `a_poids` et `a_prix_kg`, tous deux de type `xsd:decimal`, sont généralement indiquées pour les fruits et légumes, et parfois pour la viande. Le prix total doit correspondre alors – s'il n'y a pas de fraudes – au résultat d'une multiplication entre ces deux données. De même, si plusieurs produits identiques sont vendus, certains magasins les regroupent sur le ticket de caisse et précisent le nombre du même produit acheté, que nous représentons par une donnée `a_quantite` de type `xsd:integer`. Cette quantité est parfois accompagnée du prix de l'article ou du service à l'unité, représenté par `a_prix_unitaire`, de type `xsd:decimal`. Là encore, s'il n'y a pas de fraude ni d'erreurs dans l'extraction ou dans la reconnaissance de caractères, nous devrions avoir `a_prix_total` = `a_quantite` × `a_prix_unitaire`.

Les paiements intermédiaires sont caractérisés par plusieurs propriétés de type données, à savoir : `a_montant`, `a_moyen_paiement` et `a_retour_monnaie`. Le premier, de type `xsd:decimal`, représente le montant donné par le client au commerçant. Le deuxième, de type `xsd:string`, est une chaîne de caractères extraite du ticket, qui représente le moyen utilisé pour payer : espèces, chèque, carte bancaire avec ou sans contact, ticket restaurants, EMV (*European Mastercard Visa* qui est le standard international de sécurité des cartes de paiement à puces)... Si le paiement se réalise en espèces, le montant rendu au client est indiqué dans la donnée `a_retour_monnaie` de type `xsd:decimal`.

3.3 Peuplement de l'ontologie par l'extraction d'informations

Une fois le modèle de l'ontologie créé, il s'agit de le peupler automatiquement avec les données de nos tickets de caisse, c'est-à-dire de créer les instances en extrayant les informations à partir des textes de nos documents.

3.3.1 Présentation de l'approche

Pour peupler notre ontologie, nous avons programmé un algorithme en python à base d'expressions régulières qui extraient les mots ou les expressions voulus et nous avons utilisé la librairie Owlready2, créée par Lamy (2017), afin d'implémenter le modèle et de créer dynamiquement les individus et leurs relations à partir de ces extractions. Owlready2 sert à ce que l'auteur appelle de la « programmation orientée ontologie » : là où la programmation orientée objet crée des classes, la programmation orientée ontologie crée des concepts et des propriétés. Si cette librairie a été conçue pour des besoins dans le domaine du biomédical, elle est particulièrement utile pour représenter des informations aussi variées que celles de nos documents, et notamment pour créer et manipuler les concepts et les individus de façon dynamique. En effet, là où la plupart des outils permettent de créer et de modifier les individus, Owlready2 permet également de créer et modifier la structure de l'ontologie en manipulant les entités grâce à toute la puissance de la programmation et du calcul autorisés par Python.

Nous avons choisi de réaliser l'intégralité de notre processus en Python, afin de le rendre plus universel, plus intégrable à d'autres applications et à notre processus global, et moins dépendant des plateformes ou logiciels d'extraction d'information et d'édition d'ontologie que nous avons présentés en section 3.1. De plus, la nature de nos documents rend inutiles l'utilisation des ressources lexicales et des outils syntaxiques tels que des étiquetages morpho-syntaxiques, puisqu'ils ne contiennent pas, ou très peu, de phrases et que les syntagmes nominaux et expressions temporelles ou spatiales qui les composent contiennent beaucoup de noms propres et d'abréviations.

Nous créons une ontologie unique pour tous les documents contenus dans un seul dossier. Toute ontologie est éditable et des instances peuvent être ajoutées à tout moment dans l'ontologie. Par conséquent, le processus peut être dynamique et prendre de nouveaux documents en entrée à tout moment.

Afin de jongler entre les majuscules et les minuscules, toutes les expressions régulières cherchant à relever des lettres commencent par `(?i)`, ce qui permet de ne pas prendre en compte la casse. Par ailleurs, le format OWL/RDF impose des règles quant à l'écriture des différentes entités. Pour cette raison, toutes les espaces des instances ont été transformées en tirets bas (_) ainsi que les apostrophes ou guillemets simples. De même, de nombreux caractères ont été supprimés, comme les symboles €, %, les accolades ou les chevrons.

Pour chaque document texte du dossier en entrée, nous donnons son nom (qui est en fait le numéro attribué à chaque document sous ses différentes formes – image, texte brut, texte corrigé – depuis qu'il a été numérisé) à une instance de la classe `TicketCaisse`.

Nous relevons dans le ticket plusieurs types d'informations : celles qui font intervenir un prix et celles qui concernent l'entreprise.

3.3.2 Les informations de prix

Nous lisons le fichier une première fois ligne par ligne pour en extraire celles qui se terminent par un prix, c'est-à-dire, généralement, un nombre décimal suivi, ou non, d'un symbole € ou équivalent en lettres. Parfois, le symbole € tient lieu de séparateur

du nombre décimal, à la place de la virgule ou du point. Tous les prix, lorsqu'ils sont enregistrés dans l'ontologie, le sont sous forme de `float`, c'est-à-dire de nombre décimal. Il faut donc normaliser les nombres en transformant les séparateurs virgule ou symbole € en point.

Sur les tickets de Carrefour city, qui représentent un peu moins de la moitié de notre corpus, seuls trois types de lignes finissent par des prix : les lignes concernant le détail des produits, la ligne concernant le montant total des achats et enfin la ligne concernant le paiement. Sur d'autres tickets, il peut y avoir beaucoup de lignes qui finissent par des prix, comme en témoigne le tableau 3.3, présentant une image d'un ticket de caisse ayant quinze lignes finissant par un prix, dont sept qui n'entrent pas dans les trois catégories précédemment citées. En effet, dans ce document, deux lignes présentent des taux de TVA particuliers de 5,5% s'appliquant à certains types de produits et de 10% s'appliquant aux ventes à emporter, quatre lignes donnent des informations sur les montants obtenus et enregistrés sur la carte de fidélité du client et une ligne rappelle le montant des achats déclarés sur le ticket.

Afin d'extraire *tous* les noms de produit, nous avons créé une expression régulière relevant toutes les lignes qui finissent par un prix, en considérant que tout ce qui précède le prix dans la ligne est un produit. Les noms de produits peuvent contenir toutes sortes de caractères, comme nous le verrons plus en détail dans le chapitre 4. Afin d'extraire *uniquement* les noms de produits, nous avons créé dans la même lecture du document les expressions régulières pour relever le montant total et le ou les paiement(s) effectué(s), ainsi que pour relever des exceptions. Nous avons posé des conditions dans un certain ordre pour d'abord écarter les exceptions, puis pour compléter les propriétés de type données `a_nombre_produit` et `a_montant_total` du concept `TicketCaisse`, peupler le concept `PaiementIntermediaire` et ses propriétés, et enfin pour créer et/ou peupler les sous-concepts de la classe `Produit`.

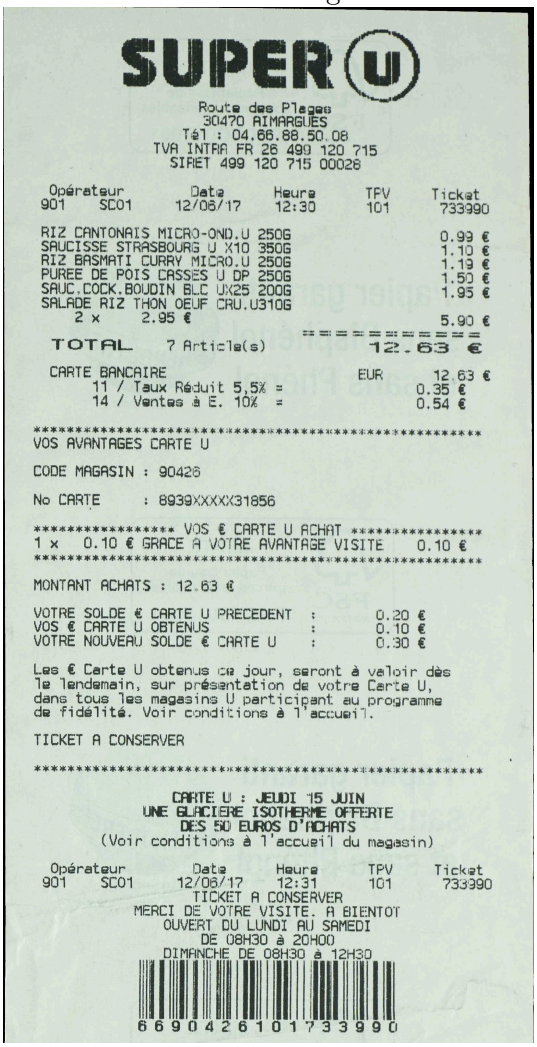
Les exceptions

Les exceptions relevées sont constituées à partir de mots qui ont peu de chance d'apparaître dans les noms de produit, comme « remise », « acquis », « solde », « avantage », « tva », « votre », « vos », « prix », « sous-total »... Nous avons fait le choix de ne pas relever les informations qui concernent les cartes de fidélité et les promotions car elles sont exprimées de façon trop différentes selon la provenance du ticket et n'apportent pas beaucoup d'indices pour la détection des fraudes.

Le total

Les lignes indiquant le total contiennent parfois le nombre d'articles, indiqué par un ou deux chiffres (nous avons toutefois mis une limite à trois chiffres, pour les très grosses courses, en supposant qu'il n'était pas raisonnable d'acheter plus d'un millier d'articles en une seule fois) suivi du mot article (au singulier, au pluriel, avec un `s` entre parenthèse, ou abrégé, avec ou sans point). Lorsque cette partie de l'expression régulière est trouvée,

Tableau 3.3 – Lignes relevant du même pattern INTITULE+PRIX



SUPER U

Route des Plages
30470 AIMARGUES
Tél : 04.66.88.50.08
TVA INTRA FR 26 499 120 715
SIRET 499 120 715 00026

Opérateur	Date	Heure	TPV	Ticket
901 SC01	12/06/17	12:30	101	733990

RIZ CANTONAIS MICRO-OND.U 250G	0.99 €
SAUCISSE STRASBOURG U X10 350G	1.10 €
RIZ BASMATI CURRY MICRO.U 250G	1.19 €
PUREE DE POIS CASSES U DP 250G	1.50 €
SAUC.COCK.BOUDIN BLC UX25 200G	1.95 €
SALADE RIZ THON OEUF CRU.U310G	
2 x 2.95 €	5.90 €
TOTAL 7 Article(s)	12.63 €

CARTE BANCAIRE	EUR	12.63 €
11 / Taux Réduit 5,5% =		0.35 €
14 / Ventes à E. 10% =		0.54 €

****
VOS AVANTAGES CARTE U

CODE MAGASIN : 90426

No CARTE : 8939XXXXX31856

********** VOS € CARTE U ACHAT **********
1 x 0.10 € GRACE A VOTRE AVANTAGE VISITE 0.10 €
****

MONTANT ACHATS : 12.63 €

VOTRE SOLDE € CARTE U PRECEDENT	:	0.20 €
VOS € CARTE U OBTENUS	:	0.10 €
VOTRE NOUVEAU SOLDE € CARTE U	:	0.30 €

Les € Carte U obtenus ce jour, seront à valoir dès le lendemain, sur présentation de votre Carte U, dans tous les magasins U participant au programme de fidélité. Voir conditions à l'accueil.

TICKET A CONSERVER

****

CARTE U : JEUDI 15 JUIN
UNE GLACIERE ISOTHERME OFFERTE
DES 50 EUROS D'ACHATS
(Voir conditions à l'accueil du magasin)

Opérateur	Date	Heure	TPV	Ticket
901 SC01	12/06/17	12:31	101	733990

TICKET A CONSERVER
MERCI DE VOTRE VISITE. A BIENTOT
OUVERT DU LUNDI AU SAMEDI
DE 08H30 à 20H00
DIMANCHE DE 08H30 à 12H30

6690426101733990

```
RIZ CANTONAIS MICRO-OND.U 250G 0.99 €
SAUCISSE STRASBOURG U XI0 350G 1.10 €
RIZ BASMATI CURRY MICRO.U 250G 1.19 €
PUREE DE POIS CASSES U DP 250G 1.50 €
SAUC.COCK.BOUDIN BLC UX25 200G 1.95 €
SALADE RIZ THON OEUF CRU.U310G
2 x 2.95 € 5.90 €

TOTAL 7 Articles(s) 12.63 €

CARTE BANCAIRE EUR 12.63 €
11 / Taux Réduit 5.5% 0.35 €
14 / Ventes à E. 10% 0.54 €

1 x 0.10 € GRACE A VOTRE AVANTAGE
VISITE 0.10 €

MONTANT ACHATS : 12.63 €
VOTRE SOLDE € CARTE U PRECEDENT :
0.20 €
VOS € CARTE U OBTENUS : 0.10 €
VOTRE NOUVEAU SOLDE € CARTE U : 0.30
€
```

le groupe qui relève le nombre est transformé en entier et vient compléter la propriété `a_nombre_produit` de l'instance du ticket de caisse en cours de traitement.

La même expression régulière sert également à relever la propriété `a_montant_total`, car les deux indications sont, la plupart du temps dans notre corpus, sur la même ligne de texte. Le montant en lui-même, un prix, est repéré par l'expression régulière suivante : `[0-9]+[.,][0-9]2`. Ce patron prend un ou plusieurs chiffre(s), suivi d'un point ou d'une virgule, puis de deux chiffres. Ce prix doit être précédé d'un mot qui signifie que c'est le montant total, comme « total », « montant », « net TTC »... Le prix est souvent, mais pas toujours, suivi d'un signifiant qui exprime la monnaie utilisée, la plupart du temps par le symbole €, mais parfois aussi par « euros », « EUR » ou encore « E ».

Les paiements

Après avoir éliminé les lignes finissant par un prix qui sont des exceptions et des lignes qui concernent le montant total du ticket de caisse, nous cherchons à peupler le concept `PaiementIntermediaire`. Sur une ligne de paiement, on trouve plusieurs éléments : le montant du paiement, le moyen de paiement et/ou des mots signifiant de quoi il s'agit : « paiement », « encaissement », « règlement »... Ces mots, comme pour les lignes qui concernent le total, peuvent être considérés comme des labels, des intitulés ou des annotations. Dans un cadre d'extraction d'informations moins supervisée, sur des documents plus variés, il pourrait être intéressant de prendre ces mots comme nom de concept, plutôt que de prédéfinir un concept figé, et établir un lien `equivalent_to` entre les différents concepts qui relèvent de la même réalité.

Nous avons recensé de nombreux moyens de paiement dans nos tickets : chèque, espèces, tickets restaurant, paiement différé (bon de commande) ou carte bancaire. Nous n'avons pas trouvé, dans notre corpus, de paiement par smartphone, peut-être parce qu'en 2016 et 2017, dates de collecte des documents, ce moyen de paiement était encore rare. Les commerçants et logiciels de caisse sont parfois inventifs quant aux dénominations de ces moyens de paiement. Outre les pluriels, les majuscules ou minuscules, les absences ou présences d'accents, qui rendent l'extraction parfois complexe, on trouve également de nombreuses variantes lexicales pour un même moyen de paiement. Par exemple, la carte bancaire peut-être abrégée en « CB » (ou en « C.B. »), mais peut également être accompagnée de la précision « EMV » et/ou de « sans contact », ou être appelée « carte bleue », « carte (de) crédit », « visa », « mastercard »... C'est pourquoi notre patron prend le mot « carte » suivi de n'importe quelle suite de caractères, en plus des autres moyens de paiement listés plus haut. Compte-tenu de la diversité du lexique, nous avons décidé de ne pas faire une propriété de type liste fermée de chaînes de caractères dans notre ontologie, mais de prendre la chaîne de caractères extraite telle quelle pour remplir la propriété `a_moyen_paiement`. Le montant du paiement s'exprime comme les autres prix, le plus souvent en fin de ligne, mais peut parfois être situé avant le moyen de paiement. Les deux cas sont pris en compte dans notre algorithme, afin de peupler la propriété `a_montant`.

Un cas particulier de l'extraction des informations de paiement est également pris en compte dans notre approche : le cas des espèces. Il est fréquent que le montant associé

à un paiement en espèces, c'est-à-dire inscrit sur la ligne de paiement, soit le montant donné au commerçant par le client. Si ce montant est plus élevé que le total à payer, le rendu de monnaie est alors généralement écrit sur la ligne suivante ou plus loin dans le document. Toutefois, l'accès à deux lignes consécutives ne peut pas s'effectuer dans une lecture ligne par ligne. Nous procédons donc à une deuxième lecture du document, qui prend l'ensemble du texte dans une seule chaîne de caractères, afin d'affecter à la propriété `a_retour_monnaie` le montant que rend le commerçant au client pour arriver au montant dû.

Une fois le ou les paiements intermédiaires extraits pour le ticket analysé, nous affectons à la propriété `a_paiement_total` du concept `TicketCaisse` soit la différence du montant et du rendu de monnaie, arrondie à deux chiffres après la virgule dans le cas d'un paiement en espèces, soit la somme des autres paiements intermédiaires du ticket.

Les produits

Il nous reste ainsi, dans la catégorie des lignes finissant par un prix, les noms de produits, ainsi que toutes les informations qui concernent les produits, c'est-à-dire la quantité, le prix à l'unité, le poids, le prix au kilogramme et le prix total. Le nom de produit est constitué de tout ce qui précède le prix, à l'exception de la quantité et du prix à l'unité. Pour créer le sous-concept de Produit qui regroupe les différents exemplaires d'un même produit, nous ne prenons en compte que les caractères alphabétiques de ce nom. Pour créer les instances en revanche, nous accolons ce nom, dans lequel les espaces sont transformés en tirets bas, à la lettre R (pour « reçu ») et le numéro du document, comme illustré dans la figure 3.3. Cela nous permet de différencier un même produit acheté lors de courses différentes, car le prix n'est peut-être pas le même, ni la quantité ou le poids.

Une fois ces instances créées, nous pouvons les relier au ticket grâce à la relation `concerne_achat` et son inverse `est_inscrit_sur`.

La quantité et le prix à l'unité, s'il y a un achat en plusieurs exemplaires d'un même produit, sont écrits sur les tickets de Carrefour city entre le nom du produit et le prix total, comme sur la deuxième ligne de produits de l'image 3.6. Pour les autres émetteurs de tickets, il peut également être écrit sur la ligne suivante, comme on peut l'observer sur le ticket du tableau 3.3. Nous extrayons donc ces informations, si elles existent, afin de peupler les propriétés `a_quantite` et `a_prix_unitaire`. S'il n'y a pas l'information, ou qu'elle est sous une autre forme, la quantité est par défaut de 1 et le prix unitaire est le prix total.

Le fait qu'une information concernant un même produit se situe sur deux lignes différentes nous oblige là encore à reparcourir le document comme une seule chaîne de caractères (méthode `read()`) ou comme une liste de chaînes de caractères (`readlines()`). Cette deuxième fonction nous permet de détecter les lignes constituées d'un poids (un nombre décimal à trois ou quatre chiffres après la virgule), de la lettre « x » puis d'un prix (nombre décimal à deux chiffres après la virgule suivi de la suite de caractères « €/kg ») et de compléter ainsi les propriétés de type données facultatives `a_poids` et `a_prix_kg`

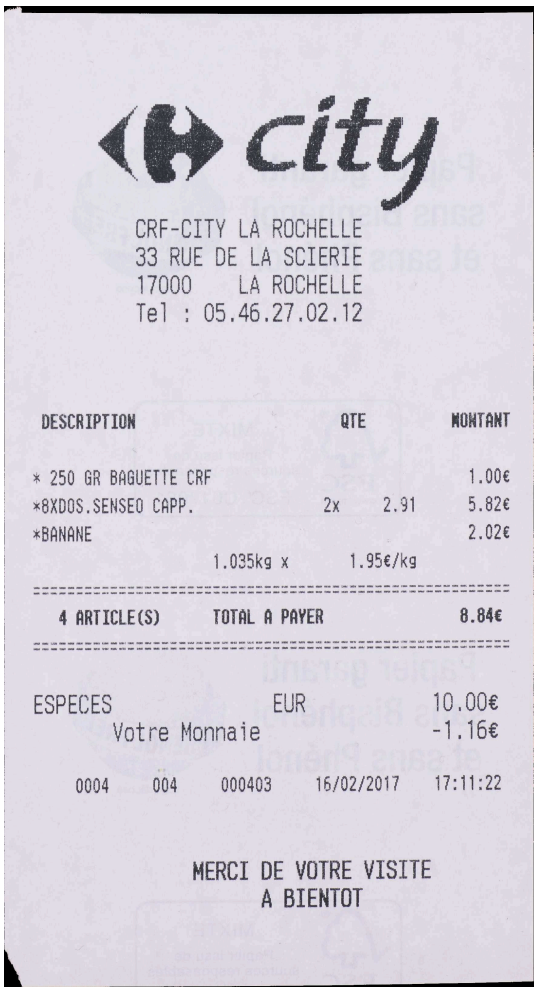


FIGURE 3.6 – Exemple de ticket de caisse concernant l'achat de produits en diverses quantités et avec un paiement en espèces

au produit inscrit sur la ligne précédente, c'est-à-dire de l'élément de liste précédent, déjà enregistré en tant qu'entité de l'ontologie.

Les expansions

Pour chaque type de produit, c'est-à-dire pour chaque sous-concept de `Produit`, on recherche l'expansion possible grâce à l'algorithme que nous détaillerons dans le chapitre 4. Les instances des produits sont alors reliées à leur possible expansion grâce à la relation `a_expansion`, et la relation inverse `est_expansion_de` relie un nom de produit non-abrégé à un ou plusieurs produits.

3.3.3 Les informations sur le document et l'entreprise

Les informations concernant les échanges commerciaux sont au cœur du ticket de caisse et sont les informations qu'on imagine facilement les plus sensibles et les plus sujettes à la fraude. Cependant, d'autres éléments du ticket de caisse peuvent être fraudés, comme les expressions de localisations temporelles et spatiales : une adresse peut être modifiée pour justifier un déplacement qui n'a pas réellement eu lieu, une date peut être altérée pour créer un alibi... Il faut donc extraire toutes ces informations également.

Les noms d'entreprise

L'entité nommée la plus complexe à extraire avec des expressions régulières dans un ticket de caisse est le nom de l'entreprise. En effet, nous n'avons pu déterminer aucune indication pour trouver un modèle générique de nom d'une entreprise. En effet celui-ci peut être composé d'un ou plusieurs mots, peut se situer sur une ou plusieurs lignes, au début ou à la fin du document, avoir une majuscule ou non... L'un des problèmes que nous avons remarqué sur notre corpus est que le nom de l'entreprise est régulièrement écrit en haut du document dans son logo, avec une police particulière. Cela a pour conséquences qu'il n'est pas systématiquement transcrit dans le texte. C'est par exemple le cas du U encerclé des magasins SUPER U, que les correcteurs ont parfois noté en majuscule, en minuscule, ou supprimé, ou laissé tel que l'OCR l'a reconnu, c'est-à-dire comme étant un ⓒ ou un ®.

Malgré ces difficultés, nous avons tout de même choisi de prendre la ou les premières lignes des tickets comme nom d'entreprise. Nous avons mis en place un certain nombre d'exceptions afin de limiter le nombre de faux positifs et de délimiter correctement le nom de l'entreprise. Ainsi, si la première ligne contient les mots ou abréviations « ticket » ou « TK », « ESP », « BP », « CB », « CS » ou « justificatif », nous passons directement à la deuxième ligne. Si cette deuxième ligne contient les expressions « à emporter », « sur place » ou « facture », on passe à la troisième ligne, qui est généralement plus satisfaisante. Si la première ligne ne contient pas d'éléments problématiques, nous la prenons comme nom d'entreprise, accompagnée de la deuxième ligne si cette dernière ne contient pas une adresse ou un numéro de téléphone.

Les immatriculations

Pour compléter les instances des sous-concepts d'immatriculation, nous cherchons assez simplement l'expression régulière qui correspond à la définition de chaque numéro, précédé de l'intitulé de ce numéro. Ainsi, pour le numéro de SIRET, dans une lecture ligne par ligne, nous cherchons l'ER `(?i)(siret) ?:? ?([0-9]{9})[-\- ]{0,3}([0-9]{5})` :

- `(?i)` : l'expression ne tient pas compte de la casse, *i.e.* « siret » peut être écrit en majuscule, en minuscule ou avec seulement une intiale en majuscule ;
- `(siret) ?:? ?` : le mot siret (obligatoire) peut être suivi ou non d'une ou de deux espaces, avec un deux-points facultatif intercalé ;
- `([0-9]{9})[-\- ]{0,3}` : exactement neuf chiffres, suivis de zéro à trois caractères présents entre les crochets, à savoir une espace, un trait d'union ou un tiret

moyen. Le trait d'union, qui est souvent réalisé par le même caractère que le signe « moins », également appelé « tiret court » ou « tiret quart de cadratin », sert à séparer les composantes d'un mot composé, par exemple. Ce n'est pas le même signe typographique que le « demi-tiret », ou « tiret demi-cadratin », qui sert, par exemple, à séparer un élément incident d'une phrase. La barre oblique inversée sert à déspécifier le caractère qui le suit, pour que le trait d'union ne soit pas compris comme devant servir d'intervalle mais bien pour qu'il soit pris comme caractère à trouver.

— `([0-9]{5})` : exactement cinq chiffres.

L'écriture du numéro de SIRET est en effet parfois décomposé pour mettre en évidence le numéro de SIREN, qui correspond au troisième élément de notre expression régulière.

Les expressions régulières qui permettent d'extraire le RCS et la TVA Intracommunautaire sont sensiblement similaires : le mot « siret » est remplacé par « rcs » ou par « tva intracommunautaire », toutes les lettres à partir du « c » et l'espace étant facultatives, pour repérer les formes abrégées observées dans le corpus. Pour le RCS, nous cherchons ensuite une suite de lettres, afin de relever n'importe quel nom de ville, puis le numéro de SIREN, composé de neuf chiffres. Pour la TVA Intracommunautaire, on cherche les lettres FR, suivies de deux chiffres (la clé), suivis de de neuf chiffres (le SIREN).

Comme tous ces numéros ont en commun le numéro de SIREN, nous l'avons choisi comme seule immatriculation d'une entreprise. Nous attribuons donc systématiquement comme instance de `SIREN` le groupe des patrons des trois autres entités correspondant. Nous cherchons également le numéro de SIREN avec une ER si aucun des trois autres numéros n'est relevé. Le numéro de SIREN étant régulièrement confondu avec le numéro de SIRET, il est fréquent que l'indication SIRET soit suivie de seulement 9 chiffres. Le patron relevant le SIREN a donc été assoupli, permettant l'extraction d'un « t » à la place du « n », et donc à la fois des mots « siret » et « siren ». Les relations `a_immatriculation` et `est_immatriculation_de` sont créées entre le nom de l'entreprise et le numéro de SIREN.

Les coordonnées

Après avoir relevé les numéros d'immatriculation, nous relevons les numéros de fax puis de téléphone, dans la même lecture ligne par ligne, ce qui permet grâce à des `elif` (sinon si) d'éliminer les lignes contenant des suites de chiffres au fur et à mesure, et ainsi limiter les faux positifs dans l'extraction des numéros de téléphone. Nous cherchons donc d'abord les numéros de fax, pour lesquels l'indication « fax » est toujours signalée, avec ou sans point, avec ou sans deux-points, avec ou sans espace. Le numéro en lui-même commence par un 0 puis par un chiffre de 1 à 9, puis par 8 chiffres séparés ou non par des espaces, des points, des barres obliques ou des tirets en groupes de deux. Cela nous permet de peupler le concept `NumeroFax` et les propriétés `a_fax` et `est_fax_de`.

Le numéro de téléphone est enfin extrait, après tous les autres numéros, car nous avons observé que l'indication « téléphone » (ou équivalent) n'est pas toujours présente sur les tickets, comme s'il était évident qu'un numéro en 10 chiffres, séparés par groupe

de deux, commençant par un zéro, situé à proximité d'une adresse, était un numéro de téléphone. Le mot « téléphone », avec ou sans accents, abrégé ou non en « tél », est donc facultatif. Seule la séquence de chiffres, identique à celle du numéro de fax, est relevée, et vient peupler le concept `NumeroTelephone`, qui est relié par les propriétés `a_telephone` et `est_telephone_de` à l'entreprise.

L'adresse mail est relevée grâce à la présence systématique d'un élément : le symbole @. Notre expression régulière relève donc deux chaînes de caractères ne contenant que des lettres, des chiffres, des tirets-bas, des traits d'union et des points, entrecoupées d'un symbole @. Lorsque cette adresse est indiquée sur le ticket et qu'elle est relevée correctement, elle vient peupler le concept `AdresseMail`, qui est relié à l'entreprise par les propriétés `a_adresse_mail` et `est_adresse_mail_de`.

L'adresse de site web contient également un élément significatif : les trois « w » qui marquent le début du nom de domaine. Les adresses mails ne sont pas identifiées sur nos tickets par des URL contenant le protocole (http par exemple), mais simplement par le nom de domaine dans le World Wide Web. L'expression régulière relève donc « www. » suivi de n'importe quelle chaine de caractères qui a les mêmes caractéristiques que celles des adresses mails et qui finit par « .com », « .fr » ou « .net ». Cette chaîne constitue l'instance de `SiteWeb`, reliée à son entreprise par `a_site_web` et `est_site_web_de`.

L'adresse postale est quant à elle constituée de trois éléments importants que sont l'intitulé, le code postal et le nom de la ville. Nous avons fait le choix de n'extraire que les adresses complètes, c'est-à-dire les adresses qui contiennent ces trois éléments. Ainsi, nous ne relevons pas de codes postaux seuls ou de noms de villes seuls, ce qui nous évite d'utiliser des dictionnaires. L'adresse s'étale souvent sur plusieurs lignes, mais pas toujours. Nous avons donc choisi de relire le document comme une seule chaîne de caractères avec la fonction `read()`.

Nous extrayons ainsi une chaîne de caractères de taille variable commençant par un saut de ligne, puis pouvant contenir des lettres, des chiffres, des virgules, des traits d'union, des points, des apostrophes et des barres obliques. Cela vient du fait qu'un magasin ou un restaurant peut occuper plusieurs numéros d'une rue et ces numéros peuvent être écrits sous la forme « 24,26 », « 14-16 » ou « 3/5 » [3]. C'est le code postal, contenant quatre ou cinq chiffres, et précédé d'un retour à la ligne, d'une espace ou d'une séquence espace–tiret–espace, qui force la reconnaissance d'une adresse par ses caractères obligatoires et immuables. Le nom de la ville suit le code postal après une espace et est constitué des mêmes éléments que la première partie de l'adresse. En effet, les noms de villes peuvent être composés de plusieurs mots, comme « La Rochelle », et donc contenir des apostrophes, comme « Saint Jean d'Angély » ou « Le Château d'Oléron », ou des abréviations, comme « Moutiers/Chantemerle » pour « Moutiers-sous-Chantemerle ».

Ces éléments viennent peupler les concepts `CodePostal`, `Ville` et`AdresseComplete`, et les relations `a_code_postal`, `est_partie_de`, `a_adresse` et `est_adresse_de` sont créées.

3. Ces trois exemples sont tirés de notre corpus.

Les dates et heures

Afin de remplir les propriétés de type données `a_date` et `a_heure` des instances du concept `TicketCaisse`, nous cherchons les expressions de localisation temporelles de type heures et dates. Pour le heures, il s'agit d'extraire les heures, les minutes et le cas échéant les secondes, comme étant des entiers positifs, afin de correspondre au standard du format `xsd:time`. Les formats des heures sont relativement simples et homogènes : deux chiffres, signifiant les heures, sont suivis d'un deux-points ou de la lettre « h », puis deux autres chiffres suivent, signifiant les minutes. Les secondes ne sont pas toujours indiquées, mais quand elles le sont, ce sont deux chiffres suivant un point ou un deux-points. Par défaut, si les secondes ne sont pas indiquées lors du peuplement de la propriété, elles sont à zéro. Les chiffres, de type `string`, sont transformés en `int`, c'est-à-dire en entiers.

Pour les dates, le format `xsd:date` prend en entrée, dans cet ordre, l'année, le mois et le jour. Les noms des jours de la semaine ne nous intéressent donc pas ici, et nous avons fait le choix de ne pas entrer de dictionnaire des noms de mois, sous formes longues ou abrégées, car nous avons pensé que le format des tickets de caisse ne se prêtait pas à une expression littéraire, ou du moins littérale, de la date, mais plus à sa forme réduite et numérale. Par ailleurs, en français, la date s'exprime dans l'ordre jour – mois – année, à la différence du format de la propriété, qui suit le modèle anglo-saxon. Nous avons observé dans notre corpus que la date subissait parfois des abréviations : les zéros en début de jour et de mois sont parfois supprimés et les deux premiers chiffres de l'année (« 20 ») peuvent également l'être. Notre expression régulière prend donc deux séquences d'un ou deux chiffres séparées d'une barre oblique, d'un point ou d'un trait d'union, puis une troisième séquence de deux à quatre chiffres séparée de la précédente par les mêmes séparateurs. Si l'année n'est pas complète, nous rajoutons les chiffres « 2 » et « 0 » devant, et si le mois ou le jour commence par « 0 », nous le supprimons, avant de transformer toutes ces chaînes de caractères en entier.

3.4 Évaluation

Il est toujours difficile de savoir quand arrêter l'amélioration de l'extraction d'information et de juger s'il faut persévérer à prendre en compte les cas particuliers ou s'en tenir à un degré d'extraction « satisfaisant ». La satisfaction est également extrêmement compliquée à évaluer : si l'on s'en tient à l'objectif de notre thèse, il faudrait extraire parfaitement toutes les informations susceptibles d'influencer la détection des faux documents. Si l'on considère que l'extraction est déjà biaisée par le fait que nos documents contiennent des erreurs de transcriptions, le seuil de satisfaction peut légitimement être plus faible.

Nous présentons dans cette section une évaluation quantitative réalisée sur un petit échantillon de tickets, puis nous détaillons les résultats obtenus dans une évaluation qualitative plus approfondie. Nous conclurons cette section par un survol des informations acquises sur notre corpus, qui viennent approfondir la présentation qui en a été faite dans le chapitre 2, et qui dressent un panorama des possibilités d'extraction sur un tel corpus.

3.4.1 Évaluation quantitative

Pour évaluer correctement une extraction d'information, il faudrait la confronter à une vérité terrain, c'est-à-dire à ce qui aurait dû être extrait. Constituer une telle base manuellement, sur l'ensemble des tickets, serait extrêmement chronophage. Nous avons donc choisi de réaliser cette vérité terrain sur une petite partie des tickets à partir des informations extraites automatiquement. Pour cela, il faut vérifier que les informations extraites sont correctes et complètes, ajouter les informations manquantes et corriger les informations incorrectes, afin de pouvoir comparer ces quelques corrections aux résultats d'extraction. Il s'agit donc plus ici de se donner une idée de la qualité de notre extraction d'information, plutôt que d'avoir un score précis et détaillé sur l'ensemble des informations des documents.

Nous avons fait évaluer les résultats d'extraction sur 100 tickets (10% du corpus utilisé pour la détection des faux documents). Les tickets évalués ont été partagés en deux : les tickets venant des magasins de l'enseigne Carrefour d'une part et les tickets des autres enseignes d'autre part. En effet, la plupart de nos règles d'extraction ont été apprises sur les tickets provenant de Carrefour, car la structure des documents est sensiblement la même. En revanche, la structure des informations des autres tickets est très variable et il est très difficile d'observer et d'implémenter toutes les possibilités d'extraction. Nous avons donc jugé utile de séparer ces deux provenances. Le choix des tickets à évaluer a été fait de manière aléatoire, sachant que la moitié des tickets utilisés proviennent de magasins de l'enseigne Carrefour.

L'extraction d'information est difficile à évaluer objectivement : il ne s'agit pas d'un problème de classification classique où le résultat est binaire, soit vrai, soit faux. Maynard et al. (2006) présentent ces difficultés ainsi que plusieurs façons d'évaluer l'extraction d'information basée sur l'ontologie. L'un des problèmes majeurs de l'évaluation d'information est de juger si une information extraite est correcte ou non, complète et précise, c'est-à-dire avec tous les caractères ou mots qui doivent être extraits et sans caractères ou mots superflus. C'est pourquoi nous avons demandé à l'évaluateur de classer les informations de chaque ticket comme étant :

— VP : information bien extraite
— VN : information non extraite car non présente sur le document
— FN : information non extraite alors qu'elle aurait dû
— FP : mauvaise information extraite
— Partiel : information extraite correcte mais mal délimitée

Les métriques classiques de précision, rappel et *accuracy* doivent donc être adaptées afin de prendre en compte cette particularité. Plusieurs possibilités s'offrent à nous : calculer deux fois la précision et le rappel selon si l'on considère les extractions « un peu vraies », ou partielles, comme vraies ou comme fausses, ou calculer ces métriques en comptant pour moitié les extractions partielles, selon les formules suivantes :

$$\text{Précision} = \frac{\text{Corrects (VP)} + \frac{1}{2}\text{Partiels}}{\text{Corrects (VP)} + \text{Indus (FP)} + \text{Partiels}}$$

$$\text{Rappel} = \frac{\text{Corrects (VP)} + \frac{1}{2}\text{Partiels}}{\text{Corrects (VP)} + \text{Oubliés (FN)} + \text{Partiels}}$$

$$\text{Accuracy} = \frac{\text{VP} + \text{VN} + \frac{1}{2}\text{Partiels}}{\text{VP} + \text{VN} + \text{FP} + \text{FN} + \text{Partiels}}$$

Les tableaux 3.4 et 3.5 présentent la précision, le rappel, la F-mesure et l'*accuracy* pour chaque information relevée dans un échantillon de 50 tickets aléatoires provenant respectivement de magasins de l'enseigne Carrefour et des autres magasins.

Tableau 3.4 – Évaluation de l'extraction d'information par information sur les tickets de l'enseigne Carrefour

Information	Précision	Rappel	F-mesure	*Accuracy*
Date	1.00	1.00	1.00	1.00
Heure	1.00	1.00	1.00	1.00
Entreprise	0.80	0.80	0.80	0.80
Adresse	1.00	1.00	1.00	1.00
Téléphone	0.58	1.00	0.73	0.58
SIREN	n/a	n/a	n/a	1.00
Nombre d'articles	1.00	0.96	0.98	0.96
Total	1.00	0.98	0.99	0.98
Paiement	1.00	0.96	0.98	0.96
Produits	0.99	0.96	0.98	0.96

Tableau 3.5 – Évaluation de l'extraction d'information par information sur les tickets de provenance diverse

Information	Précision	Rappel	F-mesure	*Accuracy*
Date	0.92	0.73	0.81	0.70
Heure	0.94	0.83	0.88	0.82
Entreprise	0.82	0.89	0.85	0.82
Adresse	0.95	0.72	0.82	0.77
Téléphone	0.92	0.87	0.89	0.84
SIREN	1.00	0.06	0.11	0.66
Nombre d'articles	1.00	0.06	0.12	0.70
Total	0.77	0.83	0.80	0.66
Paiement	0.94	0.53	0.68	0.68
Produit	0.92	0.73	0.81	0.70

Nous pouvons observer que les résultats d'extraction sont plutôt bons dans l'ensemble, surtout pour les tickets Carrefour où seuls les noms d'entreprise et les numéros

de téléphone ont des f-mesure et *accuracy* inférieures à 0.96. Pour l'ensemble des autres tickets, on observe dans l'ensemble une précision élevée, ce qui signifie que ce qui est extrait est généralement correct, mais le rappel l'est moins, ce qui signifie que beaucoup d'informations ne sont pas relevées, ce qui est particulièrement vrai pour le numéro de SIREN et le nombre d'articles. Par ailleurs, nous avons constaté que pour 110 tickets sur les 517 non-Carrefour, notre système n'a pas extrait de produits. Cela peut être dû à une absence d'intitulé de produits, ce qui est souvent le cas de tickets provenant du marché, ou à une trop grande restriction de nos règles.

Il est important de noter que l'évaluateur a compté comme Vrai Négatif les informations n'ayant pas été relevées pour cause d'erreurs OCR, puisque ce n'est pas un problème lié à notre processus d'extraction, mais un problème amont. Par conséquent, nous ne voyons pas qu'environ 3% des informations sont erronées et ne sont donc soit pas relevées (mais classées comme VN) soit relevées mais fausses (mais classées comme VP).

3.4.2 Évaluation qualitative

La difficulté de l'extraction d'information par expressions régulières réside dans le fait que chaque changement dans une règle, c'est-à-dire un ensemble de conditions et d'expressions régulières pour extraire une information, peut affecter l'ensemble des extractions, et changer du tout au tout la précision et le rappel des informations extraites. Chaque règle peut être très généraliste et relever ainsi de nombreux faux positifs : c'est le cas par exemple d'une règle qui relèverait toutes les lignes finissant par un prix et considèrerait que ce qui précède le prix est un nom de produit, alors qu'il existe de nombreuses exceptions à cette règle, puisque les lignes de montants totaux, de paiement, de TVA, etc. peuvent également finir par un prix comme nous pouvons le voir dans la figure 3.3. De même, une règle peut être trop précise (afin d'éviter les faux positifs), et passer à côté de certaines expressions qui n'apparaissent que peu dans le corpus ou qui sont trop éloignées des patterns habituels, ce qui augmente le nombre de vrais négatifs et diminue ainsi le rappel. C'est le cas par exemple de nos règles d'extraction sur les numéros de SIREN, SIRET, TVA Intracommunautaire et RCS qui ne tolèrent pas les espaces ou les points au sein de la suite de neuf chiffres du SIREN, ce qui en fait une règle trop précise. De même, l'information sur le nombre d'articles était essentiellement relevée pour les tickets provenant des magasins Carrefour, pour lesquels l'information était indiquée sur la ligne du montant total juste avant celui-ci et était précédée de l'indication « Article » ou équivalent. Or dans de nombreux cas, l'information sur le nombre d'articles est indiquée sur une autre ligne, ou n'est pas précédée de l'indication.

En ce qui concerne les numéros de téléphones des magasins Carrefour, l'erreur est double. Tout d'abord il y a une erreur d'extraction, qui fait que c'est la dernière suite de 10 chiffres commençant par un « 0 » qui est prise en compte comme numéro de téléphone lors de la lecture ligne par ligne d'un ticket, ce qui pose problème quand il y a un numéro de carte de fidélité par exemple, qui est situé vers la fin du document. Ensuite, lors de la création du tableur donné à l'évaluateur, nous n'avons donné qu'une seule instance de `NumeroTelephone` reliée au concept `Entreprise`, parmi la liste des différentes valeurs

extraites, qui ne correspond donc pas à la valeur extraite sur le ticket en question et qui est donc la même pour tous les tickets émis par cette entreprise. Ainsi, la même mauvaise information est indiquée pour 11 tickets émis par Carrefour market évalués, sans que l'information ne soit présente sur aucun. Cela est vraiment problématique dans notre modèle car nous n'avons pas le moyen de savoir quel numéro de téléphone est extrait sur chaque ticket, vu que le concept `NumeroTelephone` n'est pas relié au `TicketCaisse`. Il suffit donc d'une seule erreur, ou d'une seule fraude, sur un seul ticket, pour impacter l'évaluation de tous les tickets d'une même entreprise.

Le (relativement) faible score d'extraction des noms d'entreprise est dû principalement à la difficulté de délimiter ce qui fait partie du nom et ce qui n'en fait pas partie. Ainsi, de nombreux noms ont été classés comme « partiels » sans que ce soit véritablement faux. Par exemple, sur les 50 reçus évalués de Carrefour, 20 ont été classés comme « partiels » alors que 17 d'entre eux indiquent tout de même « Carrefour city » ou « Carrefour market », mais ne précise pas la ville (« La Rochelle » ou « Aytré »), comme d'autres peuvent le faire. La pertinence de la délimitation est donc laissée au libre arbitre de l'évaluateur, et la mesure la plus juste était de les compter comme partiellement bien extraits. De même, certains noms extraits sont trop longs, comptent trop de mots ou de lettres. C'est le cas des tickets qui commencent par « Bienvenue à ... » ou qui inscrivent le slogan à la suite du nom de l'entreprise.

Informations mal extraites

Certaines difficultés rencontrées sont dues à des erreurs d'impression, sans doute liées à des erreurs d'encodage des logiciels d'impression de ticket de caisse. Par exemple, sur un ticket de jardinerie de notre corpus, nous trouvons deux fois le symbole `{` à la place de la lettre `é`, dans les mots « tél » et « azalée » (orthographiés par conséquent `t{l` et `azal{e`). Cela est particulièrement gênant dans notre système car les entités XML, ou les concepts OWL, ne doivent contenir que des caractères alphanumériques et quelques autres caractères.

Un cas particulier qui pose également problème est le style d'écriture utilisé par les marchands sur certains tickets de caisse : pour mettre en évidence le montant total à payer, ils n'hésitent pas à écrire `T O T A L` avec des espaces entre chaque caractère. Cela nous oblige à rajouter une possibilité dans la règle d'extraction des totaux, sans quoi cela crée un produit `TOTAL`, qui est aussi le nom d'une entreprise relevée dans notre corpus. Or OWLREADY2 attribue automatiquement un même nom aux différents concepts qui lui sont rattachés. Ainsi `TOTAL` serait Entreprise $\cap$ Produit, ce qui ne correspond pas à la réalité : avoir le même nom ne veut pas dire être le même concept, le même objet, la même réalité. Si cette exception peut-être réglée aisément en modifiant une règle d'extraction, d'autres exceptions du même type sont plus délicates à lever : `BOUCHERIE` est ainsi à la fois un nom d'entreprise, certes incomplet, et un nom de produit, certes pas très explicite.

Le choix de la présentation des informations influe également beaucoup sur notre capacité à extraire l'information. La figure 3.7 par exemple montre un tableau (sans lignes) qui détaille les montants hors taxes et toutes taxes comprises sur lesquels s'appliquent les différents taux de TVA (Taxes sur la Valeur Ajoutée). En effet, il existe quatre taux

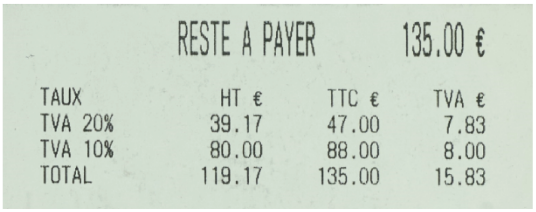


FIGURE 3.7 – Disposition en tableau, difficile à traiter avec nos règles d'extraction

de TVA différents en fonction de la nature du produit acheté. Les commerçants doivent donc prendre en compte ces différents taux dans leurs calculs, mais ne l'indiquent pas toujours sur les tickets de caisse qu'ils émettent. Dans l'exemple de la figure 3.7, le montant total du ticket de caisse est indiqué deux fois : à la suite de « reste à payer » et dans la troisième colonne de la quatrième ligne du tableau, à la suite de l'indication « total ». Notre règle prend en compte la dernière ligne qu'il trouve pouvant correspondre à un total, c'est-à-dire ici la dernière ligne, et le dernier montant de la ligne, c'est-à-dire ici 15.83.

Une autre entité difficile à extraire est la date, parfois écrite en toutes lettres, parfois avec les noms des jours et des mois abrégés, ou en anglais. Nous avons même trouvé un format de date très original, sur les tickets venant d'un restaurant proche du laboratoire. La date y est écrite dans un format très particulier : les deux chiffres du jour, suivi sans espace des trois premières lettres du mois dans sa version anglaise, une apostrophe, puis les deux derniers chiffres de l'année : « 19Jul'16 ». Nos règles ne permettent pas d'extraire ces dates.

Les tickets de caisse peuvent également contenir des informations erronées ou incomplètes, ce qui empêche l'extraction. Par exemple, l'un des tickets évalués indique un numéro de SIREN à seulement 8 chiffres, ce qui n'est pas possible. L'information n'est donc pas extraite, et a été évaluée comme Vrai Négatif.

Par ailleurs, certains tickets présentent également une mesure à laquelle nous n'avions pas pensé : la superficie et le prix au mètre carré. C'est le cas par exemple du ticket 1236, venant d'un magasin de bricolage, qui concerne l'achat du produit « MDF EP 6MM » pour « 0,12 X10,71 EUR », c'est-à-dire, un panneau de fibres à densité moyenne (*Medium Density Fiberboard*) autrement appelé « medium » d'épaisseur 6 millimètres et de taille 0.12 mètre carré, sachant que ça coûte 10.71 euros au mètre carré. Ces deux dernières informations, qui n'ont pas été relevées, ont été classées par l'évaluateur comme devant être relevées dans les propriétés de type données `a_poids` et `a_prix_kg`.

Tous ces exemples viennent souligner la difficulté d'extraire toutes les informations pertinentes des tickets de caisse, car celles-ci sont très variées et se présentent sous diverses formes et à divers emplacements. Si, au premier coup d'œil, nous avons l'impression que les tickets de caisse se ressemblent tous, il n'en est rien : notre corpus, qui n'est pourtant pas si grand, le prouve bien en ne permettant pas de dresser un modèle générique haut-niveau pour extraire les informations. Ce sont donc dans les détails bas-niveaux, à l'échelle du caractère, qu'il a fallu se situer.

Informations non inscrites

Certaines informations sont plutôt rares sur les tickets. Par exemple, le numéro de SIREN n'apparaît que 18 fois sur les 100 tickets évalués, et jamais sur les tickets Carrefour ce qui explique l'impossibilité d'évaluer l'extraction dans le tableau 3.4. De même, le nombre d'articles présents sur le ticket n'apparaît que 16 fois sur les 50 tickets qui ne sont pas de Carrefour. Plus étonnant, l'adresse n'apparait pas sur 4 tickets Carrefour et sur 12 tickets non-Carrefour.

Nous avons également évalué l'extraction des produits et des informations qui leur sont associées. Nous avons donc 302 produits sur les 50 tickets Carrefour, dont 11 qui ont été oubliés, et deux mauvaises extractions. Sur les 291 produits bien extraits, toutes les quantités et prix unitaires sont corrects. L'information du poids et du prix au kilogramme n'est présente que six fois sur tous ces produits et est à chaque fois bien extraite. Sur les autres tickets, seulement 134 produits sur 216 sont bien extraits, mais 141 produits ont la bonne quantité et 140 le bon prix unitaire. Cela est dû aux tickets qui indiquent l'intitulé du produit sur la ligne précédant toutes ces informations et le prix. Le poids et le prix au kilogramme ne sont précisés que 16 fois et ne sont jamais relevés par nos règles.

Les informations sur le paiement (moyen de paiement, montant payé) ne sont pas toujours écrites sur le ticket, ou du moins de façon explicite. Le ticket de la figure 3.8 par exemple indique que la somme a été payée mais n'indique pas comment. Ainsi, nous avons 17 tickets sur les non-Carrefour qui n'indiquent pas d'informations sur le paiement.

Problèmes liés à l'OCR et à sa correction

La majorité des difficultés vient encore des erreurs OCR qui n'ont pas été corrigées. Nous pouvons constater par exemple que les O et les 0 sont toujours très confondus, y compris après une double correction manuelle, ce qui crée des erreurs d'extraction, comme nous pouvons le voir dans la figure 3.9 qui montre les instances du (mal classé) sous-produit « TTAL ». Ces 6 lignes, venant de 6 tickets différents, auraient dû être extraites comme total de chacun de ces tickets, et non comme produit. Le problème ici vient du fait que les lettres du mot « total » ont été espacées sur le ticket et l'OCR n'a donc pas pu concevoir cette suite de caractères comme un seul mot, ce qui aurait pu le contraindre à choisir la lettre O plutôt que le chiffre 0. Ce sont donc ici six tickets qui n'ont pas de total (Faux Négatifs), et 6 produits relevés qui n'auraient pas dû l'être (Faux Positifs).

De nombreuses autres informations sont également passées entre les mailles des filets des correcteurs : des points à la place de virgules dans les adresses après les chiffres, ce qui n'est pas prévu par les règles, ou des lettres et des chiffres mal corrigés ou oubliés, ce qui ne permet pas aux informations d'être correctement relevées.

L'évaluation telle qu'elle a été faite ne permet pas de voir l'impact de ces erreurs OCR sur les résultats de l'extraction d'information, mais nous constatons que, sur les tickets Carrefour, au moins une extraction a échoué pour chaque type d'information. Cette information est importante pour la suite de notre travail car cela impactera les

FIGURE 3.8 – Absence d'informations sur le paiement

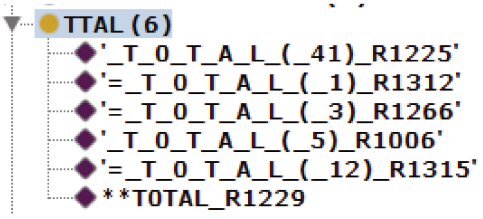


FIGURE 3.9 – Individus de la classe TTAL, sous-classe de Produit, avec un chiffre 0 pris pour une lettre O

résultats de la vérification des informations du document, notamment concernant les montants erronés ou non extraits.

3.5 Conclusion

Nous avons vu dans l'état de l'art de cette section qu'il existe de nombreux travaux autour de l'information, de la connaissance et de la modélisation des données de manière générale. Au cours des deux dernières décennies, de nombreux outils ont été mis en place pour extraire l'information de manière automatique et de nombreux standards ont été créés pour représenter les données et les modéliser. Ces méthodes d'extraction reposent essentiellement sur le pré-requis que l'information se situe dans un texte en langage naturel, ce qui permet de l'analyser syntaxiquement et sémantiquement avec les outils maintenant très fonctionnels de traitement automatique des langues et d'apprentissage automatique.

Cependant, nos documents ne sont pas en langage naturel, et ne possèdent pas non plus de structure explicite que l'on pourrait analyser pour extraire l'information comme sur des pages web. Nous avons donc adapté notre système d'extraction à notre corpus, en guidant la construction des règles par une ontologie que nous avons construite et que nous avons pu, avec des résultats tout à fait intéressants, peupler avec les informations de nos tickets de caisse.

Cette méthode, très bas-niveau dans le sens où nous avons travaillé essentiellement à l'échelle du caractère, nous a permis d'obtenir du très haut-niveau : de la sémantique. En effet, grâce à des expressions régulières, nous avons pu transformer des caractères en informations, qui répondent aux questions où, quand, qui et quoi.

Cette approche se veut générique, dans le sens où elle est adaptable : l'ontologie tolère largement l'ajout de concepts dans sa structure et nous pouvons rajouter toutes sortes d'expressions régulières ou de fonctions dans le code qui permet de peupler l'ontologie. De plus, la plupart des documents qui nous intéressent possèdent des données similaires concernant l'entreprise, ses coordonnées et son immatriculation par exemple. Nous pourrions donc rajouter des documents de type facture, bulletin de salaire ou avis d'imposition par exemple, qui comportent également des dates et des montants... Ainsi, notre implémentation s'est limitée à notre cas d'étude, mais l'approche est générique et peut s'appliquer à tout type de documents.

Par ailleurs, il serait intéressant d'intégrer à notre système d'extraction basé sur le texte les dimensions spatiale et graphique. En effet, l'emplacement physique de chaque information sur le ticket pourrait beaucoup aider à identifier l'information, tout comme la taille et la police utilisées pour écrire certaines informations, comme le total, qui se démarque régulièrement, ou le nom de l'entreprise, souvent dans la police du logo. Cependant, la diversité des styles de tickets et de leur taille rend difficile l'écriture de règles. Il faudrait donc avoir accès à des corpus annotés pour entrainer des algorithmes à reconnaitre l'information de façon supervisée.

L'extraction d'information nous a également permis de répondre à plusieurs questions que nous nous posions lors de la collecte de tickets et de la création du corpus, comme

l'origine géographique des tickets, la diversité des commerces explorés, le nombre de produits achetés ou les sommes dépensées en fonction des mois... Toutes ces statistiques, qui sont présentées et illustrées dans l'annexe C, permettent de mieux comprendre le corpus, et de se replacer dans un plan haut-niveau d'analyse du corpus, après ce chapitre dans lequel nous avons cherché à expliciter le sens des chaînes de caractères constituant nos documents.

Le chapitre suivant, qui correspond à une expérience à part entière, vient compléter la compréhension du document en apportant, toujours par des méthodes bas niveau, un peu plus de sémantique. En effet, extraire les informations telles qu'elles sont écrites sur le document, c'est intéressant, mais si on ne les comprend pas, on ne peut pas les traiter. C'est pourquoi nous avons décidé de chercher, pour chaque nom de produit abrégé, un équivalent long plus compréhensible pour l'humain et pour les systèmes sémantiques que nous présenterons dans le chapitre 5.

Chapitre 4

La gestion des abréviations

Sommaire

Dans l'objectif général de détecter les faux documents par la vérification des informations que contiennent les documents, nous les avons extraites et modélisées afin de pouvoir les comparer entre elles et les confronter à des informations que nous allons chercher dans des ressources externes. Or, les documents administratifs présentent de nombreuses contraintes, notamment en ce qui concerne la disposition du contenu et la place qui est accordée à chaque élément. En effet, les documents administratifs, à l'exception des lettres et des contrats, ont pour objectif d'être suffisamment concis et clairs pour être analysés au premier coup d'œil par l'humain.

Cette concision s'exprime souvent par une structure de mise en page contenant des tableaux et des emplacements précis pour les différents éléments habituels propres à chaque type de documents. Ainsi, nous trouverons souvent un logo dans la partie haute d'une facture, un entête de type coordonnées, un tableau avec des noms de produits ou de services à gauche, des prix à droite, un total en bas, ainsi que des « petites lignes » rappelant par exemple la législation à la fin du document. Cette structure très organisée impose des contraintes spatiales aux éléments textuels et graphiques : il s'agit de faire tenir dans un seul document, souvent une seule page pour des raisons économiques, toutes les informations nécessaires à l'utilité du document. Si les logos et autres éléments graphiques sont réduits, les informations textuelles quant à elles doivent être raccourcies tout en restant compréhensibles pour l'humain.

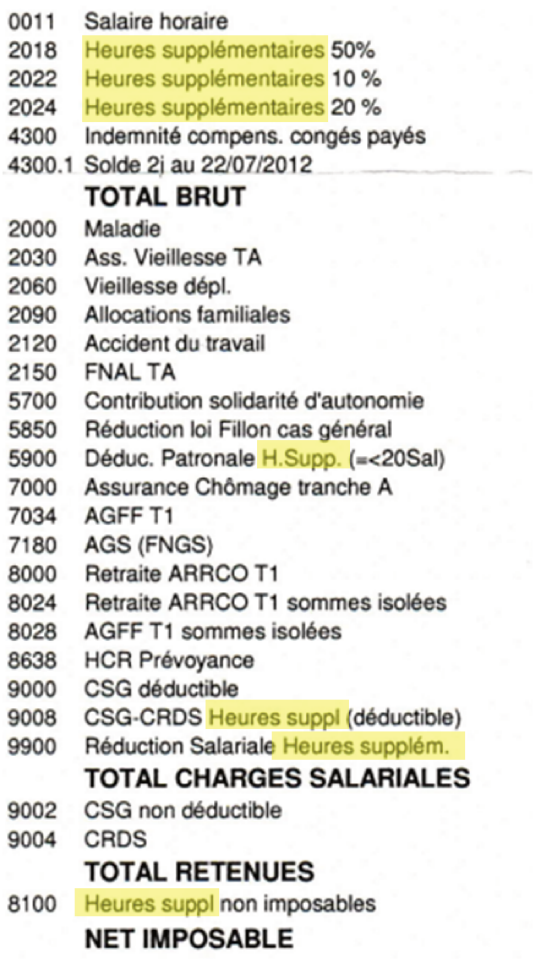
C'est à cause de cette contrainte que nous assistons dans de nombreux documents, comme les bulletins de salaire, les formulaires administratifs, les factures ou les tickets de caisse à des phénomènes linguistiques d'abréviations. La figure 4.1 montre par exemple une partie d'un bulletin de salaire contenant de nombreuses abréviations, et, encadrées en rouge, plusieurs abréviations différentes pour une seule signification (« heures supplémentaires ») :

— H. Supp.
— Heures suppl
— Heures supplém.
— Heures supplémentaires (forme complète)

Ces abréviations, présentes sous de nombreuses formes, sont un verrou à l'analyse automatique de l'information des documents. En effet, nous ne pouvons extraire et interpréter des informations qui ne sont pas explicites ou entières, et encore moins les vérifier. Nous devons donc chercher à associer les abréviations à leur forme complète, forme que nous nommerons « expansion ».

Cette tâche d'association des abréviations avec leur expansion n'est pas simple car elle met en jeu de nombreuses contraintes : l'absence de contexte, la diversité des types d'abréviations, le double niveau d'abrègement, l'ambiguïté des noms de produits...

Comme nous le montrerons dans la section 4.1, l'analyse et la mise en correspondance d'abréviations et de leur forme complète est un phénomène très peu étudié en linguistique et en traitement automatique des langues. Après la présentation de nos données et des difficultés qu'elles présentent (section 4.2), nous expliquerons notre approche pour apparier les abréviations et leurs possibles expansions (section 4.3). Enfin, la section 4.4 évaluera notre approche et discutera les résultats obtenus.



0011	Salaire horaire
2018	Heures supplémentaires 50%
2022	Heures supplémentaires 10 %
2024	Heures supplémentaires 20 %
4300	Indemnité compens. congés payés
4300.1	Solde 2j au 22/07/2012
	TOTAL BRUT
2000	Maladie
2030	Ass. Vieillesse TA
2060	Vieillesse dépl.
2090	Allocations familiales
2120	Accident du travail
2150	FNAL TA
5700	Contribution solidarité d'autonomie
5850	Réduction loi Fillon cas général
5900	Déduc. Patronale H.Supp. (=<20Sal)
7000	Assurance Chômage tranche A
7034	AGFF T1
7180	AGS (FNGS)
8000	Retraite ARRCO T1
8024	Retraite ARRCO T1 sommes isolées
8028	AGFF T1 sommes isolées
8638	HCR Prévoyance
9000	CSG déductible
9008	CSG-CRDS Heures suppl (déductible)
9900	Réduction Salariale Heures supplém.
	TOTAL CHARGES SALARIALES
9002	CSG non déductible
9004	CRDS
	TOTAL RETENUES
8100	Heures suppl non imposables
	NET IMPOSABLE

FIGURE 4.1 – Exemple d'abréviations dans un bulletin de salaire.

4.1 Travaux sur les abréviations et leur(s) expansion(s)

4.1.1 Définitions

Les grammaires de référence en linguistique ne décrivent pas beaucoup les abréviations dans la langue française, leurs usages et leurs valeurs sociolinguistiques, sémantiques et pragmatiques. La *Grammaire méthodique du Français* (Riegel et al. 2016) n'y accorde par exemple que deux pages et ne s'intéresse qu'à la troncation et à la siglaison. L'usage des abréviations semble alors anecdotique car n'appartenant qu'à l'aspect scriptural de la langue et propre à chaque scripteur : « Chaque scripteur adulte possède ses propres abréviations pour la prise de notes rapides. Il ne s'agit pas ici de créations de mots mais de commodités d'impression ou d'écriture. »

C'est l'aspect graphique qui prime également dans *Le Bon Usage* (Grevisse & Lits 2009) pour la définition de l'abréviation, marquant une différence avec le phénomène

de réduction, qui, lui, est la construction d'un nouveau signifiant à partir d'un autre ayant le même sens (ou « signifié ») par la suppression d'une partie de la forme pleine, comme dans « bio » pour « biologique », par exemple. Les auteurs relèvent tout de même une exception concernant le sigle, où l'abréviation rentre dans le langage courant en devenant un mot (ou « signifiant »), comme c'est le cas pour « CAPES ». L'abréviation n'est donc ici considérée que comme un phénomène graphique non créatif. Grevisse & Lits (2009) situent par ailleurs les unités de mesure comme étant des symboles, et non des abréviations, en faisant remarquer que les unités, en métrologie ou en chimie par exemple, ont perdu leur valeur d'abréviations pour prendre une valeur symbolique et ne sont donc plus suivies de points.

Si les abréviations pour ces deux ouvrages de références ne sont que pure pratique graphique personnelle, pour Martinet (1967) au contraire, les abréviations servent à ce qu'il appelle « l'économie de la langue », c'est-à-dire le moyen de pouvoir communiquer tout en fournissant le moindre effort. Il s'agit dans le cas des abréviations de garder le strict nécessaire pour transmettre l'information et de gagner ainsi en mémoire, en temps et en espace. L'abréviation fait donc pour lui partie intégrante de l'évolution de la langue : plus une expression est amenée à être prononcée fréquemment, plus elle a de chance d'être abrégée, par élimination des éléments non-spécifiques, par « tronquement » (que nous appellerons plutôt « troncation ») et par siglaison. Le contenu informationnel, s'il est présent dans le nouveau signifiant par l'utilisation fréquente qui en est faite, n'est parfois signifié que par un élément non-significatif du mot savant d'origine, comme c'est le cas dans « métro », dont l'extension est « chemin de fer métropolitain ».

D'un point de vue du traitement automatique des langues, les abréviations sont parfois traitées comme des entités nommées à reconnaitre ou à normaliser, comme dans les *shared tasks* du *Workshop on Noisy User-generated Text* de 2015 (WNUT) (Baldwin et al. 2015), et peuvent également être considérées comme des *Multi-Word Expressions*, c'est-à-dire des expressions polylexicales (de Marneffe et al. 2009).

Les abréviations sont également très présentes dans les micro-blogs (Twitter), les plateformes de discussions en ligne et les SMS. Ainsi, Liu et al. (2012) soulèvent le problème des abréviations pour l'extraction des entités nommées dans les tweets, qui mettent en échec leur approche. Joseph (2008) décrit les différentes abréviations présentes dans un corpus de SMS et en dresse la typologie, à la manière de Krautgartner (2003), qui s'intéressait, elle, aux abréviations dans les webchats francophones.

4.1.2 Désambiguïsation des abréviations

Nous avons trouvé peu de travaux dans le domaine du traitement du langage naturel sur l'appariement des différents types d'abréviations avec leurs expansions, sans contexte. Cependant, on trouve de nombreux travaux sur l'extraction d'acronymes, ou plus exactement de sigles, dans les textes et leur désambiguïsation en fonction du contexte. Le terme « *abbreviation* » en anglais est d'ailleurs très souvent utilisé pour ne parler que des acronymes ou des sigles.

Roche & Prince (2007) décrivent AcroDef, un système pour trouver la bonne expansion d'un acronyme alors qu'elle n'est pas présente dans le texte. Ce système utilise le

contexte : il compte le nombre de réponses à une requête contenant les mots des possibles expansions, trouvées grâce à un site recensant les sigles français courants, et des n-grammes de 1 à 3 mots des autres mots présents dans le texte à désambiguïser. Plusieurs mesures statistiques sont alors calculées afin de classifier le texte en fonction du contexte, et donc de trouver l'expansion la plus probable de l'abréviation à désambiguïser.

La désambiguïsation des acronymes est particulièrement présente en biologie et en médecine, où les acronymes sont très nombreux, tant dans les articles académiques, en particulier sur MEDLINE [1], que dans les rapports cliniques. Yeates (1999), Yu et al. (2002), Larkey et al. (2000) par exemple, présentent des approches pour tenter d'extraire les abréviations du texte. Les auteurs utilisent des règles de correspondance de formes pour faire correspondre une abréviation à sa forme complète : les acronymes sont souvent en majuscules, et parfois entre parenthèses quand ils apparaissent pour la première fois Nadeau & Turney (2005). Ils utilisent ensuite des méthodes d'apprentissage automatique basées sur le contexte pour choisir la bonne expansion à partir de dictionnaires de paires d'acronymes-expansions, comme l'expliquent Gaudan et al. (2005). Wu et al. (2012) comparent trois systèmes de traitement des abréviations sur des documents cliniques et concluent que ces systèmes ne sont pas suffisamment performants.

Les abréviations doivent également être traitées dans le domaine de la physique, où les noms de molécules sont très souvent abrégés par les auteurs d'articles scientifiques, avec parfois beaucoup de créativité (Batista-Navarro et al. 2015).

4.1.3 Méthodes de calculs de distances entre chaînes de caractères

Achananuparp et al. (2008) présentent de nombreuses mesures de similarité pour comparer deux textes. Les auteurs séparent ces mesures en trois classes : celle qui concerne la superposition des mots, celle qui prend en compte les fréquences des termes (TF-IDF) et celle qui englobe les mesures linguistiques (syntaxiques et sémantiques). Cependant, nos données ne sont que des chaînes de caractères, correspondant à des informations écrites sous formes abrégées et sous formes longues. Pour cette raison, nous ne pouvons travailler qu'avec des méthodes bas-niveau sur les chaînes de caractères, et non avec des approches sémantiques ou syntaxiques sur des textes complets.

Yu et al. (2016) ont réalisé l'inventaire des méthodes pour calculer la similarité entre deux chaînes de caractères, dans des optiques de recherche de chaîne de caractères, où l'on recherche une chaîne requête dans des chaînes sources, ou d'association de chaînes de caractères, où l'on recherche les paires de *tokens* les plus similaires dans des ensembles de *tokens*. Trois sortes de métriques sont relevées : les fonctions de similarité basées sur les caractères, celles basées sur les *tokens* et les fonctions de similarité hybrides.

La métrique la plus connue et la plus utilisée pour calculer une distance entre deux chaînes de caractères est la distance de Levenshtein (aussi connue sous le nom de distance d'édition), présentée dans Levenshtein (1966). Cette mesure donne le plus petit nombre de changements de caractères (suppression, insertion, substitution) pour passer d'une

1. Medical Literature Analysis and Retrieval System Online, `https://www.nlm.nih.gov/pubs/factsheets/medline.html`

chaîne à l'autre. Plus le résultat est faible, moins il y a de différences entre les chaînes de caractères. Par exemple, la distance d'édition entre « petit » et « plat » est de 3 : deux substitutions et une suppression. L'ordre de ces substitutions et suppressions n'a pas d'importance, comme dans les exemples suivants, dans lesquels la distance sera toujours de 3.

$$(1)\begin{cases} \texttt{p} \to \texttt{p} & \\ \texttt{e} \to \texttt{l} & \text{substitution} \\ \texttt{t} \to \texttt{a} & \text{substitution} \\ \texttt{i} \to & \text{suppression} \\ \texttt{t} \to \texttt{t} & \end{cases} \qquad (2)\begin{cases} \texttt{p} \to \texttt{p} & \\ \texttt{e} \to & \text{suppression} \\ \texttt{t} \to \texttt{l} & \text{substitution} \\ \texttt{i} \to \texttt{a} & \text{substitution} \\ \texttt{t} \to \texttt{t} & \end{cases}$$

De nombreuses variantes ont été proposées pour cette métrique (Navarro 2001). La distance de Hamming ne compte les modifications que dans des chaînes de caractères de même longueur. Il n'y a alors pas de suppression ni d'insertion. Cette distance est utile par exemple dans le cas de vérifications de bits de sécurité lors de la transmission de signaux codés. La distance de Damerau-Levenshtein, quant à elle, prend en compte les transpositions, c'est-à-dire les inversions de deux caractères, ce qui peut arriver lors d'une faute de frappe par exemple, en complément des trois types d'éditions de la distance de Levenshtein classique .

Dans notre problème, comme nous le verrons dans la section 4.3, les chaînes de caractères sont de longueurs différentes et seules les insertions nous intéressent, pour passer des abréviations à leur expansions. Nous recherchons les abréviations au niveau des caractères dans les mots, puis au niveau des mots dans les phrases. Pour cette deuxième étape, nous pouvons placer le problème dans la théorie des ensembles, en considérant la phrase comme un ensemble de mots, ou de *tokens*, ce qui nous permet d'utiliser des mesures comme l'indice de Jaccard ou le coefficient de recouvrement. Le coefficient de recouvrement ($\mathcal{R}$) calcule simplement la taille de l'intersection des deux ensembles : $\mathcal{R}(A, B) = |A \cap B|$. L'indice de Jaccard est un coefficient de similarité simple entre deux ensembles, défini comme la taille de l'intersection divisée par la taille de l'union, comme présenté dans la formule suivante. Cet indice se situe entre 0 (rien en commun) et 1 (ensembles identiques).

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \qquad 0 \leq J(A, B) \leq 1$$

Une autre mesure fréquemment utilisée est le coefficient de Dice, également basé sur des ensembles, dont la formule est la suivante :

$$D(A, B) = 2\frac{|A \cap B|}{|A| + |B|} \qquad 0 \leq D(A, B) \leq 1$$

4.2 Présentations des données et analyse des abréviations

Nous avons créé manuellement un corpus de 248 noms de produits tels que typographiés sur des tickets de caisse des magasins Carrefour City. Nous nous sommes intéressée

dans ce chapitre à un nombre restreint d'informations, provenant toutes du même magasin, afin de simplifier la construction de la vérité terrain et l'évaluation de notre approche. Cependant, la méthode proposée pour trouver la meilleure expansion possible aux noms de produits du corpus est applicable à tout type d'informations extrait de n'importe quel document, sous condition de pouvoir accéder à une base d'informations suffisamment fournie pour contenir les expansions possibles de l'abréviation recherchée.

Les noms de produits inscrits sur les tickets de caisse ont pour contrainte d'être compréhensibles pour l'homme tout en étant limités à 20 caractères, dans le cas des tickets provenant de Carrefour City. Ces noms de produits contiennent de nombreuses réductions, ou abréviations, tant au niveau des mots (suppression de lettres) qu'au niveau des syntagmes (suppression de mots).

4.2.1 Typologie des abréviations

L'abréviation est un phénomène graphique : c'est la réduction d'un ou plusieurs mots par suppression ou modification d'une partie de ses lettres afin de former une graphie plus courte, mais compréhensible, de ce terme. La seule constante que nous avons pu observer dans notre corpus d'abréviations est la présence systématique de la première lettre du terme abrégé en première place de l'abréviation. Nous dressons dans cette section une typologie des différentes formes d'abréviations observées dans notre corpus.

Apocope avec point

L'apocope est la suppression des lettres finales d'un mot, ne laissant comme dans les exemples (1a) du tableau 4.1 parfois que quelques lettres. La présence du point sert à la fois à signifier que les caractères le précédant correspondent à une abréviation et à la fois à marquer la séparation entre les termes abrégés. Nous pouvons remarquer que le point n'est suivi d'une espace que dans un seul cas dans notre corpus, ce qui est contraire aux règles typographique mais justifié vis-à-vis de la contrainte spatiale inhérente à ce type de document.

Apocope sans point

Les exemples (1b) du tableau 4.1 montre des apocopes qui ne sont séparées que par des espaces et non par des points. Cette absence de marque d'abréviation rend difficile leur détection automatique, d'autant qu'aucune règle ne semble prévaloir quant à l'apparition ou non des points. Ainsi, dans un même nom de produit, nous pouvons trouver des apocopes avec et sans points.

Syncope

La syncope est la suppression de lettres à l'intérieur du mot. Nous avons repéré plusieurs types de syncopes dans notre corpus :

Tableau 4.1 – Exemples d'abréviations du corpus

Type	Abréviation	Expansion des mots, *Expansion attendue*
(1a)	**CONF.FIG.PROV.R.**FR	CONFiture FIGues PROVence Reflet FRance *Confiture de figues de Provence Reflets de France*
(1a)	**SAUCIS.LENTILL.**CRF	SAUCISses LENTILLes CaRreFour *Plat cuisiné saucisses lentilles Carrefour*
(1b)	**TAB CHO** CRF NR **BIO**	TABlette CHOcolat CaRreFour NoiR BIOlogique *Chocolat bio noir 74% cacao Carrefour Bio*
(1b)	75**CAB ANJ** RSE MDD	75 CABernet ANJou RoSE Marque Du Distributeur *Vin rosé Cabernet d'Anjou*
(2a)	160G **BLC PLT** 4TR.F	160 Grammes BLanC PouLeT 4 TRanches Fines *Blanc de poulet Carrefour*
(2a)	190G **SDW** TR. **JBN** C	190 Grammes SanDWich TRiangle JamBoN Cheddar *Sandwich triangle jambon cheddar Sodebo*
(2b)	**PT** BEUR.CHO **LT** NOI	PetiT BEURre CHOcolat LaiT NOIsette *Biscuits petit beurre chocolat lait Petit Ecolier*
(2b)	1/2 RAVIOLI PUR **BF**	un demi ravioli pur BoeuF Ravioli pur bœuf Carrefour
(2c)	PT DEJ.FRUITS **RGES**	PetiT DEJeuner FRUITS RouGES *Biscuits petit déjeuner aux fruits rouges Carrefour*
(2c)	FILET **MAQUERX** MOUT	FILET MAQUERauX MOUTarde *Filets de maquereaux moutarde Petit Navire*

— Le cas le plus fréquent est la suppression des voyelles pour ne garder que trois consonnes du mot, qui peuvent être des consonnes sonores ou non, comme dans les exemples (2a) du tableau.
— Un autre type de syncope est le maintien de la première et de la dernière lettre du mot, principalement quand le mot est court, comme dans les exemples (2b).
— Le troisième type de syncope que nous pouvons relever dans notre corpus est la suppression de certaines voyelles du mot, parfois au début, parfois à la fin, comme nous pouvons le constater dans les exemples (2c).

Quelques-unes des syncopes relevées dans le corpus sont suivies de points, ce qui est contraire aux règles typographiques et peut créer des confusions pour le traitement automatique des abréviations, car cela signifie que le point ne vient pas remplacer des lettres supprimées en fin de mot, mais seulement marquer l'abréviation. Une autre difficulté vient du fait que certains termes abrégés sont d'origine anglaise comme *Brick Square* pour signifier une brique de jus, abrégé en BRK SQR.

Troncation

La troncation est une abréviation particulière dans le sens où elle est un phénomène non seulement graphique, mais également phonétique et linguistique. Le mot sous sa

forme raccourcie est tellement utilisé qu'il existe maintenant dans le langage courant. Nous trouvons fréquemment dans notre corpus le mot BIO, qui est une abréviation de « issu de l'agriculture biologique », et dans une moindre mesure MAYO, pour « mayonnaise », qui fait partie du langage courant de registre familier, ainsi que THAI, souvent utilisé pour parler de la cuisine thaïlandaise.

Sigle

Le sigle est l'assemblage des premières lettres de chaque mot d'un terme composé ou d'un nom de marque composé de plusieurs mots. Si le sigle est prononcé comme un mot, c'est un acronyme. Dans notre corpus, le sigle est parfois suivi d'un point et toutes les occurrences relevées sont constituées de trois lettres, dont parfois l'initiale d'une préposition. Le tableau 4.2 présente quelques exemples de sigles du corpus.

Tableau 4.2 – Exemples de sigles

Sigles	Expansion
2X30CL VELOUT.**PDT.**	Velouté de Pomme De Terre
PT L'EVEQUE **RDF**	Pont l'Eveque Reflet De France
YAF PANIER FRAM/MU	Yaourt Aux Fruits Panier de Yoplait framboise/mûre
MORBIER **AOP**	Morbier d'appellation d'origine protégée Juraflore

Symboles

Les mots sont parfois représentés par des symboles mathématiques, ou en contiennent. Nous retrouvons ainsi fréquemment les symboles « + » (« plus ») et « / » (barre oblique). Ces caractères spéciaux posent des problèmes de segmentation. Leur usage non-uniforme ne permet pas de procéder à un traitement systématique. En effet, dans l'exemple (1) du tableau 4.3, nous pouvons constater que la barre oblique peut séparer des mots (abrégés ou non) afin de signifier un mélange d'ingrédients. Dans ce cas, il suffit de considérer la barre oblique comme un séparateur, ou comme la conjonction de coordination « et ».

L'exemple (2) montre une barre oblique entre deux initiales S. Il s'agit, tout comme dans les exemples (1), d'un séparateur entre deux initiales, mais la sémantique n'est pas la même. Si dans le cas précédent, la barre oblique séparait deux unités de sens, elle s'interpose ici entre deux mots (préposition et nom) d'une même unité, reprenant le modèle des abréviations souvent utilisées dans la signalisation routière du « /s » et « s/ » symbolisant respectivement les prépositions « sous » et « sur ». On doit noter que cette abréviation « s/s » peut correspondre dans notre corpus au bigramme « sans sucres » mais également « sans sel ».

La barre oblique se trouve également sous sa forme signifiante de symbole mathématique (3). Dans ce cas, nous ne pouvons définir la barre oblique comme séparateur car la suite de caractères de type chiffre/barre oblique/chiffre, qui correspond à une fraction mathématique, constitue une seule unité sémantique (ici, « demi »).

D'autres symboles mathématiques sont fréquemment utilisés dans les abréviations des tickets de caisse, comme le signe plus et la lettre X, utilisée à la place de la croix de multiplication. Dans l'exemple (4), le signe « + » a une vraie valeur mathématique : il s'agit de signifier une addition entre deux quantités chiffrées. Nous pouvons noter que dans cet exemple, le caractère « + » est compris entre deux espaces, ce qui lui confère un statut de token à part entière. En revanche, dans l'exemple (5), le caractère « + » s'insère entre deux mots (abrégés), désignant un produit constitué de deux entités. Le terme « duo » permet de compléter le sens du caractère « + » en signifiant que c'est un lot de deux produits qui ne sont pas toujours associés. Ce caractère représente donc, dans les deux cas, la combinaison de deux unités qui ne sont pas habituellement assemblées, cependant le traitement à effectuer n'est pas le même dans les deux exemples. Dans le premier cas le caractère doit être interprété comme un mot à désambiguïser, alors que dans le second, le caractère doit être considéré comme un séparateur à ne pas intégrer aux mots l'encadrant.

Tableau 4.3 – Exemples de symboles

Type	Initialisms	Expansion and *English gloss*
(1)	BURGER EMM/CHEDDAR	*Burger emmental* ***et*** *cheddar*
	YAF PANIER FRAM/MU	*Yaourt Aux Fruits FRAMboises et MÛres*
(2)	KRISPROLL.S/S	*Krisprolls* ***sans sucre***
(3)	BEURRE BIO 1/2 SEL	*Beurre bio* ***demi****-sel*
	1L LT 1/2EC LACTEL	*1 litre de lait* ***demi****-écrémé Lactel*
(4)	CHOCOLATINE X4 + 1 G	*4 chocolatines* ***et*** *une gratuite*
(5)	DUO TERRIN+MOUS.CD	*Duo terrine* ***et*** *mousse de canard*

Chiffres

Le mélange de chiffres et de lettres est particulièrement problématique dans nos données car nous ne pouvons pas appliquer une règle générale pour le traiter. En effet, il existe plusieurs cas de figures où des chiffres et des lettres sont mêlés :

1. Les expressions de quantités, associant un nombre et l'abréviation classique d'une unité de mesure ou ses variantes :
   — 85G : 85 grammes
   — 200GR : 200 grammes
   — 1KG : 1 kilogramme
   — 75CL : 75 centilitres
   — 750ML : 750 millilitres
   — 1L : 1 litre
   — 30M : 30 mètres
2. Les expressions de quantités, associant un nombre et l'abréviation d'un mot :
   — 4T : 4 tranches

   - 4TR.F : 4 tranches fines
   - 32D : 32 dosettes
   - 4FRT : 4 fruits

3. Les expressions de quantités, associant un nombre et un mot complet sans espace séparatrice : 4FROMAGES
4. Le nombre d'unités, représenté par la lettre « X » signifiant ici « fois » accolée avant ou après le nombre :
   - X6 : 6 unités
   - 4X : 4 unités
   - 5X100 : 5 unités de 100
   - 4X125G : 4 fois 125 grammes
5. Les abréviations symbolisées mathématiquement, associant une fraction et des lettres : 1/2 pour « demi »
6. Les noms de produits ou de marques contenant des chiffres et des lettres, qu'il ne faut donc pas séparer :
   - 3D'S : Biscuit apéritifs Bénénuts
   - C4 : taille d'enveloppe
   - T45 : type de farine
7. L'absence d'espaces entre lettres et chiffres due au manque de place sur le ticket et au fait que l'humain est capable d'analyser la différence lettres/chiffres et de segmenter correctement les mots même sans séparateur :
   - CAMEMB.BIO CRF250G : Camembert bio(logique) Carrefour de 250 grammes
   - 75 VDP OC RG BIO09 : Vin de pays d'Oc rouge bio(logique) de 2009

Nous pouvons remarquer dans ce dernier exemple que les nombres peuvent également être abrégés, notamment quand il s'agit d'années, fréquemment présentes pour les achats de bouteilles de vin.

4.2.2 Autres caractéristiques problématiques

Si les différentes formes de réductions présentes dans le corpus présentent des enjeux de segmentation et d'analyse du texte, d'autres caractéristiques propres au format du document doivent faire l'objet d'une attention particulière. En effet, la segmentation en mots des tickets de caisse est rendue difficile par divers éléments typographiques, telles que la casse utilisée, la présence de points d'abréviations ou encore la concaténation de chiffres et d'unités de mesure. Dans de rares cas, aucun caractère ne permet d'identifier la séparation entre différents mots, comme dans l'exemple ci-dessous où les lettres X et G sont chacune des unités sémantiques et doivent être séparées de l'abréviation « DESS. ».

Nom de produit : 4X100GDESS.PANACHE

Segmentation : 4 X 100 G DESS. PANACHE

Signification : quatre fois cent grammes de crèmes dessert panachées

Éléments typographiques

La casse Les caractères de ce corpus sont tous en majuscule, ce qui ne permet pas de visualiser aisément les sigles par exemple, comme c'est le cas dans la plupart des travaux sur les sigles et acronymes dans des textes en langage naturel. Cela ne permet pas non plus de distinguer les noms propres, comme les noms de marques par exemple, ce qui aurait pu aider à l'analyse des documents.

Les signes diacritiques Le corpus ne présente aucun signe diacritique, c'est-à-dire aucun accent ou aucune cédille. Cette absence est également un handicap pour l'analyse car elle peut provoquer l'ambiguïté de certains mots, comme PATE, qui peut correspondre à « pâté » ou « pâte ». Nous pouvons aussi observer dans notre corpus un cas problématique probablement lié à l'encodage des accents : BAGUETTE C R ALES (Baguette céréales Carrefour). Les caractères accentués n'ont pas dû passer dans le logiciel de caisse ou d'impression et sont donc remplacés par des espaces. Il s'agit cependant d'un *hapax*.

La ponctuation Certains noms de produits comportent des points qui font à la fois office de caractères de segmentation et d'indices de troncation. Dans l'exemple ci-dessous, il faudrait séparer les mots « DOS. » et « SENSEO » tout en intégrant le point au premier, à l'inverse de l'espace, pour éventuellement forcer la recherche de la forme longue du mot par la suite. En effet, dans cet exemple, la distance la plus faible trouve le mot « dos » au lieu du mot « dosette », ce qu'on sait impossible grâce à la présence du point qui signifie que le mot a été abrégé.

Nom de produit : 8XDOS.SENSEO CAPP.

Segmentation : 8 X DOS. SENSEO CAPP.

Signification : huit dosettes Senseo cappuccino

Il faut toutefois faire attention à ne pas appliquer la segmentation lorsqu'un point est compris entre deux chiffres, comme dans l'exemple « CRISTALINE 1.5L », car il s'agit d'un nombre décimal, et donc d'une seule unité sémantique.

Éléments pseudo-syntaxiques

Au-delà des réductions à l'échelle du mot et des éléments typographiques problématiques expliqués précédemment, nous avons également besoin d'analyser les processus de réductions à l'échelle du nom de produit, que nous pouvons considérer comme des expressions multi-mots afin de mieux comprendre les difficultés d'alignement entre formes courtes et formes longues. Ces éléments concernent les procédés de réductions utilisés, non plus d'un point de vue graphique mais d'un point de vue syntaxique et sémantique. Nous verrons par exemple que les réductions peuvent être ambigües, ne satisfaisant pas à des règles systématiques et cohérentes, ou encore que l'agencement des mots dans le nom du produit n'est pas régulier.

Ellipse Les noms de produits ne comportent généralement que les noms et adjectifs nécessaires à la discrimination du produit. Les prépositions et déterminants ne sont donc pas présents, sauf dans de rares cas, comme EMMENTAL EN TRANCH, PINEAU DES CHARENTES ou encore PENNE A LA BOLOGN. Nous assistons donc à des ellipses syntaxiques, comme :

— 140G TARTE POMMES : Tarte aux pommes
— BEIGNETS CALAMARS : Beignets de calamars

Ces ellipses n'empêchent pas la compréhension des noms de produits, car ces mots sont quasiment vides de sens. Certaines ellipses concernent des mots plus importants pour la compréhension, car porteurs de sens, comme dans les exemples suivants. Nous appellerons cela des ellipses sémantiques.

— BLANC SOIF D'EVASION : Vin blanc Sauvignon Soif d'Evasion
— 205G DESSERT NOIR : Chocolat noir Nestlé Dessert ou Chocolat dessert noir Carrefour
— X4 POM.MURE MATERN : Compotes pomme mûre s/sucres ajoutés Materne
— VAN CACAO CRAQU : Dessert glacé vanille cacao craquant Viennetta
— CRISTALINE X6 : Eau de source Cristaline

L'ellipse du mot « vin » dans les expressions « vin rouge », « vin blanc » et « vin rosé » par métonymie dans le langage courant ou familier est fréquente. Cette réduction n'est donc pas très gênante pour la compréhension humaine. L'absence des mots « chocolat », « compote », « dessert glacé » et « eau », en revanche, est plus gênante car ce sont les termes « clé » pour comprendre ce qu'est le produit : ceux qu'on utilise quand on parle du produit (« j'ai acheté du chocolat/de la compote... »). De même, l'absence de l'unité de mesure dans certains noms de produit, qui est également une ellipse sémantique peut causer une ambiguïté : 75 CAB ANJ RSE MDD signifie « bouteille de 75 centilitres de Cabernet d'Anjou rosé Marque de distributeur » (et non pas 75 bouteilles...)

Ordre des mots aléatoire L'ordre des mots dans les noms de produits n'est pas toujours intuitif. Par exemple, dans 300G MOUSSAKA BARQ (Barquette de Moussaka), le complément du nom se situe après le nom complété. De même, l'ordre des mots dans des produits similaires, mais de tailles différentes par exemple, n'est pas toujours le même, comme nous pouvons le voir dans les exemples suivants :

1. KINDER SCHOKOBONS : Kinder Schoko-Bons
   200G SHOKOBONS KI : 200 grammes de Kinder Schoko-Bons
2. BRK SQR 1L PJ POMM : BRicK SQuaRe 1 Litre Pur Jus de POMMe
   BRK SQR PJ MULT 1L : BRicK SQuaRe Pur Jus de MULTifruits 1 Litre
   P.J.MULTIFRUITS 1L : : Pur Jus de MULTIFRUITS 1 Litre

Ambiguïté des réductions Les réductions peuvent être ambigües, comme nous l'avons vu dans la partie 4.1. Ainsi, nous trouvons dans notre corpus certaines abréviations de mots identiques qui ne correspondent pas au même mot.

1. **PDT** BEURRE PLQ 82% : Plaquette de beurre doux **Président** 82%

PDT GROSSE KG : Grosses **pommes de terre**

2. **PT** DEJ.FRUITS RGES : Biscuits **petit** déjeuner aux fruits rouges Carrefour
   PT L'EVEQUE RDF : **Pont** l'Evêque Reflets de France

Nous pouvons voir que les procédés d'abréviation utilisés dans le premier exemple ne sont pas les mêmes (le premier est une syncope alors que le deuxième est une siglaison). Cela peut poser problème dans un processus d'apprentissage automatique : une forme correspond à deux résultats. Le contexte, ou ici les quelques autres mots du nom de produit, est donc important.

Plusieurs réductions pour une seule forme longue Pour certains produits récurrents dans les listes de courses de notre corpus, nous avons pu remarquer la diversité des formes réduites : pour un même mot, nous pouvons trouver jusqu'à six abréviations différentes.

— Pizza : PIZZ., PIZZA
— Salade : SLD., SLDE, SALADE
— Jambon : JBN, JBON
— Sandwich : SDW, SANDW., SANDWICH
— Rosé : ROSE, RSE, ROS
— Yaourt : YT, YRT, YAF (yaourt aux fruits)
— Carrefour : CR, CRF, CRF., CARF, CARREFO, CARREFOUR

Nous pouvons remarquer que certaines formes réduites ne le sont finalement pas tant que ça en terme de caractères économisés. C'est le cas par exemple de PIZZ. qui contient autant de caractères que PIZZA.

Variantes orthographiques Les tickets de caisse contiennent également des variantes graphiques ou orthographiques de certains mots. Parfois, ces variantes allongent le mot, comme c'est le cas pour HALLAL (au lieu de « halal », forme plus conventionnelle). De même, il est plus aisé sur les tickets de simplifier la graphie de certaines marques. C'est le cas de SCHOKOBONS (pour « Schoko-Bons ») ou MINIBABY (pour « Mini Babybel »), qui est à la fois une contraction de deux mots et une abréviation.

L'hétérogénéité des réductions des noms de produit, que ce soit dans la diversité des formes réduites, ou plus généralement dans la variété des procédés de réductions, de la présence ou non des points ou encore du mélange de caractères spéciaux, de chiffres et de lettres, rend l'analyse du corpus riche et son traitement automatique complexe. Pour tenter de trouver la forme longue des produits qui correspond aux formes réduites présentes sur les tickets de caisse, nous avons créé un corpus de référence contenant une liste de produits dont les noms sont détaillés et complets.

4.2.3 Construction d'un corpus de référence

Obtenir des listes de références vendues par un petit magasin d'une chaîne de distribution n'est pas chose aisée. En effet, même si le magasin franchisé accepte de nous

aider, il n'a pas la main sur les noms des produits qu'il vend. En d'autres termes, le système d'information des magasins franchisés est géré depuis la centrale nationale (voire internationale) de l'enseigne, qui fournit à chaque magasin la liste des références et des étiquetages de chaque produit.

Nous avons donc décidé de récupérer une liste de produits vendus par l'enseigne par d'autres moyens. Nous avons ainsi extrait une liste d'intitulés de produits distribués par le service Drive d'un magasin de la même enseigne. Le service Drive permet de faire ses courses en ligne en choisissant les produits grâce aux renseignements que le site fournit : nom détaillé du produit, prix, prix au kilogramme, poids du produit, promotions et photographie. Ces informations sont précieuses pour qui veut faire ses courses, mais également pour qui veut obtenir une liste de produits détaillée.

Pour cela, nous avons parcouru de nombreuses pages du site web d'une grande surface de la même enseigne dont nous avons ensuite analysé le code HTML afin d'en extraire les informations qui nous intéressent : le nom du produit, la quantité, le prix et le prix au poids.

Le fait que le *Drive* soit celui d'une grande surface apporte un avantage et un inconvénient : il permet d'avoir un plus grand choix de produits, et donc un corpus de référence plus large, mais dans le même temps, les produits extraits ne sont pas toujours les mêmes que ceux achetés dans les magasins de proximité de l'enseigne. En effet, les consommateurs ne sont pas les mêmes dans une petite surface de proximité, située dans un quartier étudiant, et dans un drive d'une grande surface située en périphérie de la ville. Les produits vendus ne sont donc pas les mêmes non plus : nous trouverons principalement dans la première des plats tous prêts (salade, sandwichs, plats cuisinés), des petits gâteaux et des boissons, à l'unité et en quantité individuelle, alors que nous trouverons plus de produits frais ou non-préparés, en quantité familiale, dans la deuxième. Ce choix impacte donc directement la qualité de la comparaison entre les produits extraits des tickets de caisse et les produits issus d'Internet, autant au niveau de la recherche du produit en lui-même que de la comparaison des prix qui seront nécessairement différents selon le format du produit.

Nous avons ainsi pu extraire 13 888 produits ayant des nomenclatures différentes, ce qui correspond à 21 698 produits si l'on prend en compte les différentes quantités proposées et les différents prix auxquels ils sont vendus. En effet, les sites marchands proposent régulièrement des promotions. Ainsi, nous pouvons obtenir pour un seul produit plusieurs prix et plusieurs quantités.

Les noms de produits que nous avons récupérés différent beaucoup des noms de produits extraits des tickets de caisse. Si la brièveté est requise sur le ticket de caisse, c'est le détail et la précision qui est de rigueur sur le site web d'achat en ligne. Il s'agit d'être le plus exhaustif possible pour que le client sache quel produit il achète sans l'avoir sous les yeux. Les noms de produits issus du Web sont donc longs (37 caractères en moyenne) et contiennent en moyenne 5,5 mots. Ainsi l'exemple « Biscuits Petit beurre tablette chocolat fin Petit Ecolier » est particulièrement précis sur la composition de ce que tout un chacun appelle simplement « Petit Ecolier ». Chaque nom de produit contient un mot-clé qui définit le type de produit, souvent suivi d'un ou de plusieurs

termes qualificatifs (adjectif ou nom) et se termine par le nom propre du produit ou de la marque, comme dans « Céréales goût caramel/chocolat Lion ».

Ce corpus d'expansions potentielles pour nos noms de produits abrégés contiennent également quelques mots abrégés, notamment des sigles courants ou des troncatures telles que « choco » ou « PDT », des ellipses de prépositions comme « chocolat lait » au lieu de « chocolat au lait », ou encore des chiffres et des symboles.

Cette petite analyse préalable des noms de produits et de leurs possibles expansions nous permet de comprendre les difficultés à surmonter dans le traitement automatique de ces informations. En effet, nous ne pouvons chercher à vérifier une information que nous ne comprenons pas, ou qui n'existe pas en l'état, sans la transformer au préalable. Il nous faut donc trouver la bonne expansion des noms de produits abrégés avant la suite de la procédure de vérification d'information, et maintenant que nous avons une liste d'abréviations d'une part, et une liste d'expansions possibles d'autre part. Nous présentons dans la section suivante les approches que nous avons testées, puis nous détaillerons l'évaluation et la vérité terrain associée à ce corpus dans la section 4.4.

4.3 Approche proposée

Nous cherchons à trouver la bonne expansion parmi les 13 888 noms de produits extraits du web pour chacune des 248 expressions abrégées de notre corpus d'étude. Pour cela, nous pré-traitons les données dans un premier temps, puis, comme nous avons vu que les noms de produits contenaient des abréviations aussi bien au niveau des mots qu'au niveau du syntagme, notre algorithme produit dans un premier temps une liste de mots complets possibles, puis dans un deuxième temps, à partir de cette liste, notre approche trouve la meilleure expansion possible de la liste extraite du web.

4.3.1 Pré-traitement automatique

Notre analyse précédente nous permet de noter les principales différences entre nos deux ensembles de données. Tout d'abord, la forme abrégée est totalement en majuscules et ne contient pas de diacritique. Afin de rendre la source (termes extraits des tickets) et la cible (termes extraits du web) comparables, nous avons normalisé tous les caractères cibles pour les mettre en majuscules et supprimer tous les diacritiques.

Ensuite, nous devons segmenter à la fois les termes des tickets et les termes du web en mots, considérant les difficultés expliquées dans la section 4.2 concernant les symboles et l'absence éventuelle d'espace. Selon les cas, on considère un ou deux tokens autour d'un symbole comme `+`, `/` ou `..` En effet, comme nous l'avons vu, ces symboles peuvent séparer des chiffres ou des lettres. Nous prenons donc en compte uniquement les symboles entourés d'au moins une lettre avant et après le symbole. En revanche, l'apostrophe ne doit pas être traitée comme un séparateur de mots, car elle apparait la plupart du temps dans des noms de produits tels que PIM'S ou PT L'EVEQUE qui sont des noms propres dont les mots sont indissociables.

Nous séparons également en deux tokens les nombres suivis de plus de deux lettres, telles que « 4FROMAGES ». Nous choisissons une limite de deux lettres afin d'éviter de séparer les chiffres suivis d'une unité de mesure, comme dans 75CL (« 75 centilitres »).

Une fois cette segmentation des mots effectuée, nous comparons chacun des mots du corpus d'abréviations qui n'est pas suivi d'un point à la liste des mots du corpus d'expansions possibles. Si une correspondance est trouvée, cela signifie que le mot existe et n'est probablement pas abrégé. Nous avons relevé certaines exceptions cependant, comme DOS pour « dosette » qui existe dans le corpus d'expansions pour les dos de cabillauds et de colins.

Nous définissons enfin la liste des trigrammes les plus fréquents des termes extraits du web, c'est-à-dire la liste de trois mots consécutifs qui reviennent plus de huit fois dans les expressions extraites du web, afin de créer un dictionnaire des sigles les plus probables. En effet, nous avons remarqué précédemment que tous les sigles de notre corpus de tickets de caisse sont composés de trois lettres, et correspondent à des expressions multi-mots fréquentes, comme PDT pour « pomme de terre ». Le seuil de huit a été choisi afin de limiter la durée du traitement, en ne prenant que 1% des trigrammes du corpus, et de limiter l'ambiguïté des abréviations en faisant en sorte que chaque sigle soit unique. Ainsi, pour chaque token qui n'est pas un mot complet et qui est composé de trois lettres dans les expressions abrégées, nous le comparons aux éléments de la liste des sigles créés à partir des mots du web. Si un sigle correspond, alors le token abrégé est remplacé par les trois mots du sigle. Tous les sigles sont ainsi correctement désambiguïsés dans notre corpus.

4.3.2 Appariement automatique des mots

Après avoir écarté les mots non abrégés et les sigles, nous comparons les mots de forme raccourcie et chaque mot extrait du web commençant par la même lettre que les noms de produits. Ne garder que les mots partageant la même initiale permet de diminuer le temps de traitement particulièrement long de cette approche.

Nous avons choisi, en terme de comparaison, de calculer pour chaque mot abrégé une distance pondérée de Levenshtein ($\mathcal{LP}$) avec chaque mot du web et de sauvegarder le mot complet ayant la plus petite distance. La distance pondérée de Levenshtein nous permet, à la différence de la distance de Levenshtein classique ($\mathcal{L}$), de mettre un poids différent à chaque opération d'édition : la substitution, la suppression et l'insertion. Dans le cas de la recherche d'une distance minimale entre une forme raccourcie et une forme longue d'un même mot, le poids de l'insertion doit être plus faible que celui des deux autres opérations. En effet, si l'abréviation contient des lettres qui ne sont pas dans la forme longue testée (par substitution ou suppression), la probabilité que ce soit la bonne expansion est plus faible que si toutes les lettres de l'abréviation se retrouvent dans la forme longue et que seules des insertions de lettres ont eu lieu. Par exemple, nous désirons que pour l'abréviation PLT, la distance avec « POULET » (3 insertions, 0 substitutions, 0 suppressions, $\mathcal{L}(PLT, POULET) = 3$) soit moins grande que celle avec « PETIT » (2 insertions, 1 substitutions, 0 suppressions, $\mathcal{L}(PLT, PETIT) = 3$).

L'algorithme de Wagner-Fischer que nous utilisons pour calculer cette distance de Levenshtein pondérée Klein (2016) entre deux chaînes de caractères s et t construit une matrice $\mathcal{M}$ de taille *longueur de* s+1×*longueur de* t+1. Cette matrice est initialisée à 0 pour $i = 0, j = 0$, puis chaque valeur est calculée suivant les règles suivantes :

$$\mathcal{M}(0,0) = 0$$
$$\mathcal{M}(i,0) = i \times a$$
$$\mathcal{M}(0,j) = j \times b$$
$$\mathcal{M}(i,j) = minimum \begin{cases} \mathcal{M}(i-1,j) + a \\ \mathcal{M}(i,j-1) + b \\ \mathcal{M}(i-1,j-1) + c & \text{si} \quad s[i] \neq t[j] \\ \mathcal{M}(i-1,j-1) & \text{si} \quad s[i] = t[j] \end{cases}$$

où (a, b, c) sont respectivement les coûts de suppression, d'insertion et de substitution.

Nous constatons, après avoir testé les différentes combinaisons de coefficients de pondération, que la meilleure doit fixer ces coefficients à 11 pour les suppressions et les substitutions et laisser les insertions avec un coût de 1. Ces coefficients nous permet d'obtenir le meilleur appariement. Ainsi, pour l'exemple précédent, nous obtenons les matrices suivantes :

		P	E	T	I	T
	0	1	2	3	4	5
P	11	**0**	1	2	3	4
L	22	11	**11**	12	13	14
T	33	22	22	**11**	**12**	**13**

		P	O	U	L	E	T
	0	1	2	3	4	5	6
P	11	**0**	**1**	**2**	3	4	5
L	22	11	11	12	**2**	**3**	4
T	33	22	22	22	13	13	**3**

Les insertions s'incrémentent donc horizontalement, les suppressions, verticalement et les substitutions en diagonale. Le dernier chiffre correspond ainsi à la distance pondérée minimale entre deux chaînes de caractères. Nous avons donc :

$$\begin{cases} \mathcal{L}(PLT, PETIT) = \mathcal{L}(PLT, POULET) = 3 \\ \mathcal{LP}(PLT, PETIT) = 13 \\ \mathcal{LP}(PLT, POULET) = 3 \end{cases}$$

.

Pour chaque nom de produit, nous enregistrons le mot ayant la plus petite distance avec le mot abrégé dans une liste intermédiaire avec les expansions des autres mots de l'expression.

Nous avons également fixé un seuil minimum de 20 au-delà duquel nous jugeons la distance trop grande entre le mot abrégé et le mot testé. En effet, cela correspondrait à une insertion de 20 lettres, ce qui est trop pour un mot de la langue française présent sur un ticket de caisse, ou à des suppressions et substitutions, ce qui n'est pas souhaitable. Si aucun mot ne passe sous ce seuil, c'est-à-dire si aucune distance n'est satisfaisante, aucun mot n'est sauvegardé et c'est le mot tel qu'il est inscrit sur le ticket qui est enregistré dans la liste des mots complets de l'expression initiale donnée en entrée de l'étape suivante.

4.3.3 Appariement automatique des syntagmes

Le pré-traitement des noms de produits et l'appariement des mots abrégés avec leur probable expansion permet d'obtenir une liste intermédiaire de mots complets pour chaque syntagme extrait des tickets de caisse afin de les comparer aux syntagmes venant du web. En effet, ces derniers contiennent souvent beaucoup plus de mots, comme nous l'avons vu dans la section 4.2. Nous proposons cette fois-ci trois approches différentes pour cette comparaison : l'une utilisant la distance de Levenshtein pondérée, une autre utilisant le coefficient de Jaccard et une troisième utilisant le coefficient de Dice. Ces trois mesures nous semblent les plus pertinentes au vu de la nature des objets à comparer.

La première approche est sensiblement la même que celle présentée précédemment pour les mots, à cela près qu'elle s'applique à une chaîne de caractères plus longue, puisque nous concaténons la liste de mots probables obtenue précédemment pour obtenir un syntagme que nous pouvons ensuite comparer avec chaque syntagme du web. Là aussi, les meilleurs coefficients de pondération sont de 11 pour les substitutions et les suppressions et 1 pour l'insertion. Nous fixons en revanche le seuil minimal à 500, vue la longueur des noms de produits venant du web.

La deuxième approche consiste à calculer le coefficient de Jaccard en prenant la liste des mots obtenue et la liste des mots des syntagmes du web comme deux ensembles dont les éléments. Le coefficient de Jaccard ne calcule pas une distance entre deux chaines de caractères mais une similarité entre deux ensembles, avec une valeur comprise entre 0 et 1. Nous cherchons donc l'expansion possible ayant le coefficient le plus élevé, c'est-à-dire celle partageant le plus de mots avec la liste de mots intermédiaire.

La troisième approche est très similaire à la deuxième mais calcule le coefficient de Dice à la place de celui de Jaccard, ce qui a pour effet d'obtenir plus de valeurs différentes.

4.4 Évaluation

Afin d'évaluer notre approche, nous avons créé une vérité terrain, qui nous permet d'obtenir des résultats que nous présentons et discutons dans cette section.

4.4.1 Vérité terrain

Nous avons créé une vérité terrain en associant manuellement chaque nom de produits extraits des tickets de caisse à un ou plusieurs noms de produits extraits du web dans un format JSON. Pour cela, nous avons cherché pour les 248 noms de produit la ou les meilleure(s) expansion(s) parmi les 13 888. Il peut en effet y avoir plusieurs expansions possibles, quand les produits sont similaires et que seule la marque change ou quand l'intitulé est imprécis, comme c'est le cas dans l'exemple suivant :

PINEAU DES CHARENTES :

— Pineau des Charentes blanc Domaine du Feynard
— Pineau des Charentes blanc Jules Gautret
— Pineau des Charentes blanc Moulin de la Grange

— Pineau des Charentes rosé Jules Gautret
— Pineau des Charentes rouge Domaine du Feynard
— Pineau des Charentes rouge Moulin de la Grange

Dans ces cas, il est impossible de choisir une expansion plutôt qu'une autre, et nous avons donc choisi de mettre toutes les expansions possibles. Une autre difficulté que nous avons rencontrée lors de la création de cette vérité terrain est de connaitre, ou reconnaitre, toutes les variantes des noms de produits. Par exemple, « BEIGNETS CALAMARS » correspond à la fois à :

— Anneaux de calamars à la romaine Cité Marine
— Anneaux de calamars panure croustillante Costa
— Beignet à la romaine Carrefour

Dans les trois cas, nous n'avons qu'un mot sur les deux présents sur le ticket de caisse (« beignet » ou « calamars »). De même, « VACHE A BOIRE YRT » correspond aux yaourts à boire de la marque Michel et Augustin et à l'expansion « Yaourt vanille de Madagascar Michel et Augustin » (les autres goûts sont précisés sur les tickets). Dans ce cas, aucun mot n'est en commun, et nous avons dû nous aider de connaissances extérieures pour trouver les correspondances entre les noms de produits dans les deux exemples précédents. Il est d'ailleurs évident que nos approches ne peuvent pas fonctionner dans ces cas, puisqu'elles n'utilisent pas d'éléments de contexte.

Dans d'autres cas, le produit vendu par la supérette qui émet les tickets sur lesquels nous travaillons n'est pas vendu par le service Drive de l'enseigne. C'est le cas des « fougassettes », petites pizzas locales, et de quelques alcools locaux (Guignette et Rocheloise). Dans ce cas, nous n'avons pas de correspondance et nous devons alors retirer ces produits de l'évaluation, puisqu'ils ne trouveraient pas de correspondance.

Pour rendre l'évaluation plus intéressante, nous avons également demandé à un humain de réaliser une vérité terrain « à l'aveugle ». Cet humain avait pour mission d'écrire le nom de produit le plus complet possible à partir de ses connaissances et pouvait s'aider d'internet. Les procédures pour construire ces deux vérités terrain sont détaillées dans l'annexe B.

4.4.2 Résultats

Les résultats sont mesurés sur la base de la correspondance binaire entre les expansions proposées et la vérité terrain, qui peut retourner soit « Vrai » soit « Faux », selon si l'expression longue proposée est la bonne ou non. En outre, les cas où aucune vérité terrain n'est associée à l'abréviation sont ignorés (cas qui n'ont pas été annotés ou cas où même l'évaluateur humain ne peut pas effectuer l'évaluation). Ainsi, nous calculons un pourcentage de bonnes réponses, qui est de 36,87 % pour la meilleure combinaison de poids dans la méthode utilisant la distance de Levenshtein pondérée, 33,64% pour la méthode utilisant l'indice de Jaccard et 31,72% pour la méthode utilisant le coefficient de Dice.

Nous observons que ces deuxième et troisième approches ne fonctionnent pas aussi bien que la méthode basée sur la distance de Levenshtein pondérée. Ceci est, en partie, dû aux marques du pluriel sur de nombreux mots : alors que la distance d'édition

augmente seulement de 1 lorsqu'un « s » est présent en fin de mot dans le syntagme étendu et absent dans la liste de mots intermédiaires, le coefficient de Jaccard et celui de Dice, eux, considèrent que ce sont deux mots différents et ils n'appartiennent donc pas à l'intersection des deux ensembles, diminuant ainsi le coefficient. Par exemple :

Ticket OIGNON JAUNE

Liste intermédiaire [OIGNON (singulier), JAUNE (singulier)]

Web « Oignons jaunes » (pluriel)

Le coefficient de Jaccard entre l'ensemble de mots intermédiaire et l'ensemble de mots du nom venant du web est égal à zéro, ce qui entraîne une mauvaise réponse. La méthode du Levenshtein pondéré donne un score de 2 et fournit donc la bonne réponse.

Les mesures traditionnelles, à savoir la précision, le rappel et la F-mesure, ne sont pas les plus appropriées pour évaluer la performance de notre système, parce qu'il ne procure pas de prédiction, seulement des scores correspondant à des distances. Si nous sélectionnons les critères $distance = 0$ pour une prédiction correcte et $distance \neq 0$ pour une prédiction erronée, nous obtenons les résultats présentés dans la troisième ligne du tableau 4.4.

Les trois premières lignes de résultats nous indiquent que lorsque la distance est faible, ou la similarité élevée, le résultat de l'algorithme est toujours correct (précision élevée), mais aussi qu'il y a très peu de cas de ce genre (rappel très faible). Les trois lignes suivantes indiquent la F-mesure maximale que nous pouvons trouver en modifiant les critères de sélection des éléments pertinents : ainsi, si nous choisissons de dire que les expansions trouvées avec une distance de Levenshtein pondérée inférieure à 50 forment l'ensemble des prédictions correctes (et le reste, des prédictions incorrectes), nous obtenons une F-mesure de 56%. De même, si nous considérons toutes les expansions trouvées avec un coefficient de Jaccard supérieur à 0,33 comme étant des prédictions correctes, nous atteignons 68% de F-Mesure et 76% d'*accuracy*.

Les métriques présentées dans le tableau 4.4 sont calculées comme suit, avec V = vrai et F = faux, c'est-à-dire bien prédit ou non selon le critère, et P = positif et N = négatif, c'est-à-dire satisfaisant le critère ou non.

$$\text{Précision} = \frac{\text{VP}}{\text{VP+FP}} \qquad \text{Rappel} = \frac{\text{VP}}{\text{VP+FN}}$$

$$\text{Rejet} = \frac{\text{VN}}{\text{VN+FP}} \qquad \text{Accuracy} = \frac{\text{VP+VN}}{\text{VP+FP+VN+FN}}$$

$$\text{F-Mesure} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Étant donné que ces résultats ne permettent pas vraiment de tirer des conclusions quant à la qualité de notre approche, nous proposons d'utiliser des mesures qui permettent une approche plus qualitative des résultats. Seulement un peu plus d'un tiers de notre corpus de syntagmes abrégés est correctement trouvé, mais lorsque nous analysons de plus près les résultats, nous constatons que la performance n'est pas aussi faible, puisque le(s) mot(s) qui provoque(nt) l'échec correspond(ent) souvent au nom de la marque, alors que la majeure partie du syntagme est souvent correctement identifiée.

Tableau 4.4 – Évaluation avec les mesures traditionnelles

Méthode	Précision	Rappel	F-mesure	Rejet	*Accuracy*
Coef. Jaccard = 1	1.0	0.06	0.10	1.0	0.68
Coef. Dice = 1	1.0	0.06	0.11	1.0	0.70
Distance $\mathcal{LP}$= 0	1.0	0.04	0.07	1.0	0.65
Coef. Jaccard ≥ 0.33	0.60	0.79	0.68	0.74	0.76
Coef. Dice ≥ 0.44	0.52	0.81	0.63	0.65	0.70
Distance $\mathcal{LP} \leq 50$	0.47	0.69	0.56	0.54	0.59

Le tableau 4.5 montre la moyenne des coefficients de Jaccard, taux de recouvrement ordonné et distances de Levenshtein, calculés entre les syntagmes trouvés et la vérité terrain selon la méthode utilisée. Le taux de recouvrement ordonné est calculé selon la formule suivante :

$$\mathcal{R}(A, B) = \frac{|C|}{|A|}$$

où A et B sont deux chaînes de caractères (ou listes ordonnées de caractères), $|A| \leq |B|$ et C est la liste des caractères appartenant à A et à B et formée en parcourant A et B dans l'ordre de leurs éléments. Ce taux tient compte également des doublons. Ce taux de recouvrement est compris entre 0 (A n'est pas une abréviation de B) et 1 (A peut être une abréviation de B car toutes les lettres de A se retrouvent, dans l'ordre, dans B).

Nous pouvons observer qu'il n'y a en moyenne qu'entre 15 et 16 opérations d'édition entre les deux chaînes de caractères (colonne Levenshtein) et que près de la moitié des caractères des deux phrases est commune (colonne Recouvrement). La quatrième ligne du tableau 4.5 correspond aux mesures effectuées entre l'annotation humaine expliquée dans la section précédente et la vérité terrain. Nous observons que la distance est plus importante qu'avec notre algorithme et que le recouvrement, prenant en compte l'ordre des mots, est très faible. Ceci s'explique par l'ajout par l'humain de nombreux mots-clés décrivant le produit, comme « boîte de conserve ».

Tableau 4.5 – Évaluation avec les mesures de distance

Méthode	Taux de réussite	Jaccard	Recouvrement	Levenshtein
Dice	31.72	0.42	0.43	15.38
Jaccard	33.64	0.44	0.45	15.98
Levenshtein pondéré	36.87	0.49	0.43	15.46
Humain		0.19	0.06	18.34

4.5 Conclusion

L'hétérogénéité des réductions de noms de produits, que ce soit dans la diversité des formes abrégées ou plus généralement dans la variété des procédés d'abréviation, la pré-

sence ou l'absence de ponctuation ou le mélange de caractères spéciaux, de chiffres et de lettres, enrichit l'analyse du corpus et complique son traitement automatique. L'appariement automatique de mots et de phrases raccourcis avec leur expansion sans contexte est un défi indéniable. En effet, la majorité des travaux sur les abréviations et leurs expansions utilise le contexte pour trouver des relations syntaxiques ou sémantiques entre elles grâce aux techniques de traitement du langage naturel. Nous proposons cette approche innovante dans Artaud et al. (2018).

L'analyse qualitative de nos résultats expérimentaux est très prometteuse et nous pensons que la combinaison de l'approche présentée avec une analyse automatique des éléments contextuels (comme d'autres produits achetés, l'heure des achats...) nous permettrait d'améliorer le choix de la meilleure expansion pour chaque syntagme abrégé.

Par ailleurs, nous avons défini dans notre approche un périmètre très particulier pour analyser ces abréviations, en utilisant volontairement un vocabulaire restreint : celui des noms de produits sur les tickets de caisse d'un seul magasin. Nous pensons cependant que cette approche peut être adaptée pour résoudre des problèmes de compréhension du document dus à des abréviations présentes sur tous types de documents, à condition d'avoir accès à une base de connaissances externe contenant les bonnes expansions. Notre approche est donc très appliquée à nos données, et il serait intéressant de l'évaluer sur un ensemble de données plus large et plus varié. Cependant, l'évaluation de cette approche est longue et fastidieuse, et nécessiterait donc beaucoup de temps, ainsi que l'accès dans le même temps à des expressions abrégées et à une liste contenant les expansions possibles de toutes ces expressions.

Il serait également intéressant d'obtenir de nombreux autres corpus d'abréviations et expansions avec la vérité terrain associée pour pouvoir entrainer des algorithmes d'apprentissage automatique. Compte-tenu du nombre d'abréviations possibles présentées dans la deuxième section de ce chapitre, il faudra fournir à de tels algorithmes en effet beaucoup de données pour qu'ils puissent apprendre des modèles.

Ce travail expérimental nous permet d'ajouter à notre ontologie l'expansion la plus probable pour chaque produit des magasins Carrefour extrait. Trouver ces bonnes expansions permet de rendre les informations que les documents contiennent beaucoup plus claires, aussi bien pour l'humain que pour l'approche que nous allons présenter dans le chapitre 5. En effet, nous chercherons dans ce chapitre à vérifier les informations contenues dans le document, notamment en les comparant à ce que l'on pourrait trouver dans des ressources externes.

Chapitre 5

Vérification de la cohérence des informations des documents

Sommaire

Une fois les documents récoltés, préparés et falsifiés, les informations extraites et modélisées et les abréviations interprétées, nous avons enfin tous les éléments pour vérifier les informations contenues dans nos documents et détecter les anomalies. Pour cela, nous avons mis en place plusieurs stratégies pour détecter la diversité des fraudes présentes dans notre corpus.

Avant de présenter ces stratégies, nous avons cherché à obtenir quelques résultats de l'état de l'art sur la détection automatique des faux documents qui soient comparables à nos travaux. Comme nous l'avons vu dans les chapitres précédents, aucun corpus, et donc aucun résultat, n'est adapté à nos travaux. Nous avons donc décidé d'organiser une compétition sur notre corpus afin d'obtenir un nouvel état de l'art, c'est-à-dire des nouveaux résultats, comparables aux nôtres, avec de nouvelles méthodes.

Nous présenterons ainsi dans un premier temps l'organisation de la compétition ainsi que les approches proposées par les candidats et leurs résultats. Nous expliquerons ensuite les différents types d'approches que nous avons testés et détaillerons les différents indices de détection des fausses informations que nous utilisons. Nous testerons dans un troisième temps plusieurs combinaisons d'indices et nous évaluerons les meilleures d'entre elles. Enfin, nous discuterons nos résultats à la lumière de l'état de l'art et proposerons des perspectives.

5.1 Apport à l'état de l'art de la détection des faux documents : la compétition *Find it !*

Nous avons vu tout au long de ce manuscrit que peu de travaux ont été réalisés sur la détection automatique des faux documents et seuls des corpus d'images sont disponibles pour cette tâche. L'un des aspects importants de cette thèse a donc été de créer notre corpus, et de le diffuser grâce à l'organisation d'une compétition internationale dans le cadre de l'International Conference on Pattern Recognition (ICPR) que nous présentons dans Artaud et al. (2018).

En effet, il est difficile d'évaluer nos résultats sans les comparer à des résultats d'expériences dans des conditions similaires. C'est pourquoi nous avons organisé cette compétition : pour créer une base de référence, avec des méthodes innovantes à la fois sur l'image et sur le texte, pour que tout le monde puisse proposer de nouvelles approches et comparer ses résultats.

Nous présenterons donc dans un premier temps de cette section le matériel fourni aux participants : les corpus d'apprentissage et de test, la vérité terrain associée au corpus d'apprentissage et les métriques d'évaluation. Nous présenterons ensuite l'ensemble des méthodes proposées lors de cette compétition, ainsi que les résultats obtenus. Nous avons enfin proposé la même tâche de détection des faux documents sur le corpus de test à des humains : nous présenterons donc leurs résultats afin d'avoir une vue d'ensemble de la difficulté de la tâche.

5.1.1 Corpus, vérité terrain et métriques d'évaluation

La compétition *Find-it !*[1] proposait deux tâches aux participants : une première tâche de détection des faux documents parmi les vrais, et une tâche de localisation des altérations dans des faux documents. Pour cela, nous avons fourni aux participants un corpus unique d'images de documents, vrais et altérés, et pour chaque image, sa transcription : les textes utilisés dans nos travaux.

En effet, nous avons organisé cette compétition avec un double objectif :

— Tout d'abord, diffuser une base de données ne contenant aucune information privée pour qu'elle puisse être utilisée sans contrainte et sans modification de l'image (flou, masquage ou autre).

— Deuxièmement, fournir un corpus parallèle d'images et de textes et une base de référence unique pour tester et évaluer des méthodes basées sur l'image et/ou sur le texte pour la détection de faux documents.

Ce deuxième point est particulièrement important, car notre corpus vient combler une lacune reconnue dans le domaine de la détection des faux documents. De plus, nous souhaitions pouvoir comparer les méthodes basées sur du traitement de l'image et/ou sur du traitement de l'information textuelle sur un même corpus, unique. Les données à traiter ont donc été organisées sous la forme d'un ensemble d'images et de fichiers texte :

— Un fichier image formaté en png, représentant un reçu pouvant contenir une ou plusieurs contrefaçons,

— Un fichier texte contenant une transcription textuelle du contenu du ticket.

Les participants pouvaient utiliser soit les images, soit le texte, soit les deux. Dans notre thèse, nous n'avons utilisé que le texte, mais il serait intéressant de rajouter des indices provenant de l'analyse de l'image.

Pour la première tâche, nous avons fourni comme corpus d'apprentissage un ensemble de 500 documents, contenant 6% de documents altérés. Un fichier XML de vérité terrain indique le nom des documents et s'ils sont authentiques ou frauduleux. Pour la seconde tâche, un corpus d'apprentissage de 100 documents (images & textes) a été fourni avec une vérité terrain concernant la localisation des informations modifiées dans le texte d'une part et dans l'image d'autre part. Il n'y avait pas de chevauchement des documents frauduleux entre le corpus de la tâche 1 et le corpus de la tâche 2, de sorte que les participants pouvaient ajouter les 100 documents de la deuxième pour améliorer leur apprentissage pour la première. Nous ne nous étendrons pas sur la deuxième tâche car, bien qu'elle soit très intéressante, très peu de candidats s'y sont risqués et nous-même n'avons pas traité cette question.

Afin d'entraîner au mieux leurs algorithmes, nous avons fourni aux candidats le code qui évalue la détection de faux documents parmi d'autres, dans un ensemble de documents contenant à la fois des documents authentiques et des documents modifiés. Le script d'évaluation produit un fichier CSV avec les résultats de précision, rappel et f-mesure sur la détection de faux documents et, pour information, l'identifiant de chaque ticket et son

1. Le site de la compétition, sur lequel on peut également remplir un formulaire pour télécharger le corpus, est disponible à l'adresse : `http://findit.univ-lr.fr/`.

statut (vrai positif, faux négatif, vrai négatif, faux positif). Les « vrais » correspondent à une bonne prédiction et les « faux » à une mauvaise ; les « positifs » correspondent à des documents marqués comme « modifiés », les « négatifs » comme « authentiques ». Il ne s'agit donc pas de résultats sur la classification globale en vrai ou en faux, qui serait très élevée étant donné que nous n'avons que 6% de faux documents, mais de résultats sur la classification des faux documents uniquement.

La vérité terrain est évidemment la même pour la détection des faux documents à partir de l'image ou du texte puisque les documents sont les mêmes, et seule la vérité terrain du corpus d'apprentissage a été donnée aux participants. Le corpus de la phase test de la première tâche respectait les mêmes proportions que celui de la phase d'apprentissage, soit 30 faux documents sur 500 (6%).

Les deux corpus, d'apprentissage et de test, sont composés pour moitié de tickets de caisse de l'enseigne Carrefour. Étant donné que notre extraction d'informations fonctionne beaucoup mieux sur ces tickets que sur les autres, et que certains indices ne sont applicables qu'à ces documents, nous avons séparé en deux sous-corpus le corpus d'apprentissage et le corpus de test.

5.1.2 Approches proposées par les candidats

La compétition a suscité un vif intérêt dans la communauté du document, et nous avons reçu 36 inscriptions, provenant pour deux tiers du monde académique et pour un tiers, de l'industrie, de diverses institutions (police, justice) et de personnes non-affiliées. Si de nombreuses équipes se sont initialement inscrites à la compétition, nous n'avons reçu que 5 soumissions pour la première tâche et 2 pour la deuxième, utilisant toutes des méthodes différentes.

Méthode 1 La première méthode n'utilise que les images et combine de l'apprentissage profond avec des techniques de détection de la fraude pour obtenir plus de 85% de bonnes suppositions sur le corpus d'apprentissage. Les images sont tout d'abord pré-traitées en utilisant une combinaison de méthodes de détection d'images trafiquées :

— analyse du niveau d'erreur
— transformée en ondelette discrète
— niveaux de gris

Ces matrices à trois dimensions, pour les trois méthodes, ont ensuite été fournies au réseau de neurones Resnet152 (He et al. 2016).

Méthode 2 Cette méthode vise à détecter des parties de l'image d'un document qui sont dupliquées, par exemple en cas de modification d'une chaîne de caractères par copier-coller de quelques-uns des caractères. Sur la base de quelques travaux antérieurs appliqués sur des images de scènes naturelles (Fridrich et al. 2003), la méthode développée est basée sur la transformée en cosinus discrète, qui est souvent utilisée dans les algorithmes de compression d'images en raison de sa capacité à projeter une image (ou une partie d'une image) avec d'excellentes propriétés pour regrouper les niveaux d'énergie. Cela permet

par conséquent d'avoir des informations majeures sur seulement quelques coefficients. Cet algorithme utilise cette propriété pour détecter et identifier les zones des images qui possèdent des coefficients similaires, signifiant que l'information est identique.

Méthode 3 L'approche proposée ici se compose de neuf modules de contrôle, chacun concernant un type spécifique de fraude. Ces modules sont basés soit sur le texte soit sur l'image :

- Modules basés sur le texte :
  - Vérification de variation des prix : recherche des prix aberrants
  - Vérification du total à payer : examen des incohérences dans les prix des articles et le montant à payer
  - Vérification de texte manquant : recherche des mots-clés qui impliquent un élément d'information spécifique, mais cette information est manquante
  - Vérification des remises : recherche des incohérences dans les promotions
  - Vérification des quantités : recherche des incohérences dans la formule `quantité` $\times$ `prix de l'article = somme`
  - Vérification des dates : recherche des dates non valides
- Modules basés sur l'image (utilisant OpenCV) :
  - Contrôle des couleurs : recherche de saturation artificielle, de noirceur, ou de « bruits de poivre » (pixels noirs)
  - Vérification des parties effacées : recherche des zones blanches non naturelles ou de grandes zones homogènes (qui n'ont pas de bruit)
  - Vérification des copier-coller : recherche des composantes connexes identiques dans les images binarisées

Chaque module renvoie une valeur de probabilité de fraude comprise entre 0 et 1. La fusion des modules reporte une fraude si la somme de toutes les valeurs est supérieure ou égale à 1. Par conséquent, la fraude est détectée si l'un ou l'autre des modules est très confiant ou si beaucoup de modules ont une petite valeur.

Le bruit dans les données textuelles, venant de la sortie OCR partiellement corrigée, constituait un véritable défi. Une normalisation du texte a été effectuée (suppression des espaces dans les prix, correction des points dans les décimales, etc.) mais ceci pourrait être étendu pour couvrir davantage d'incohérences. Les paramètres ont été réglés manuellement, mais cela pourrait être automatisé à l'avenir.

Méthode 4 Cette méthode utilise trois approches pour détecter les images falsifiées :

- Détection des fraudes de copier-coller et de dessin sur l'image par des techniques de zones denses (*dense-field techniques*) (Cozzolino et al. 2015)
- Empreinte de bruit : il s'agit d'extraire la signature de la camera à travers un réseau profond qui enlève le contenu haut-niveau de l'image (Cozzolino & Verdoliva 2018*b*,*a*). Si une image a été trafiquée, une anomalie peut être découverte en comparant l'empreinte de bruit de cette image avec l'empreinte de référence extraite d'un ensemble d'images authentiques.

— Indices de stéganalyse : cette approche, proposée par Cozzolino et al. (2014), détecte les falsifications grâce à des indices locaux de l'image et à de la classification par des machines à vecteurs de support (SVM) linéaires. Les indices locaux, proposés à l'origine dans des travaux de stéganalyse par Fridrich & Kodovsky (2012), captent des micro-patrons expressifs dans l'image sur laquelle a été appliquée un filtre passe-haut.

Méthode 5 Cette approche est basée sur les caractéristiques stéganographiques extraites à partir de l'image entière et utilisées pour former un ensemble de classificateurs SVM pour faire la distinction entre les images altérées et non altérées. Fridrich & Kodovsky (2012) ont présenté un ensemble de 39 filtres stéganographiques . Cozzolino et al. (2014) appliquent un ensemble de ces filtres sur l'image, et un descripteur de matrice de co-occurrences est formé pour l'image filtrée en entier. Chaque filtre est évalué par validation croisée, et les caractéristiques produites par les filtres les plus performants sont concaténées en un classificateur final.

La méthode 5 suit une approche similaire : une validation croisée est effectuée sur le corpus d'entraînement pour trouver les filtres de Fridrich & Kodovsky (2012) les plus performants pour le corpus, mais ici, les classificateurs individuels sont entraînés pour chaque indice. Le résultat final est obtenu par un vote majoritaire sur toutes les sorties du classificateur. Les filtres stéganographiques de Fridrich & Kodovsky (2012) qui ont démontré les meilleures performances dans la validation croisée sont :

— s5x5 spam14hv q1
— s5x5 minmax22v q1
— s3x3 minmax22v q1
— s3x3 minmax24 q1
— s3 spam14hv q1
— s3 minmax34v q1
— s3 minmax22v q1
— s2 spam12hv q1
— s1 spam14hv q1

Chaque modèle est entraîné en utilisant *bagging*. La sortie de chaque modèle est calculée par la moyenne de toutes les sorties de sacs, et le résultat final est issu d'un vote majoritaire sur l'ensemble des modèles.

Résultats des méthodes candidates

Certains de nos candidats ont utilisé des algorithmes d'apprentissage automatique. Les premiers résultats qu'ils ont soumis étaient basés sur des algorithmes qui avaient été entraînés sur le corpus d'apprentissage de la tâche 1 auquel ils avaient ajouté les documents frauduleux de la tâche 2. Cela correspond à 130 faux documents pour 470 documents authentiques, soit environ 22% du corpus, au lieu de 30 faux documents sur un total de 500 (6% du corpus). Cela n'était pas interdit, mais nous avons demandé aux participants de ré-entraîner leur modèle uniquement sur le corpus d'apprentissage de la première tâche, afin de pouvoir les comparer aux autres de manière plus équitable.

Tableau 5.1 – Résultats des méthodes proposées

Candidats	Précision	Rappel	F-Mesure
Méthode 1	0.364	0.933	0.523
Méthode 2	0.857	0.4	0.545
Méthode 3	0.882	0.5	0.638
Méthode 4 T1	0.906	**0.967**	0.935
Méthode 5 T1	0.964	0.9	0.931
Méthode 5 T1 équilibrée	**1.0**	0.9	**0.947**
Méthode 4 T1+T2	0.935	0.967	0.951
Méthode 5 T1+T2	**1.0**	**1.0**	**1.0**

Le tableau 5.1 présente les résultats obtenus par les différents participants, avec et sans les documents de la tâche 2 inclus dans le processus d'apprentissage. Pour la cinquième méthode, le candidat nous a renvoyé deux résultats : le premier utilise des paramètres identiques à ceux de l'entraînement sur les deux corpus, tandis que le second utilise un équilibrage de classe plus représentatif pendant l'entraînement pour tenir compte de la réduction des échantillons altérés.

La première méthode détecte la plupart des bons documents (très bon rappel) mais classifie également comme faux beaucoup de documents authentiques (mauvaise précision), contrairement à la deuxième méthode qui ne cherche – et trouve – que les documents qui contiennent des fraudes de type duplication de contenu au sein du document.

Le dernier résultat du tableau 5.1 montre un score de détection parfait : la méthode utilisée trouve parfaitement les 30 documents frauduleux. Ce résultat surprenant est certainement dû au fait que le corpus est très spécialisé. En effet, les documents ont tous été numérisés par la même caméra, avec des paramètres presque identiques. Il serait donc intéressant de voir si cette méthode permet d'obtenir des scores équivalents sur un corpus composé d'images provenant de différents appareils photographiques dans différentes conditions d'éclairage et d'inclinaison.

5.1.3 Détection humaine des faux documents

Afin de comparer les résultats de nos méthodes et des méthodes proposées lors de la compétition aux capacités qu'ont les humains pour détecter les faux documents, nous avons demandé à cinq collègues, non-spécialistes de la fraude, mais ayant participé à la session fraude quelques mois auparavant (*cf.* 2.4), de détecter les faux documents sur le corpus de test de la tâche 1. Pour ce faire, une interface web leur fournissait une image de ticket de caisse du corpus de test et ils devaient cliquer sur le bouton « Vrai » ou « Faux », ce qui leur permettait d'afficher l'image suivante, comme montré dans la figure 5.1. Chaque annotateur a traité les 500 images du corpus, proposées dans un ordre aléatoire, en connaissant le taux de documents frauduleux dans ce corpus (6%). Les annotateurs disposaient de plusieurs jours pour traiter l'ensemble du corpus et ont eu un retour sur leurs résultats et ceux des autres annotateurs à mi-parcours. Ils pouvaient donc

FIGURE 5.1 – Interface pour la détection humaine des faux documents

traiter les documents par petites tranches ou d'une seule traite, comme ils le souhaitaient. Afin de rendre la tâche plus agréable, des « récompenses » apparaissaient toutes les vingt images traitées, sous la forme de gifs aléatoires de chat [2], d'articles Wikipédia aléatoire [3], d'articles du Tumblr *Ciel mon doctorat* aléatoires [4] ou des liens de « sites inutiles » toujours aléatoires [5].

Le tableau 5.2 présente leurs scores de précision, rappel et f-mesure. Ces scores montrent qu'il est difficile pour un non-spécialiste humain de détecter un faux document : nous observons beaucoup de faux négatifs, d'où un faible rappel. Ces résultats révèlent également que beaucoup de documents semblent suspects même s'ils sont authentiques : ce sont les faux positifs, qui entrainent une faible précision.

Nous observons que le temps moyen de traitement d'un ticket de caisse dans ces conditions est de 20 secondes par ticket, lorsque les annotateurs sont concentrés sur cette seule tâche. En effet, au-delà de l'inspection rapide pour détecter les anomalies visibles (caractères d'une couleur ou d'une police anormale, traces étranges...), les annotateurs ont vérifié si les informations étaient cohérentes entre elles (sommes des prix correspondant au total et au paiement, bon nombre d'articles affichés, etc) et s'il n'y avait aucune information aberrante.

2. `http://thecatapi.com/api/images/get?format=src&type=gif`
3. `https://fr.wikipedia.org/wiki/Sp%C3%A9cial:Page_au_hasard`
4. `http://cielmondoctorat.tumblr.com/rando`
5. `https://www.uselessweb.com/jump2.php`

Tableau 5.2 – Résultats des détections par des humains

Candidats	Précision	Rappel	F-Mesure
Humain 1	**0.75**	**0.5**	**0.6**
Humain 2	0.64	0.47	0.54
Humain 3	0.69	0.37	0.48
Humain 4	0.55	0.37	0.44
Humain 5	0.45	0.33	0.38

Sur les 500 documents du corpus, les annotateurs ne sont pas en accord sur l'authenticité de 49 d'entre eux. De plus, ils se trompent tous sur la classification de 9 autres de ces 500 reçus (tous faux négatifs, c'est-à-dire faux documents non détectés). Nous avons également calculé le Kappa de Fleiss. Cette mesure est utilisée pour calculer l'accord inter annotateur et se situe dans un intervalle de 0 à 1 ou inférieur à 0 si les annotateurs ne sont pas du tout d'accord. Ici, $\kappa = 0,4375$, ce qui montre que les quatre annotateurs ne sont que modérément d'accord sur les fraudes qu'ils détectent. En d'autres termes, certains voient la fraude là où d'autres ne la voient pas.

Une analyse plus détaillée des résultats des candidats de la compétition montre que les erreurs de classement ne concernent pas les mêmes documents. En effet, tous les faux documents ont été détectés par au moins deux méthodes. Les faux positifs n'ont été détectés que par une seule méthode à chaque fois. En tout, ce sont 81 documents qui ont été mal classés : 70 par un seul candidat, 8 par 2 candidats et 3 par 3 candidats. Les trois derniers ont également été mal classés par 5, 4 et 2 humains respectivement. Ils contiennent les types de fraude CPI+CUT pour deux d'entre eux et CPI+IMI pour le dernier. On constate également que 5 faux documents sont parfaitement détectés comme faux par les 5 méthodes automatiques, alors que l'un d'entre eux ne l'est pas par 4 des 5 humains.

5.2 Approches proposées

Nous allons dans cette nouvelle section présenter nos propres approches pour détecter les faux documents. Nous avons séparé nos approches pour détecter les faux documents en trois types : la vérification interne au document, la vérification entre les documents d'un même corpus, et la vérification externe. D'abord, nous cherchons à vérifier que les informations contenues dans un même document sont cohérentes les unes avec les autres. En effet, nous avons remarqué que les fraudeurs n'étaient pas toujours très rigoureux et oubliaient régulièrement de modifier toutes les informations en conséquences. Par exemple, si le total est modifié, le montant payé devrait l'être également. Dans un deuxième temps, nous cherchons à vérifier la constance des informations au sein du corpus, c'est-à-dire le fait que les informations d'un document ne sont pas aberrantes par rapport à celles des autres documents. Par exemple, nous pouvons vérifier qu'un produit a toujours le même prix, ou que la variation est minime. Pour finir, nous proposons de chercher à estimer la vraisemblance des informations en les comparant à des informations externes au corpus.

Nous présenterons dans cette section les deux premières approches, en évaluant la pertinence des indices fournis, et nous étudierons plusieurs possibilités pour combiner ces deux types d'indices dans la troisième partie. Le troisième type d'approche sera présenté dans la section suivante.

5.2.1 Vérification interne au document

En observant les fraudes effectuées dans le cadre de la session fraude décrite dans la section 2.4, nous avons constaté que beaucoup de tickets falsifiés présentaient des incohérences. En effet, nous avons analysé 75 tickets fraudés sur les 250 obtenus et nous avons constaté que 45 d'entre eux avaient des informations incohérentes. 13 d'entre eux possèdent par exemple un mauvais total : un produit a été changé, ou un prix, et la somme n'a pas été recalculée pour que le total soit cohérent avec la somme des prix. De même, 6 tickets présentent une incohérence entre le montant total et le paiement effectué. Ces erreurs dans les fraudes ne sont donc pas anodines et ces informations méritent d'être vérifiées.

Nous avons donc proposé une liste de sept indices, qui sont tous binaires : ils retournent 0 si la condition est satisfaite, 1 si la condition ne l'est pas.

$\sum$prix = total Nous calculons d'abord la somme des prix de chaque produit du ticket et nous la comparons au montant total que nous avons extrait. En effet, il est fréquent dans notre corpus que les fraudeurs aient oublié de modifier les totaux en conséquence du changement des prix des produits.

$\sum$articles = Nombre articles Le deuxième indice porte sur le nombre extrait d'articles, comparé au nombre calculé de produits extraits. Il est possible que les fraudeurs ne fassent pas attention à ce genre de détails quand ils suppriment ou ajoutent un produit. La suppression de produit peut par exemple avoir lieu dans le cas de remboursement de frais de mission, où il n'est pas forcément bien vu de prendre certaines boissons alcoolisées. Cette information n'est cependant pas relevée sur tous les tickets de caisse, ce qui implique que cet indice ne pourra être pertinent que pour les tickets Carrefour.

Total = paiement Le troisième indice concerne le montant payé extrait et le montant total extrait. En effet, nous avons remarqué dans notre corpus qu'il est fréquent que les fraudeurs oublient de reporter le montant total fraudé sur le montant payé, probablement parce que les abréviations utilisées rendent difficile l'identification de cette information pour les profanes, qui ne voient pas dans des « CB EMV », « TR » et autres « ESP » la signification que notre outil, lui, peut relever.

Quantité × prix unitaire = prix Sur les tickets de Carrefour, nous relevons la quantité et le prix unitaire des produits. Certaines fraudes portent sur ces informations et nous cherchons donc à vérifier que les fraudeurs n'ont pas fait d'erreurs dans la multiplication

de ces deux informations : le prix total du produit doit en effet correspondre au résultat de cette multiplication.

Poids × prix au kilogramme = prix Nous vérifions également, dans le cinquième indice, l'égalité du résultat de la multiplication du poids d'un produit par son prix au kilogramme et du prix total, quand ces informations existent et sont relevées.

Date et heure Le sixième indice porte sur la vérification du format de la date et de l'heure : si une date ou une heure ne correspondent pas à une date ou à une heure possible, une propriété `est_suspect` est ajoutée à l'ontologie pour le concept `TicketCaisse` lors de l'extraction. Par exemple si le numéro du mois est supérieur à 12, ou si le fraudeur a oublié qu'il n'y a que 28 ou 29 jours en février, la date ne peut pas être entrée dans l'ontologie car elle ne correspond pas au format attendu. Elle est donc rentrée en tant que chaîne de caractères.

Loi de Benford Le dernier indice cherche à vérifier la « loi de Benford » sur les prix des produits. La loi généralisée de Benford établit que la distribution de certains chiffres dans de nombreuses séquences de chiffres de la vie réelle ne suit pas une distribution uniforme. Cela signifie que, dans un ensemble de données sur n'importe quel domaine (comptabilité, démographie, presse, articles scientifiques...), il est fréquent qu'un chiffre soit sur-représenté par rapport aux autres (Nigrini 2012). D'après Durtschi et al. (2004), cette loi est souvent utilisée pour détecter les fraudes, avec plus ou moins de résultats, lors d'audits comptables. Le constat est que la répartition du premier chiffre significatif d'un nombre est logarithmique : il y a ainsi plus de nombres commençant par 1 que par 2, plus par 2 que par 3... Nous avons donc testé cette loi sur chacun des tickets de notre corpus comme septième indice, ainsi que sur l'ensemble de notre corpus (voir l'annexe C).

Pertinence des indices

Le tableau 5.3 présente les résultats des indices de vérification interne sur les documents du corpus d'apprentissage que nous avons séparés en fonction de leur provenance. En effet, étant donné la difficulté de collecter certaines informations sur des tickets de caisse d'origines diverses, certains indices ont clairement moins d'impacts voire d'intérêts sur les documents ne venant pas de Carrefour. Nous avons donc décidé d'appliquer nos indices séparément sur les tickets provenant de Carrefour et sur les autres, pour lesquels nous avons extrait plus d'informations que pour les autres tickets. Nous testons également les indices sur l'ensemble du corpus d'apprentissage.

Les chiffres en gras sont les meilleurs scores de chaque métrique pour chaque sous-corpus. Nous pouvons constater que le taux de bonne classification (l'*accuracy*) ne représente pas grand-chose au vu des autres résultats. En effet, cette métrique prend en compte la bonne classification de tous les documents. Elle compte donc tous les biens classés (les « vrais »), qu'ils soient « positifs » ou « négatifs », c'est-à-dire qu'ils soient modifiés ou authentiques. Comme notre corpus contient seulement 6% de faux documents, on peut

Tableau 5.3 – Évaluation des indices de vérification interne sur les documents du corpus d'apprentissage

Provenance	Indice	Précision	Rappel	F-mesure	*Accuracy*
Carrefour	$\sum$prix	0,20	**0,81**	0,33	0,78
	NbArticles	0,14	0,56	0,23	0,75
	Paiement	0,45	0,31	**0,37**	0,93
	Quantité	**0,50**	0,06	0,11	**0,94**
	Poids	0,40	0,13	0,19	0,93
	DateHeure	0,00	0,00	0,00	0,94
	Benford	0,07	0,69	0,12	0,37
Autres	$\sum$prix	0,05	0,57	0,09	0,34
	NbArticles	0,00	0,00	0,00	0,88
	Paiement	0,05	0,57	0,10	0,42
	Quantité	0,00	0,00	0,00	**0,94**
	Poids	0,00	0,00	0,00	**0,94**
	DateHeure	**0,50**	0,07	**0,13**	**0,94**
	Benford	0,07	**0,71**	0,12	0,41
Ensemble	$\sum$prix	0,09	**0,70**	**0,16**	0,56
	NbArticles	0,09	0,17	0,12	0,85
	Paiement	0,08	0,43	0,14	0,68
	Quantité	**0,50**	0,03	0,06	**0,94**
	Poids	0,40	0,07	0,11	**0,94**
	DateHeure	**0,50**	0,03	0,06	**0,94**
	Benford	0,07	**0,70**	0,12	0,39

obtenir un taux de bonne classification de 94% en ne trouvant aucun faux document ou très peu, autrement dit en classant tous les documents ou presque comme vrais. C'est ce que l'on peut constater avec l'indice DateHeure pour les tickets Carrefour (détection d'une date ou heure suspecte), ou avec les indices Quantité (quantité × prix unitaire), Poids (poids × prix au kilogramme) et DateHeure pour les autres tickets.

Nous remarquons cependant que l'indice DateHeure a également la meilleure f-mesure sur le sous-corpus des autres documents : cela est dû au fait que la moitié des documents relevés par cet indice se révèlent réellement falsifiés. Il n'y a cependant pas beaucoup d'occurrences de cette fraude dans le corpus, ce qui explique le faible rappel sur ce sous-corpus, et le rappel nul sur le sous-corpus Carrefour qui crée une précision incalculable (que nous avons noté « 0 » par commodité).

Nous pouvons également constater que l'indice NbArticles retourne des résultats nuls pour le deuxième sous-corpus. Cela vient du fait que le nombre d'articles est difficilement extrait sur ces documents (voir chapitre 3). Cet indice n'est donc pas pertinent pour ces documents bien qu'il soit plutôt intéressant pour les tickets Carrefour. De même, les indices Quantité et Poids sont très spécifiques au sous-corpus Carrefour, car nos règles d'extraction se sont concentrées sur les formats d'écriture de ces tickets. Ces trois indices

ne sont donc pas pertinents pour le sous-corpus Autres.

Dans les deux sous-corpus, les trois indices les plus intéressants d'un point de vue rappel sont les indices $\sum$prix, Paiement et Benford. En effet, la majorité des faux documents sont relevés par ces indices. Cependant, de nombreux documents authentiques sont également relevés, ce qui rend l'indice très peu discriminant et la précision parfois très faible, surtout sur le sous-corpus des autres magasins.

5.2.2 Vérification intra-corpus

Nous avons également constaté dans notre corpus de faux documents que les informations n'étaient pas toujours similaires aux autres informations semblables se trouvant sur d'autres documents. Par exemple, un même produit peut avoir des prix différents. Parfois il s'agit seulement d'une variation minime, parfois il s'agit clairement d'une fraude. C'est le cas par exemple du produit « NOUILLES THAI » que l'on trouve à 14.11€ sur un ticket (fraudé), et à 4.11€ sur deux autres (authentiques). Nous avons donc levé une alarme si le prix unitaire du produit est différent de la moyenne des montants des autres instances du produit (indice PrixMoy), une autre si le prix unitaire est plus de 10% supérieur ou inférieur à la moyenne des montants des autres instances du produit (indice PrixMoy10) et une troisième si le prix n'est pas dans la liste des montants des autres instances du produit (indice PrixListe).

Le deuxième indice a pour objectif de limiter la détection des faibles variations de prix levées par la première (et ainsi augmenter la précision). Ces faibles variations peuvent être dues, entre autres, à des re-négociations des prix entre producteurs et vendeurs, à la variation des cours des bourses et des marchés pour certains produits ou à l'inflation générale. Nous avons exclu les produits qui contiennent une information sur le poids et le prix au kilogramme, car le prix unitaire varie alors trop fortement. Nous n'avons pas pris en compte non plus le prix au kilogramme car la fluctuation des cours est trop importante sur les fruits et légumes selon les saisons et les années.

S'il est facile de comparer des informations chiffrées, comme des prix, de façon statistique, ça l'est beaucoup moins avec des informations textuelles. Nous avons pourtant tenté de comparer divers couples d'autres informations, comme les différentes adresses relevées pour une même entreprise, les numéros de téléphone, les numéros de SIREN, ou bien les noms d'entreprises qui partagent ces mêmes informations. Pour chaque ticket de caisse de notre sous-corpus, nous récupérons donc d'abord ces quatre informations dans notre ontologie.

Nous avons remarqué dans notre corpus que certains noms d'entreprise avaient été modifiés, mais les adresses sauvegardées. Pour les détecter, nous extrayons la liste des noms d'entreprises qui ont la même adresse que le ticket examiné. Dans un monde idéal, cette liste ne devrait être composée que d'un seul nom d'entreprise, vu qu'une adresse devrait être suffisamment précise pour correspondre à un seul endroit, et qu'une entreprise ne devrait avoir qu'un seul nom. C'est pourtant loin d'être le cas dans notre corpus, comme le révèlent les figures 5.2 et 5.3 qui montrent les résultats des inférences proposées par le raisonneur HermiT1.3.8.413 dans Protégé5.5.0 respectivement pour l'instance `Chloé city`, qui est une fraude, et l'instance `C_City_CRF-CITY_LA_ROCHELLE`, qui est

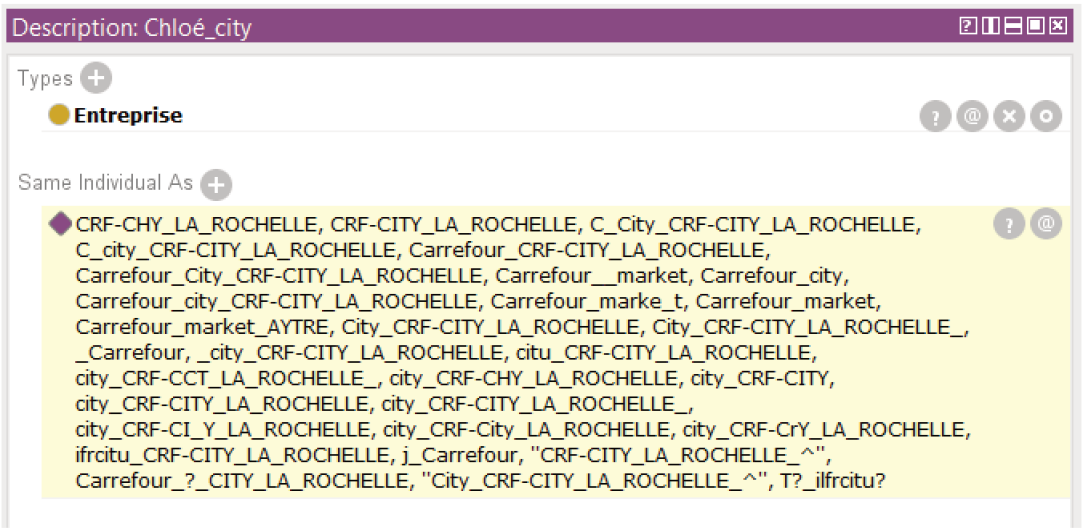


FIGURE 5.2 – Résultats des inférences proposées par le raisonneur HermiT1.3.8.413 dans Protégé5.5.0 pour l'individu `Chloé city`

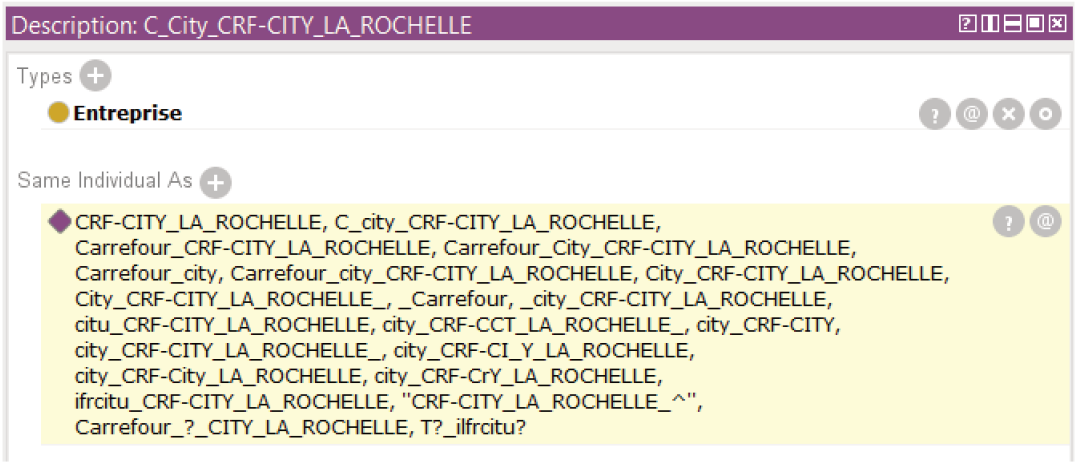


FIGURE 5.3 – Résultats des inférences proposées par le raisonneur HermiT1.3.8.413 dans Protégé5.5.0 pour l'individu `C_City_CRF-CITY_LA_ROCHELLE`

une des multiples variations de « Carrefour city ». Nous pouvons effectivement voir dans ces images que pour une même entreprise (Carrefour city en l'occurrence), nous avons plus d'une vingtaine de variations, ou synonymes, classés comme `Same Individual As`, c'est-à-dire comme « équivalent à » par le raisonneur. La majorité des variations est due aux erreurs non corrigées de l'OCR et à l'hétérogénéité des choix sur la transcription ou non des caractères inscrits dans les logos. De même, la présence d'un « C » en début de nom vient de la partie graphique du logo Carrefour, qui représente un C blanc inscrit dans un losange noir (ou rouge et bleu ailleurs que sur des tickets de caisse).

Une autre variation vient du fait que le logo des magasins Carrefour city a changé au cours de notre collecte, affectant ainsi l'extraction du nom du magasin : il est donc fréquent que le nom avant le 17/11/2016 soit « Carrefour city », et qu'il devienne « city CRF-CITY LA ROCHELLE » par la suite, comme nous pouvons le voir dans le tableau 5.4. Nous pouvons d'ailleurs remarquer dans notre corpus que le logo-nom « Carrefour city » est plus souvent transcrit que le logo-nom « city »

Tableau 5.4 – Variation temporelle du logo, et donc du nom, du magasin

Nous voyons également dans les figures 5.2 et 5.3 que le raisonneur considère certains Carrefour market comme équivalents à « Chloé city », ce qui est une erreur qui est peut-être due à la volonté de notre part de ne pas spécifier toutes les relations de disjointure et d'unicité dans notre ontologie concernant les magasins (justement pour des raisons de variations nominales), ou bien au fait que certains produits sont vendus à la fois par Carrefour city et par Carrefour market, ce qui pourrait marquer une proximité conceptuelle entre les deux entreprises.

A cause de toutes ces variations, l'indice qui nous servirait à détecter des falsifications de noms d'entreprise doit être adapté. Nous comparons deux à deux les noms d'entreprises qui ont la même adresse dans les indices « Entreprise/adresse 1 » et « Entreprise/adresse 2 », et le même numéro de téléphone pour l'indice « Entreprise/téléphone ». Dans ces trois indices, si le nom de l'une des entreprises se retrouve dans le nom de l'autre, alors nous considérons qu'il n'y a pas de fraudes. Sinon, nous calculons la distance de Levenshtein entre les deux noms d'entreprises comparés : si la distance est supérieure à 20 et que la date de l'un ou de l'autre n'est pas inférieure au 17 novembre 2016, alors nous levons une alarme. Après avoir comparé deux à deux toutes les entreprises ayant les mêmes coordonnées, nous calculons l'indice « Entreprise/adresse 1 » qui indique « vrai » pour le ticket si aucune alarme est levée, c'est-à-dire s'il n'y a pas de nom d'entreprise qui soit à une distance de Levenshtein supérieure à 20. L'indice « Entreprise/adresse 2 » indique « faux » si au moins quatre alarmes sont levées. L'indice « Entreprise/téléphone » est le même que l'indice « Entreprise/adresse 1 » mais se base sur un numéro de téléphone identique pour déterminer qu'une entreprise devrait être identique.

Les trois indices suivants se basent sur le nom de l'entreprise comme dénominateur commun pour comparer les tickets : il s'agit alors de vérifier que l'adresse (indice « Adresse/entreprise »), le numéro de téléphone (« Téléphone/entreprise ») et le numéro de SIREN (« SIREN/entreprise ») sont les mêmes entre le ticket à vérifier et les tickets

émis par la même entreprise. Pour cela, nous levons une alarme si la distance de Levenshtein est supérieure à un seuil (6 pour les adresses, 3 pour les numéros de téléphone et de SIREN) afin d'éviter de lever des fausses alarmes pour des simples erreurs de reconnaissance de caractères, et nous prédisons le document comme étant « faux » si une alarme est levée. Ses seuils sont fixés de façon empirique : si 3 numéros diffèrent sur les 9 du numéro de SIREN ou les 10 du numéro de téléphone, nous considérons que l'erreur est trop grande pour qu 'il s'agisse d'une erreur liée à l'OCR. Ses seuils pourraient cependant être réévalués par apprentissage.

Pertinence des indices

Tableau 5.5 – Évaluation des indices de vérification inter-documents sur les documents du corpus d'apprentissage

Provenance	Indice	Précision	Rappel	F-mesure	*Accuracy*
Carrefour	PrixMoyenne	0,09	**0,75**	0,16	0,51
	PrixMoyenne10	0,13	0,69	0,22	0,70
	PrixListe	0,22	0,31	**0,26**	0,88
	Entreprise/adresse1	0,21	0,25	0,23	0,89
	Entreprise/adresse2	**0,67**	0,13	0,21	**0,94**
	Entreprise/téléphone	0,10	0,06	0,08	0,90
	Adresse/entreprise	0,04	0,19	0,06	0,65
	Téléphone/entreprise	0,05	0,63	0,09	0,23
	SIREN/entreprise	0,00	0,00	0,00	**0,94**
Autres	PrixMoyenne	0,08	**0,50**	0,14	0,65
	PrixMoyenne10	**0,09**	**0,50**	**0,15**	0,68
	PrixListe	0,08	**0,50**	0,13	0,64
	Entreprise/adresse1	0,06	0,43	0,10	0,56
	Entreprise/adresse2	0,00	0,00	0,00	**0,94**
	Entreprise/téléphone	0,03	0,14	0,05	0,67
	Adresse/entreprise	0,00	0,00	0,00	0,88
	Téléphone/entreprise	0,00	0,00	0,00	0,84
	SIREN/entreprise	0,00	0,00	0,00	**0,94**
Ensemble	PrixMoyenne	0,09	**0,63**	0,15	0,57
	PrixMoyenne10	0,10	0,60	**0,18**	0,67
	PrixListe	0,10	0,37	0,16	0,77
	Entreprise/adresse1	0,08	0,33	0,13	0,73
	Entreprise/adresse2	**0,67**	0,07	0,12	**0,94**
	Entreprise/téléphone	0,04	0,10	0,05	0,79
	Adresse/entreprise	0,03	0,10	0,05	0,76
	Téléphone/entreprise	0,04	0,33	0,08	0,53
	SIREN/entreprise	0,00	0,00	0,00	0,94

Le tableau 5.5 présente les résultats des différents indices sur le corpus d'apprentissage, et ses deux sous-corpus. Nous pouvons tout d'abord voir que, pour les deux sous-corpus, les trois premiers indices présentent des résultats intéressants, même si la précision n'est pas toujours très élevée. Les prix, sans grande surprise, sont donc des informations à regarder de près lorsque l'on cherche des documents falsifiés. Nous pouvons constater que lorsque l'on augmente l'inégalité des prix par rapport à la moyenne des autres prix de produits similaires, la précision augmente, même si nous perdons légèrement en rappel, ce qui signifie que nous avons éliminé quelques faux positifs qui étaient certainement des variations classiques des prix.

Nous pouvons remarquer que dans le cas des tickets provenant de Carrefour, l'indice concernant la détection des variantes des noms d'entreprise avec strictement plus de deux alarmes est très précis par rapport aux autres indices. Toujours dans ce sous-corpus, le seul indice qui ne relève rien est celui qui porte sur le numéro de SIREN, ce qui est normal car cette information n'apparait pas sur les tickets Carrefour. Tous les autres indices relèvent des faux documents, et parfois nombreux, mais relèvent aussi comme « Faux » beaucoup de documents authentiques (faible précision). On peut remarquer le rappel élevé pour l'indice « Téléphone/entreprise », qui est directement dû à la très mauvaise précision de l'extraction des numéros de téléphone (voir chapitre 3).

La comparaison des noms d'entreprises par les adresses et numéros de téléphone communs donne également des résultats intéressants. Sur les tickets des autres commerces en revanche, les indices sur la comparaison des coordonnées par rapport aux noms d'entreprises identiques donnent des résultats nuls. Cela vient certainement du fait qu'il y a peu de documents émis par chaque entité, ce qui limite les possibilités de vérification au sein du corpus. Par ailleurs, certaines adresses ne sont pas assez précises et sont alors attribuées à plusieurs entreprises totalement différentes. C'est le cas des Z.A.C. par exemple, Zone d'Activités Commerciales, qui sont des grands espaces commerciaux où se regroupent de nombreuses grandes enseignes. C'est également le cas des gares, qui possèdent de plus en plus de galeries marchandes, ou des aires d'autoroutes. Ainsi, près de la moitié des adresses que nous avons extraites n'ont pas de numéro de rue, et sont donc potentiellement partagées avec d'autres entreprises.

5.2.3 Combinaison des approches et apprentissage

Les résultats des indices seuls, sur chacun des deux sous-corpus, ne dépassent pas les 37% de f-mesure. Ce score, bien que déjà significatif pour un seul indice, ne suffit pas à détecter les faux documents de manière satisfaisante. Nous avons donc cherché comment les assembler, les combiner, afin d'optimiser les chances de trouver les faux documents.

Les stratégies des participants à la compétition pour combiner les résultats des différents indices utilisés sont variés : apprentissage automatique avec des SVM, seuils sur la somme des valeurs produites par les indices, système de vote majoritaire... Toutes ces méthodes mériteraient d'être approfondies d'un point de vue théorique, mais pour des raisons de temps, nous ne présentons ici que quelques approches de combinaison et fusion des indices pour obtenir le meilleur système de détection des faux documents.

Cette question de combinaison des indices est cruciale dans notre problème : devant un juge, il faut apporter un faisceau de preuves, un ensemble d'indices qui prouve que le suspect est coupable. Parfois, une seule preuve suffit, mais il faut qu'elle soit indéniable. Dans notre corpus, la pertinence de chaque indice est difficile à évaluer, du fait de la faible représentativité de la fraude qu'il peut révéler. Ce sont donc dans cette section différentes approches que nous testons, tout en gardant à l'esprit que notre échantillon n'est certainement pas assez grand pour en apprendre des modèles génériques.

Nombres d'indices convergents

Tout d'abord, nous calculons les métriques habituelles sur la somme des valeurs binaires fournies par chacun des indices, en fixant un seuil qui détermine la justesse de la classification sur la classe « faux document ». Ainsi, si le seuil est de 1, cela signifie qu'il faut au minimum une alarme pour que le document soit prédit comme faux. Les « positifs » sont les documents prédits faux documents, les « vrais » sont les documents bien prédits. La figure 5.4 illustre les scores de précision, rappel et f-mesure pour les différents seuils à partir desquels le document est prédit comme faux, sur l'ensemble du sous-corpus. La meilleure f-mesure est située à 5 et est de 0,20, ce qui ne semble pas très élevé, et donc pertinent pour obtenir de très bons résultats. Sur les deux sous-corpus, le meilleur seuil est respectivement de 7 (F = 0,32) pour les documents Carrefour et de 5 (F = 0,14) pour les autres tickets.

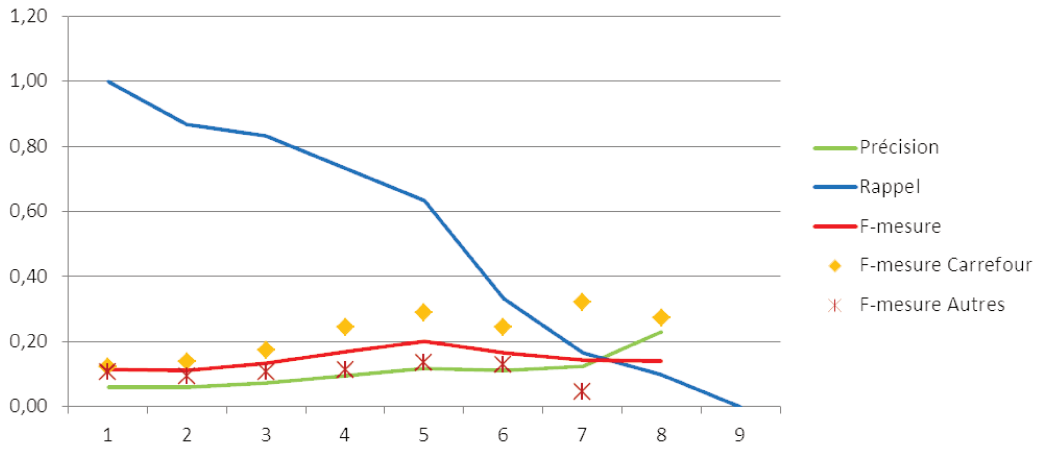


FIGURE 5.4 – Scores des seuils de 1 à 9 sur la somme des indices

Nous pouvons constater qu'à partir de 9 indices, le nombre de vrais positifs est nul, ce qui explique l'arrêt des courbes de précision et de rappel. Par ailleurs, plus le nombre d'indices nécessaire à la prédiction des faux documents est élevé, plus la précision est élevée car on est sûr que ce sont des faux documents. Cependant, dans le même temps, le rappel diminue fortement, car de moins en moins de documents sont concernés par un nombre d'indices élevé. Le vote majoritaire se situerait à la moitié de 16 indices,

c'est-à-dire qu'il faudrait au minimum 8 indices pour déclarer un document comme étant faux. Si cela signifie sur le premier sous-corpus qu'un document sur 2 déclaré faux est bien faux, ce qui est le meilleur score obtenu par la méthode de seuillage, le rappel n'est pour autant pas très bon, voire nul dans le second sous-corpus. Il semblerait en effet que certains de nos indices soient trop spécifiques, trop précis, pour être efficaces pour repérer beaucoup de documents.

La figure 5.5 montre les résultats des métriques pour chaque nombre d'indices précis sur l'ensemble du corpus d'apprentissage. La meilleure f-mesure est calculée sur un total d'indices égal à 5, et se situe à 0,18. Les mêmes indices sur le sous-corpus Carrefour indiquent que si nous avons exactement 8 indices qui disent que le document est faux, alors ces huit indices ont raison (précision à 100%). Par contre, environ 4 faux documents sur cinq ne sont pas repérés. Pour les documents des autres commerces, c'est quand il y a 6 indices que la précision et le rappel sont meilleurs.

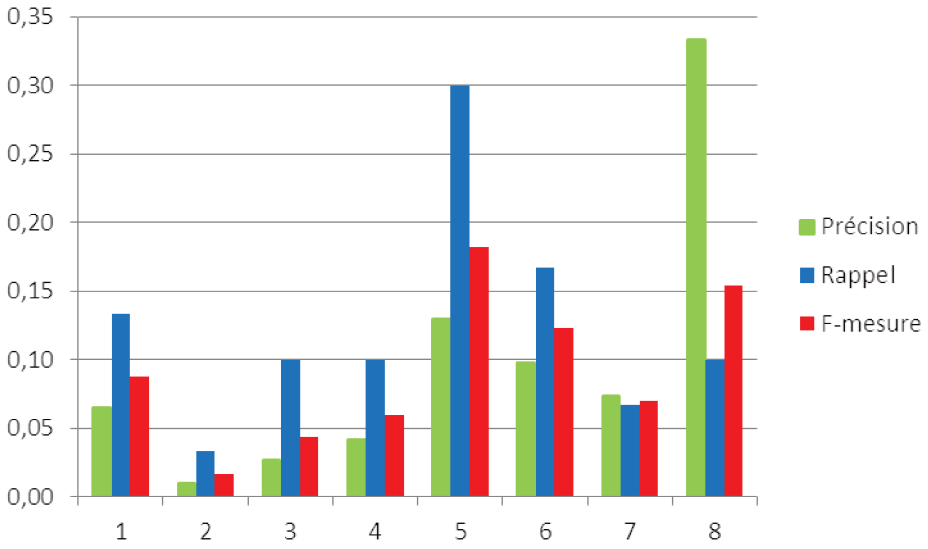


FIGURE 5.5 – Scores des nombres d'indices de 1 à 8

Combinaison d'indices

Nous avons ensuite cherché la meilleure combinaison d'indices possible sur les deux sous-corpus d'apprentissage. Pour cela, nous avons regardé plusieurs combinaisons de scores selon plusieurs critères pour déterminer si un document doit être prédit comme faux. Ces critères sont :

— 1 ou 2 indices parmi 2 ($i_1 + i_2 \geq 1 \Leftrightarrow \mathcal{P}(doc) = \text{Faux}$)
— 2 indices parmi 2 ($i_1 + i_2 = 2 \Leftrightarrow \mathcal{P}(doc) = \text{Faux}$)
— 2 ou 3 indices parmi 3 ($i_1 + i_2 + i_3 \geq 2 \Leftrightarrow \mathcal{P}(doc) = \text{Faux}$)
— 3 indices parmi 3 ($i_1 + i_2 + i_3 = 3 \Leftrightarrow \mathcal{P}(doc) = \text{Faux}$)

— 3 ou 4 indices parmi 4 ($i_1 + i_2 + i_3 + i_4 \geq 3 \Leftrightarrow \mathcal{P}(doc) = \text{Faux}$)
— 4 indices parmi 4 ($i_1 + i_2 + i_3 + i_4 = 4 \Leftrightarrow \mathcal{P}(doc) = \text{Faux}$)

Avec 16 indices à tester selon ces 6 possibilités, nous obtenons 5000 combinaisons possibles.

Sur le sous-corpus Carrefour, le meilleur score en terme de f-mesure est obtenu par une combinaison des 3 ou 4 indices parmi $\sum$prix, Paiement, Benford et PrixListe. Cette combinaison obtient 0,58 de précision, 0,44 de rappel et 0,50 de f-mesure. Trente-trois autres combinaisons obtiennent des f-mesures comprises entre 0,40 et 0,50.

Deux combinaisons obtiennent un rappel de 100%, composées de Benford et/ou respectivement de PrixMoyenne et PrixMoyenne10, avec des précisions de 0,08, mais, comme nous l'avons déjà vu, avoir un fort rappel et une très faible précision n'est pas très utile : cela limite peut-être le périmètre des recherches de faux documents, mais de très peu.

La précision est à 100% pour 258 combinaisons. Elles concernent deux faux documents pour 15 cas, et un seul pour les 243 autres. Plusieurs combinaisons peuvent concerner le même document : ce n'est pas parce que deux indices sont positifs pour la détection que deux autres ne le sont pas. De même, les combinaisons de type « ou » englobent les combinaisons de nombres d'indices inférieurs. Ainsi, si i_1, i_2 et $+i_3$ valent tous 1, les conditions $i_1 + i_2 = 2$, $i_1 + i_3 = 2$, $i_2 + i_3 = 2$, $i_1 + i_2 + i_3 = 3$ et $i_1 + i_2 + i_3 \geq 2$ sont respectées.

Sur l'autre sous-corpus, la meilleure f-mesure est significativement moins bonne, comme nous pouvions le deviner au vu des autres résultats. C'est la combinaison des trois ou quatre indices parmi DateHeure, Benford, PrixMoyenne10 et PrixListe. Là encore, ce n'est pas très étonnant, étant donné que ces quatre indices sont parmi ceux qui obtiennent les meilleures f-mesures sur ce sous-corpus. Les résultats de cette combinaison sont donc : 0,12 de précision, 0,43 de rappel et 0,18 de f-mesure. Dix-neuf autres combinaisons ont plus de 0.16 de f-mesure.

Sur ce sous-corpus, aucune combinaison n'atteint les 100% de rappel. Le meilleur rappel est obtenu par une combinaison de 1 ou 2 indices parmi Benford (qui obtient 0,71 seul) et Entreprise/adresse1 (0,43 seul) et atteint 0,93. Cela signifie qu'il n'y a qu'un seul faux document qui ne lève aucun des deux indices. Cependant, beaucoup d'autres documents (authentiques donc) sont également relevés par cette combinaison d'indices.

La précision est à 100% pour 77 combinaisons sur ce sous-corpus mais ne concerne à chaque fois qu'un seul faux document (rappel = 0,07). Nous pouvons noter également que 3972 combinaisons ne relèvent pas de faux documents sur ce sous-corpus, ce qui renforce l'analyse faite précédemment de l'inutilité de nombreux indices sur ce sous-corpus, probablement due au fait que nous n'avons pas extrait les informations vérifiées, ou qu'elles n'étaient pas présentes sur le ticket. Sur le sous-corpus Carrefour, ce sont 2954 combinaisons qui retournent des résultats nuls.

Pondération des indices

Nous avons également souhaité calculer les résultats de la détection des faux documents si, au lieu de fournir des valeurs entières, on fournissait des valeurs pondérées.

Nous avons donc affecté à chaque indice la valeur de sa précision sur l'ensemble du corpus, puis sur le sous-corpus donné. Ainsi, si la précision d'un indice sur le (sous-)corpus est de 0,75, le score de l'indice pour chaque document ne sera plus 0 ou 1, mais 0 ou 0,75. Cela permet de favoriser les indices qui sont vraiment discriminants, qui sont plus sûrs de détecter des faux documents, par rapport à ceux qui détectent de tout.

La figure 5.6 présente, pour le sous-corpus Carrefour, les courbes de précision, rappel et f-mesure en fonction des seuils minimaux pour lesquels la somme des valeurs pondérées est supérieure ou égale. Ainsi, si la somme est supérieure ou égale à 0,7, la f-mesure est 0,57, le rappel de 0,75 et la précision, de 0,46. La pondération des indices est, pour l'instant, la meilleure optimisation que nous avons testée, devant la meilleure combinaison.

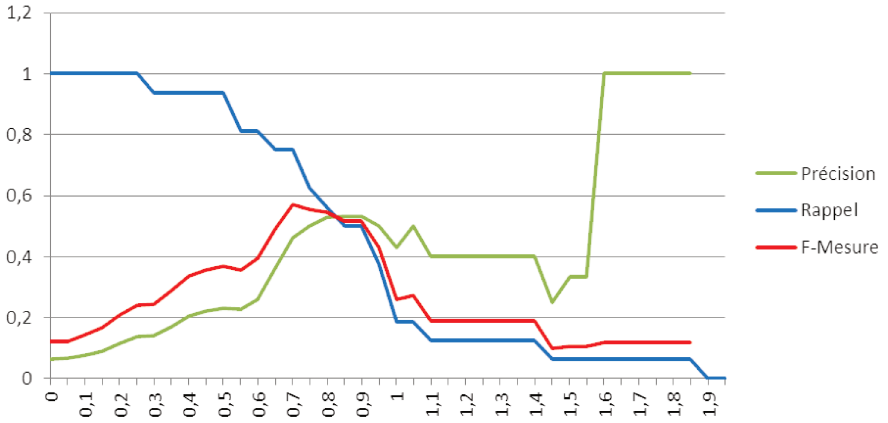


FIGURE 5.6 – Graphique des métriques calculées sur la somme des valeurs pondérées des indices en fonction des seuils de 0 à 1,9 sur le sous-corpus Carrefour

Pour le second sous-corpus, la meilleure f-mesure, avec les valeurs de pondération adaptées au sous-corpus, se situe à un seuil minimal de 0,35, et est de 0,17, avec une précision de 0,10 et un rappel de 0,50, ce qui n'est pas une forte amélioration étant donné que le meilleur indice seul donne déjà une f-mesure de 0,15 et la meilleure combinaison, une f-mesure de 0,18. Nous pouvons remarquer que les sommes des pondérations, sur l'axe des abscisses, sont beaucoup moins importantes sur ce sous-corpus que sur le premier. Cela est dû aux précisions plus faibles de chaque indice (voir les tableaux 5.3 et 5.5) sur ce corpus, et au fait que chaque faux document soit relevé par moins d'indices que les faux documents du corpus Carrefour (voir 5.4).

Sur l'ensemble du corpus d'apprentissage, la meilleure f-mesure est de 0,28, pour un seuil de 0,60. Ainsi, si la somme des indices pondérés est supérieure ou égale à 0,60, on récupère 30% des faux documents et on a 26% de chance que le document soit bien un faux.

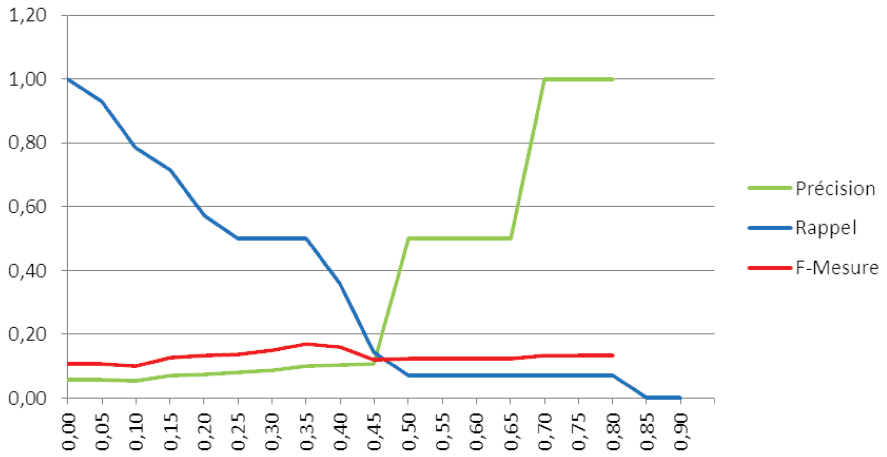


FIGURE 5.7 – Graphique des métriques calculées sur la somme des valeurs pondérées des indices en fonction des seuils de 0 à 1,9 sur le sous-corpus Autres

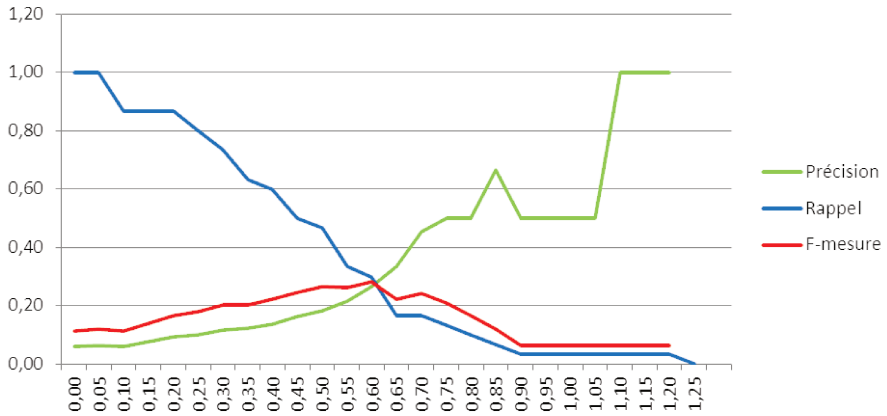


FIGURE 5.8 – Graphique des métriques calculées sur la somme des valeurs pondérées des indices en fonction des seuils de 0 à 1,9 sur l'ensemble du corpus d'apprentissage

Apprentissage automatique

Afin d'optimiser au mieux nos indices, nous avons décidé d'entrainer des classifieurs implémentés dans le logiciel d'apprentissage automatique Weka (Eibe et al. 2016). Ce « poste de travail » contient de nombreux algorithmes d'apprentissage automatique et de traitement de données et propose une interface simple d'utilisation qui permet d'apprendre et de tester rapidement différents indices sur n'importe quel jeu de données.

Les algorithmes d'apprentissage cherchent à optimiser l'*accuracy*, c'est-à-dire à rendre la classification générale la meilleure possible : la majorité des documents à classer doit être classée au mieux. Or dans notre cas, le problème n'est pas tant la classification de

l'ensemble des documents que la bonne classification des faux documents. En d'autres termes, le taux de bonne classification, comme nous l'avons vu précédemment, ne nous intéresse pas beaucoup étant donné qu'elle peut être élevée (94% de bonne classification, puisqu'il y a 94% de documents authentiques) sans qu'on détecte un seul faux document.

Nous devons donc utiliser un des algorithmes de meta-apprentissage, qui sont des algorithmes qui utilisent les classifieurs pour apprendre encore mieux. Le méta-apprenant que nous utilisons est *ThresholdSelector* (« sélecteur de seuil ») et permet de choisir la métrique (parmi la f-mesure, le rappel, la précision, le taux de bonne classification et le taux de positifs) que nous souhaitons optimiser ainsi que la classe qui nous intéresse. Ainsi nous pouvons chercher à optimiser la bonne classification des faux documents, ou plus exactement leur détection.

Il faut donc choisir ce que l'on cherche à optimiser dans un contexte de détection des fraudes : est-ce qu'on veut récupérer le maximum de documents potentiellement fraudés et donner l'ensemble des documents un peu suspects pour une investigation plus approfondie par des humains, est-ce qu'on souhaite que le système détecte seulement des faux documents, quitte à en oublier un certain nombre ou est-ce qu'on souhaite équilibrer ces deux métriques, de façon à avoir le maximum de faux documents détectés tout en ayant le minimum de faux positifs, c'est-à-dire de documents authentiques accusés à tort ?

Ce méta-apprenant permet également de choisir quel classifieur utiliser. Nous avons donc testé plusieurs possibilités de classifieurs et de méta-classifieurs sur l'ensemble des 16 indices ainsi que sur la somme des indices et la somme des indices pondérés. Nous présentons les résultats des différentes métriques, obtenues en cherchant à optimiser la f-mesure avec *ThresholdSelector* en utilisant les classifieurs mentionnés dans le tableau 5.6. En plus de nos deux sous-corpus, nous présentons également le résultat de l'entrainement sur l'ensemble du corpus d'apprentissage.

Les deux premiers algorithmes testés sont des classifieurs bayésiens : *BayesNet* apprend un réseau bayésien, qui est un modèle graphique probabiliste qui calcule le probabilité des liens entre les critères (Friedman et al. 1997), et *NaiveBayes* implémente un classifieur bayésien naïf, c'est-à-dire un modèle probabiliste qui a pour hypothèses que chaque critère, ou indice, est indépendant (John & Langley 1995). Nous pouvons voir dans le sous-corpus Autres que le réseau bayésien ne retourne aucun document comme « faux », et a donc une précision incalculable. Ce classifieur a pourtant un bon rappel pour le sous-corpus Carrefour. Cela s'explique certainement par le fait qu'il y a une plus grande cohérence et corrélation des indices sur le sous-corpus Carrefour, ce que nous avons vu par exemple en combinant les indices, que sur le second sous-corpus. Le classifieur naïf en revanche est plutôt stable d'un point de vue f-mesure, même s'il perd beaucoup en précision.

Les deux classifieurs suivants sont rangés dans la catégorie « Arbres ». J48 est une implémentation d'arbre de décision et permet de hiérarchiser les critères (nœuds) et leurs valeurs (décision) (Quinlan 1993). Weka permet de visualiser l'arbre obtenu, comme présenté dans la figure 5.9. Nous pouvons voir dans cette figure que le critère SommePond, qui correspond à la somme des indices pondérés, est particulièrement décisif : si elle est

Tableau 5.6 – Évaluation des algorithmes d'apprentissage automatique avec le méta-algorithme *ThresholdSelector* sur le corpus d'apprentissage

Provenance	Classifieur	Précision	Rappel	F-mesure	*Accuracy*
Carrefour	BayesNet	0,476	0,625	0,541	0,932
	NaiveBayes	0,400	0,125	0,190	0,932
	J48	**0,917**	0,688	0,786	0,976
	RandomForest	0,789	**0,938**	0,857	0,980
	Bagging	0,538	0,875	0,667	0,944
	MultilayerPerceptron	0,875	0,875	**0,875**	0,984
Autres	BayesNet	?	0,000	?	0,944
	NaiveBayes	0,154	0,143	0,148	0,908
	J48	?	0,000	?	0,944
	RandomForest	0,256	**0,786**	0,386	0,859
	Bagging	0,138	0,286	0,186	0,859
	MultilayerPerceptron	**0,278**	0,714	**0,400**	0,880
Ensemble	BayesNet	0,134	0,733	0,227	0,699
	NaiveBayes	0,138	0,133	0,136	0,898
	J48	?	0,000	?	0,944
	RandomForest	0,283	**0,933**	0,434	0,854
	Bagging	0,115	0,833	0,202	0,603
	MultilayerPerceptron	**0,365**	0,633	**0,463**	0,912

inférieure ou égale à 0,68, alors le document est classé comme « vrai », ce qui permet d'écarter d'emblée 220 vrais documents et il n'y a que 4 faux documents qui sont alors mal classés. Le classifieur RandomForest est en réalité un méta-classifieur qui crée des forêts aléatoires en assemblant des arbres de décisions aléatoires (Breiman 2001). Cette méthode est une technique de *bagging*, comme le classifieur suivant dans notre tableau, c'est-à-dire de ré-échantillonnage pour améliorer par des méthodes statistiques des classifieurs moins performants et moins stables (Breiman 1996). De fait, comme nous pouvons le voir sur le second sous-corpus, la forêt aléatoire permet de classer de manière beaucoup plus efficace les faux documents que l'arbre de décision, qui n'arrive pas à trouver de critère discriminant.

Le Bagging que nous utilisons est également basé sur un arbre de décision : le REPTree. Ce classifieur construit un arbre en utilisant les gains d'information puis l'élague. Le méta-classifieur Bagging a pour objectif de réduire la variance du classifieur échantillonné.

La dernière méthode utilisée, MultilayerPerceptron, est un réseau de neurones qui est entrainé par rétro-propagation (Ruck et al. 1990). C'est cette méthode qui apporte les meilleures f-mesures sur les deux sous-corpus. Néanmoins, le meilleur rappel est levé par la RandomForest dans les deux cas et la meilleure précision, par l'arbre J48 dans le premier sous-corpus.

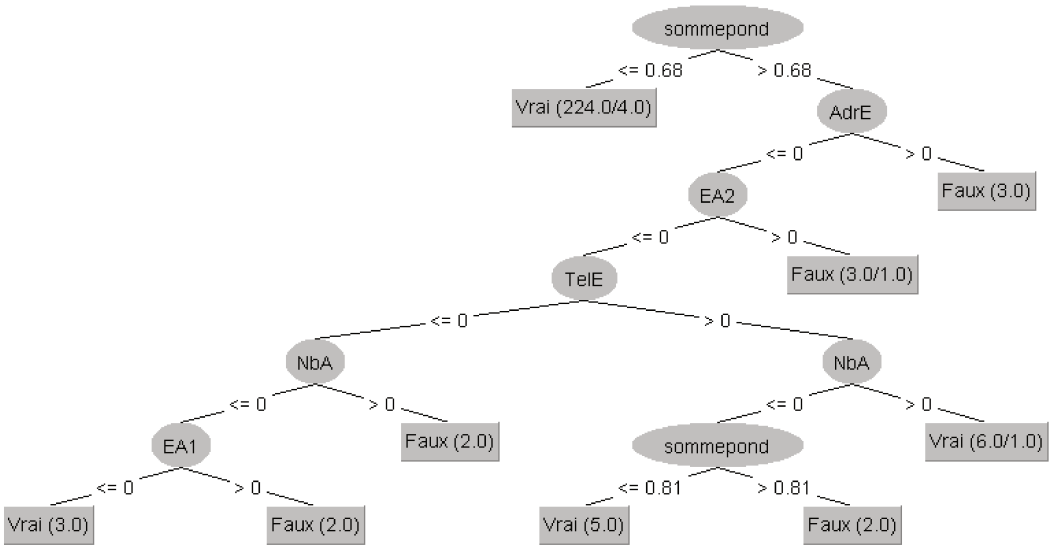


FIGURE 5.9 – Arbre de décision de sous-corpus d'apprentissage Carrefour

5.3 Évaluation

Les différents indices et les différents algorithmes que nous avons entrainé et dont nous avons analysé les performances sur le corpus d'apprentissage doivent maintenant être évalués sur le corpus de test, avec les mêmes paramètres. Nous avons vu dans la section précédente que les résultats dépendaient beaucoup de la provenance des tickets de caisse, à cause de la variété de leur format, des informations qu'ils contiennent, ou qu'ils ne contiennent justement pas, et de la diversité des fraudes, d'un point de vue sémantique.

Nous évaluerons donc dans un premier temps les indices, combinaisons et algorithmes d'apprentissage automatique déjà présentés sur le corpus d'apprentissage. Nous discuterons ensuite ces résultats en les comparant à ceux que nous avons présentés dans la section 5.1, puis nous proposerons des pistes pour aller plus loin dans le traitement sémantique des documents.

5.3.1 Résultats sur le corpus de test

Nous avons vu dans la section précédente que tous les indices et toutes les mesures testées n'étaient pas pertinents sur le corpus d'apprentissage. Néanmoins, après analyse des résultats, nous avons pu constater que certains résultats qui étaient faibles sur le corpus d'apprentissage se révélaient pertinents sur le corpus de test, et inversement. Nous montrons donc dans cette section tous les résultats obtenus sur le corpus de test, afin de pouvoir les analyser.

Vérification interne

Le tableau 5.7 indique les résultats de chaque indice de vérification interne, seul, sur le corpus de test. Comme pour l'apprentissage, nous avons séparé les documents provenant de Carrefour et d'ailleurs, puis nous calculons les résultats pour l'ensemble des documents (voir tableau 5.3, page 138).

Tableau 5.7 – Évaluation des indices de vérification interne sur les documents du corpus de test

	Indices	Précision	Rappel	F-mesure	*Accuracy*
Carrefour	$\sum$prix	0,21	**0,57**	**0,31**	0,84
	NbArticles	0,14	0,29	0,19	0,85
	Paiement	0,33	0,21	0,26	0,93
	Quantité	0,00	0,00	0,00	0,94
	Poids	**1,00**	0,07	0,13	0,94
	DateHeure	0,50	0,07	0,13	0,94
	Benford	0,04	0,43	0,07	0,33
Autres	$\sum$prix	0,06	**0,75**	**0,12**	0,31
	NbArticles	0,00	0,00	0,00	0,90
	Paiement	**0,07**	0,69	**0,12**	0,41
	Quantité	0,00	0,00	0,00	0,94
	Poids	0,00	0,00	0,00	0,94
	DateHeure	0,00	0,00	0,00	0,94
	Benford	0,05	0,50	0,10	0,44
Ensemble	$\sum$prix	0,09	**0,67**	**0,15**	0,56
	NbArticles	0,10	0,13	0,11	0,88
	Paiement	0,08	0,47	0,14	0,65
	Quantité	0,00	0,00	0,00	0,94
	Poids	**1,00**	0,03	0,06	0,94
	DateHeure	0,33	0,03	0,06	0,94
	Benford	0,05	0,47	0,08	0,39

La meilleure f-mesure sur l'ensemble du corpus est obtenu par l'indice interne $\sum$prix, comme lors de l'apprentissage. De même, sur le sous-corpus Carrefour, ce sont les mêmes mesures qui obtiennent les meilleurs résultats : Paiement pour la f-mesure et $\sum$prix pour le rappel. Néanmoins, ces résultats sont moins bons sur le corpus de test que sur le corpus d'apprentissage : de 0,16 à 0,15 pour la première f-mesure, de 0,37 à 0,26 pour la f-mesure du Paiement sur Carrefour, et de 0,81 à 0,57 pour le rappel de la somme des prix sur Carrefour.

Sur l'ensemble du corpus, les f-mesures ont toutes baissé. En revanche, sur les tickets ne venant pas de Carrefour, la $\sum$prix et le Paiement ont tous deux augmenté, notamment grâce au rappel, mais également un peu par la précision. Un autre indice qui a augmenté est l'indice DateHeure sur le corpus Carrefour, qui ne relevait rien pendant l'entrainement, mais qui s'avère intéressant pendant le test, avec une f-mesure de 0,13.

Les indices qui avaient les meilleures précisions lors de l'apprentissage, Quantité et DateHeure notamment, se révèlent finalement trop précis pour le corpus de test et ne relèvent que peu de faux document.

Vérification inter-documents

Le tableau 5.8 présente les résultats des indices de vérification inter-documents sur le corpus de test. Il fait écho au tableau 5.5, page 142.

Tableau 5.8 – Évaluation des indices de vérification inter-documents sur le corpus de test

Provenance	Indice	Précision	Rappel	F-mesure	*Accuracy*
Carrefour	PrixMoyenne	0,08	0,71	0,14	0,47
	PrixMoyenne10	0,05	0,29	0,08	0,62
	PrixListe	0,00	0,00	0,00	0,86
	Entreprise/adresse1	0,19	0,21	**0,20**	0,90
	Entreprise/adresse2	0,00	0,00	0,00	**0,94**
	Entreprise/téléphone	**0,20**	0,07	0,11	0,93
	Adresse/entreprise	0,13	0,21	0,16	0,86
	Téléphone/entreprise	0,06	**0,79**	0,10	0,18
	SIREN/entreprise	0,00	0,00	0,00	**0,94**
Autres	PrixMoyenne	0,06	0,38	0,10	0,60
	PrixMoyenne10	0,05	0,31	0,09	0,62
	PrixListe	**0,08**	**0,44**	**0,13**	0,66
	Entreprise/adresse1	0,05	**0,44**	0,10	0,50
	Entreprise/adresse2	0,00	0,00	0,00	**0,94**
	Entreprise/téléphone	0,05	0,25	0,09	0,70
	Adresse/entreprise	0,00	0,00	0,00	0,90
	Téléphone/entreprise	0,00	0,00	0,00	0,81
	SIREN/entreprise	0,00	0,00	0,00	**0,94**
Ensemble	PrixMoyenne	0,07	**0,53**	**0,12**	0,53
	PrixMoyenne10	0,05	0,30	0,08	0,60
	PrixListe	0,06	0,20	0,09	0,75
	Entreprise/adresse1	0,07	0,33	0,11	0,69
	Entreprise/adresse2	0,00	0,00	0,00	**0,94**
	Entreprise/téléphone	0,06	0,17	0,09	0,80
	Adresse/entreprise	**0,10**	0,13	0,11	0,88
	Téléphone/entreprise	0,05	0,37	0,08	0,51
	SIREN/entreprise	0,00	0,00	0,00	**0,94**

Là encore, dans l'ensemble, les indices qui montraient les meilleurs scores dans le corpus d'apprentissage donnent de moins bons résultats dans le corpus de test. C'est le cas notamment des indices portant sur les prix sur le sous-corpus Carrefour. Il semblerait

donc que ces fraudes soient moins présentes dans ces documents que dans ceux du corpus d'apprentissage, ou bien qu'elles soient moins repérables.

On peut également repérer plus d'indices qui ne relèvent aucun faux documents : Entreprise/adresse2, qui obtenait une bonne précision sur le corpus d'apprentissage, ainsi que PrixListe sur Carrefour.

Trois indices sortent du lot et améliorent leur résultat sur le corpus de test : Entreprise/téléphone, qui gagne sur les deux sous-corpus, et Adresse/entreprise et Téléphone/entreprise, qui restent nuls sur le sous-corpus Autres, mais qui passent respectivement d'une f-mesure de 0,06 à 0,16 et de 0,09 à 0,10 sur le sous-corpus Carrefour.

Nombre d'indices

En ce qui concerne le nombre d'indices minimum nécessaire pour bien détecter les faux documents, il est à 6 (f-mesure = 0,25) pour le sous-corpus Carrefour et à 3 (f-mesure = 0,12) pour le sous-corpus Autres, ce qui ne correspond pas à la figure 5.4. Pour l'ensemble du corpus de test, ce seuil à partir duquel on peut considérer qu'un document est faux est le meilleur à minimum 3 indices avec une précision à 0,07, un rappel à 0,80 et une f-mesure à 0,14. Ces scores sont semblables, aux mêmes seuils, que sur le corpus d'apprentissage, mais il y avait des seuils plus élevés sur le corpus d'apprentissage qui permettaient d'obtenir de meilleurs scores.

D'un point de vue nombre précis d'indices à utiliser, nous sommes là encore en-dessous de ce que le corpus d'apprentissage nous avait proposé, allant dans le sens de nos observations précédentes. La meilleure f-mesure sur l'ensemble du corpus d'apprentissage se situait à 5 indices (*cf* figure 5.5) ; elle se situe sur l'ensemble du corpus de test à 3 indices, bien qu'elle soit également de 0,18. Sur le sous-corpus Carrefour, elle est à 0,22 pour 7 indices, et sur les autres tickets, elle a un maximum de 0,15 pour exactement 3 indices.

Combinaisons d'indices

La meilleure combinaison d'indices sur le sous-corpus Carrefour contenait 3 ou 4 indices parmi $\sum$prix, Paiement, Benford et PrixListe. Cette combinaison obtenait 0,58 de précision, 0,44 de rappel et 0,50 de f-mesure. Elle obtient sur le sous-corpus de test équivalent une précision de 0,14, un rappel de 0,07 et une f-mesure de 0,10. En revanche, il existe de nombreuses autres combinaisons qui ont une meilleure f-mesure (1048, pour être précise). La meilleure est de 0,40 et est une combinaison de 2 ou 3 indices parmi $\sum$prix, Paiement et Adresse/entreprise. Nous constatons que, bien que les indices sur les prix aient largement perdu en pertinence, ils restent néanmoins parmi les meilleurs indicateurs pour détecter les faux documents.

Sur le second sous-corpus, la combinaison qui remportait la palme de la meilleure f-mesure fait également pâle figure sur le corpus de test équivalent : la précision est de 0,06, le rappel de 0,19 et la f-mesure de 0,09. 352 autres combinaisons ont une meilleure f-mesure, dont la meilleure est à 0,15, avec une précision de 0,09 et un rappel de 0,75.

Les critères présentés jusque-là présentent globalement de moins bons résultats de détection des faux documents que sur le corpus d'apprentissage, bien qu'il n'y ait pas eu d'apprentissage à proprement parler : les indices sont les mêmes, et, s'ils ont été choisis pour obtenir des scores sinon bons, au moins intéressants à analyser, ils n'ont pas fait l'objet d'un algorithme d'apprentissage ou de paramètres propres au corpus. Nous pouvons donc légitimement penser que le corpus de test contient de nombreuses fraudes moins facilement décelables, portant sur des informations que nous n'avons pas vérifiées, étant plus discrètes ou étant plus cohérentes.

Pondération des indices

Dans la section précédente, nous avions pondéré les indices en fonction de la précision obtenue pour chacun d'eux sur les différents échantillons du corpus d'apprentissage. Si nous reprenons les seuils des meilleures f-mesures en utilisant les mêmes pondérations sur les échantillons correspondants du corpus de test, nous obtenons les résultats présentés dans le tableau 5.9. Ce tableau reprend également, pour rappel, les valeurs sur corpus d'apprentissage, et affiche les valeurs des métriques pour le seuil ayant la meilleure f-mesure.

Tableau 5.9 – Évaluation des seuils sur la somme des valeurs pondérées des indices

Provenance	Corpus	Seuil	Précision	Rappel	F-mesure
Carrefour	App.	0,70	0,46	0,75	0,57
	Test	0,70	0,23	0,21	0,22
	Test	0,60	0,19	0,36	0,24
Autres	App.	0,35	0,10	0,50	0,17
	Test	0,35	0,06	0,25	0,09
	Test	0,15	0,08	0,81	0,14
Ensemble	App.	0,60	0,26	0,30	0,28
	Test	0,60	0,09	0,07	0,08
	Test	0,20	0,08	0,77	0,14

Nous pouvons constater que les seuils trouvés sur le corpus d'apprentissage sont trop élevés sur les corpus de test, et que si nous devons avoir ces minima requis pour classer un document comme « faux », nous n'en trouverions plus beaucoup (faible rappel). De plus, les valeurs des meilleures f-mesures sont moins élevées que sur le corpus d'apprentissage, ce qui laisse supposer que les valeurs de pondération apprises ne sont pas satisfaisantes. En effet, les précisions des indices seuls sur les corpus de test ne suivent pas les mêmes proportions et la même distribution que sur le corpus d'apprentissage.

Apprentissage automatique

Nous avons enfin évalué les algorithmes d'apprentissage automatique sur nos corpus de test. Les résultats sont présentés dans le tableau 5.10.

Tableau 5.10 – Évaluation des algorithmes d'apprentissage automatique avec le méta-algorithme *ThresholdSelector* sur le corpus de test

Provenance	Classifieur	Précision	Rappel	F-mesure	*Accuracy*
Carrefour	BayesNet	0,231	**0,214**	0,222	0,909
	NaiveBayes	**0,500**	0,071	0,125	0,940
	J48	?	0,000	?	0,927
	RandomForest	0,100	0,143	0,118	0,871
	Bagging	0,250	**0,214**	0,231	0,914
	MultilayerPerceptron	0,375	**0,214**	**0,273**	0,931
Autres	BayesNet	?	0,000	?	0,940
	NaiveBayes	0,073	**0,750**	0,133	0,414
	J48	?	0,000	?	0,940
	RandomForest	0,070	0,688	0,126	0,433
	Bagging	0,059	0,688	0,109	0,332
	MultilayerPerceptron	**0,085**	0,438	**0,143**	0,687
Ensemble	BayesNet	0,077	0,433	0,131	0,656
	NaiveBayes	0,077	0,067	0,071	0,896
	J48	?	0,000	?	0,940
	RandomForest	0,113	0,400	**0,176**	0,776
	Bagging	0,081	**0,600**	0,142	0,566
	MultilayerPerceptron	**0,136**	0,200	0,162	0,876

Les meilleures f-mesures sont obtenues par l'algorithme MultilayerPerceptron pour les deux sous-corpus, et par la RandomForest sur l'ensemble du corpus de test. On peut observer, en comparant avec le tableau 5.6, que les résultats sont vraiment moins bons qu'espérés pendant l'apprentissage. Encore une fois, l'une des explications possibles est que les faux documents du corpus de test ne présentent pas les mêmes caractéristiques que les faux documents du corpus d'apprentissage. Le fait de n'avoir que 30 documents fraudés parmi les 500 du corpus rend également l'apprentissage très difficile, comme nous allons en discuter dans la section suivante.

5.3.2 Perspectives d'améliorations de nos résultats

Les meilleures f-mesures que nous obtenons pour la détection des faux documents sur l'ensemble du corpus de test sont donc obtenues avec le RandomForest (0,176), suivi de l'indice seul $\sum$prix (0,15). Ces résultats sont très faibles, surtout si on les compare aux résultats obtenus par les candidats de la compétition. Nous avons cependant plusieurs pistes pour remédier à cela.

Augmenter le corpus d'apprentissage

Tout d'abord, notre corpus d'apprentissage, ou du moins la proportion de tickets fraudés, est beaucoup trop petit pour la diversité de la sémantique des fraudes : autant

les types de manipulations de l'image se comptent sur les doigts de la main, autant les informations modifiées sont extrêmement nombreuses (prix, adresses, dates, noms d'entreprise...) et les modifications apportées sont extrêmement variées. En effet, on observe dans nos données de nombreuses fraudes sur les informations : augmentation, diminution ou suppression pour les valeurs chiffrées, suppression ou ajout de seulement quelques caractères ou au contraire de plusieurs mots, répercussion ou non des informations modifiées sur les informations corrélées...

Afin de pallier ce problème, nous avons, comme l'ont fait certains participants, ajouté à notre corpus d'apprentissage les cent tickets du corpus d'apprentissage de la seconde tâche, qui étaient tous fraudés puisque cette tâche consistait à localiser les informations. Nous avons testé les différents algorithmes d'apprentissage automatique vus précédemment, que nous avons appliqués à l'ensemble du corpus de test. Le tableau 5.11 présente les résultats de l'apprentissage sur la classe « faux document » avec ce nouveau corpus d'apprentissage augmenté, sur trois corpus : le corpus d'apprentissage, le corpus de test de la première tâche, et le corpus de test augmenté des 80 documents du corpus de test de la seconde tâche.

Tableau 5.11 – Évaluation des algorithmes d'apprentissage automatique avec le méta-algorithme *ThresholdSelector* avec un corpus d'apprentissage augmenté

Corpus	Classifieur	Précision	Rappel	F-mesure	*Accuracy*
App. T1+T2	BayesNet	?	0,000	?	0,723
	NaiveBayes	0,253	0,754	0,378	0,462
	J48	0,217	**1,000**	0,357	0,217
	RandomForest	0,310	0,969	0,470	0,526
	Bagging	0,306	0,685	0,423	0,594
	MultilayerPerceptron	**0,391**	0,800	**0,525**	0,686
Test T1	BayesNet	?	0,000	?	0,940
	NaiveBayes	0,070	0,767	0,128	0,376
	J48	0,060	**1,000**	0,113	0,060
	RandomForest	0,058	0,933	0,109	0,080
	Bagging	0,063	0,667	0,115	0,386
	MultilayerPerceptron	**0,084**	0,667	**0,149**	0,542
Test T1+T2	BayesNet	?	0,000	?	0,723
	NaiveBayes	0,222	0,836	**0,350**	0,412
	J48	0,190	**1,000**	0,319	0,190
	RandomForest	0,188	0,764	0,302	0,331
	Bagging	**0,228**	0,600	0,330	0,538
	MultilayerPerceptron	0,217	0,591	0,317	0,517

La comparaison des résultats de l'apprentissage avec ceux que l'on a obtenus avec le corpus d'apprentissage de la seule première tâche (tableau 5.6, page 150) montre que ce corpus augmenté permet d'apprendre plus. Cependant, les résultats sur le corpus de test, pour lequel la proportion est restée à 6% de faux documents, contre 21,7%

dans l'apprentissage, sont moins bons que précédemment. Il est donc probable que les algorithmes aient trop appris et soient devenus ainsi trop spécifiques aux documents du corpus d'apprentissage. Il est également possible que les documents de ce corpus de test soient particulièrement difficiles à détecter par rapport aux autres.

La troisième partie du tableau montre de meilleurs résultats que tous les autres. La proportion de faux documents est dans ce corpus de test de 19%, ce qui se rapproche plus des conditions d'apprentissage. De plus, la variété d'informations fraudées est nécessairement plus grande que sur le corpus de test de la première tâche, et recoupe certainement plus les cas appris dans le corpus d'apprentissage augmenté. On remarque par ailleurs que ce ne sont plus les algorithmes RandomForest et MultilayerPerceptron qui donnent les meilleurs résultats, mais le classifieur bayésien naïf et le Bagging.

La taille des bases de données que l'on peut fournir en entrée de l'apprentissage reste, et restera probablement longtemps, l'enjeu principal du problème de détection des faux documents : il faudrait peut-être des milliers de documents falsifiés dans un vrai cadre de fraude, ce qui est, comme nous l'avons vu dans le chapitre 2, très difficile à obtenir.

Augmenter le nombre d'indices

Une autre solution pour améliorer la détection des faux documents serait de fournir d'autres indices sur les documents. Nous avions d'ailleurs imaginé pour cela, dans Artaud (2016) et Favre et al. (2016), de calculer des indices en comparant les informations des documents avec des connaissances externes, afin d'estimer la vraisemblance d'une information. En effet, nous aimerions pouvoir trouver qu'une bouteille d'eau, par exemple, ne peut pas coûter 90€.

Nous sommes partie du constat que, aujourd'hui, le premier réflexe de la plupart des gens quand ils souhaitent vérifier une information, c'est de chercher sur internet, et plus particulièrement sur un moteur de recherche. Internet est une base de connaissances immense, où l'on peut trouver à peu près toutes les informations que l'on cherche, à condition d'utiliser un bon moteur de recherche, et de lui fournir les bonnes requêtes. Généralement, lorsque les mots de la requête sont bien choisis, la page de résultats (SERP, pour Search Engine Results Page) donne une idée de la véracité de l'information. Sinon, il faut explorer les liens ou reformuler la requête.

Nous pouvons pour cela construire des requêtes à partir de notre ontologie en combinant des mots clés, qui sont en fait des noms d'entités de l'ontologie, comme Adresse qui est un concept ou Prix qui est une propriété de type données, avec des individus de notre ontologie. Par exemple, puisque l'on sait qu'un produit a un prix, grâce à la propriété `a_prix_total`, et que nous avons proposé une technique pour trouver les expansions des noms d'articles abrégés (*cf.* chapitre 4), nous pouvons générer la requête « Prix Pâtes tortellini pesto basilic Rana » que nous envoyons sur un moteur de recherche afin d'en récupérer la page de résultat (SERP, pour *Search Engine Result Page*).

Selon les moteurs de recherche et la sémantique qu'ils intègrent dans leur système, la SERP donne directement des résultats avec ce qui est cherché par la requête, par exemple des montants en euros pour des requêtes commençant par « prix », ou alors la SERP renvoie des résultats contenant textuellement le mot « prix ». Ce sont plutôt ces premiers

moteurs qui nous intéressent, afin de pouvoir extraire les informations directement depuis la SERP, ou depuis les sites proposés dans celle-ci. Cette méthode comporte cependant un gros inconvénient : la plupart des moteurs de recherche bloquent les robots, et limitent le nombre de requêtes possible avant de demander à remplir un captcha. C'est notamment le cas de l'un des moteurs de recherche que nous avons testé qui ne permet qu'environ 200 requêtes en quelques minutes. C'est un inconvénient car c'est également celui qui renvoie le plus de résultats pertinents. Pour rappel, nous avons environ 4700 produits dans notre corpus. Il faudrait donc payer pour accéder au service ou bien trouver le moyen de se faire passer pour plus humain qu'on ne l'est. Dans tous les cas, cela prendrait beaucoup de temps pour obtenir les indices nécessaires pour chaque document.

Plusieurs approches existent pour extraire des informations sur le web depuis des SERP comme dans l'approche proposée par Srikantaiah et al. (2013), ou celle, plus spécifique à la détection des fausses informations, proposée par Magdy & Wanas (2010). Il s'agit généralement de fournir des requêtes au moteur de recherche, puis de récupérer de façon récursive les pages web ciblées par des liens présents dans la SERP, puis celles ciblées par les premières et ainsi de suite jusqu'à atteindre une profondeur déterminée. Il s'agit ensuite d'extraire l'information de toutes ces pages web, en utilisant les techniques de traitement automatique des langues et/ou les techniques de parcours de la structure HTML de la page web.

C'est – également – à ce niveau-là que notre capacité à utiliser cette méthode se complique : comme précédemment dans notre thèse, nous rencontrons des difficultés pour extraire l'information, qui est rarement dans un contexte textuel de langage naturel. Tant qu'il s'agit de prix, la règle d'extraction est relativement simple, mais il est compliqué de savoir à quel produit celui-ci s'applique. En effet, quand vous cherchez le prix d'un produit sur n'importe quel moteur de recherche, vous n'êtes pas à l'abri d'obtenir des produits qui n'ont rien à voir, ou d'obtenir un produit similaire mais d'une autre gamme de prix.

Quand il s'agit des autres informations, qui contiennent plus de caractères alphabétiques, comme une adresse par exemple, il est beaucoup plus complexe de la trouver dans une SERP. En effet, les adresses subissent également des abréviations, comme « av. » pour « avenue » ou « bld » pour « boulevard », et beaucoup d'autres variations au niveau des caractères, comme nous l'avons vu dans le chapitre 3. Les outils de reconnaissance d'entités nommées traditionnels peuvent alors être utilisés pour extraire les informations.

Il faut ensuite pouvoir comparer la similarité entre l'information recherchée et les informations extraites d'internet, et nous pourrions nous inspirer des travaux de Fotsoh (2018), et calculer un score qui permet de distinguer faux documents et documents authentiques à travers la probabilité des informations. Il s'agit essentiellement de vérifier les liens entre les informations. Par exemple pour un commerce, on cherchera à savoir s'il est bien situé à l'adresse indiquée sur le document, pour vérifier par exemple que notre employé était bien en mission où il devait être. On peut alors pour cela créer une requête composée d'une adresse, puis chercher dans la SERP la fréquence de l'apparition du nom de l'entreprise.

On pourrait ainsi vérifier de nombreuses informations qui sont souvent présentes en

ligne, comme les horaires des magasins, les prix des menus proposés par un restaurant, les coordonnées d'un commerce... Certaines informations sont contextuelles et varient au cours du temps ou selon la zone géographique. C'est notamment le cas des fruits et légumes par exemple. Pour vérifier leur prix, il serait intéressant de relever les cours des fruits et légumes que nous pourrions comparer grâce à l'ontologie selon les dates de leurs achats.

Dans un contexte d'enquête approfondie, on pourrait même extraire de nos tickets de caisse le prénom des employés, ce qui pourrait permettre de vérifier l'authenticité d'un document en prouvant qu'il y a bien un employé du prénom extrait qui travaillait ce jour-là, à cette heure-là dans le commerce en question. On pourrait vérifier cela en comparant les informations des réseaux sociaux professionnels (LinkedIn, Viadeo...) pour chercher les employés de l'entreprise, et les réseaux sociaux classiques (Facebook, Twitter, Instagram...) pour vérifier l'emploi du temps de l'employé. Cela serait toutefois à la fois extrêmement complexe techniquement et à la fois compliqué légalement, car cela pourrait être considéré comme une atteinte à la vie privée.

5.3.3 Discussion

Nous nous sommes aperçue, avec les résultats de la compétition, que les images de documents que nous avons fournies étaient probablement trop simples pour cette tâche de détection des faux documents. En effet, le corpus a été construit principalement dans l'optique de traiter les informations du document, et nous avons fait en sorte d'obtenir les images les plus homogènes possibles afin de faciliter la reconnaissance de caractères et avoir moins de corrections post-OCR à effectuer. Les documents ont donc tous été numérisés avec le même appareil, le même éclairage, le même carton coloré en fond. Cela implique que les images sont très homogènes et que par conséquent, le moindre écart par rapport à la norme est remarquable. De plus, les possibilités de fraudes sur l'image sont assez limitées. Nous l'avons vu dans le chapitre 2, seulement quatre stratégies de fraudes sur l'image ont été recensées sur notre corpus : le copier-déplacer, l'intégration d'un morceau d'un autre document, l'imitation de caractères et la suppression de caractères, par ajout de zones similaires au fond ou par ajout de graphismes sur les caractères à cacher. Par conséquent, ce corpus d'images ne présente pas beaucoup d'originalités par rapport aux autres corpus d'images modifiées qui existent et les méthodes de l'état de l'art, déjà testées et approuvées sur ces autres corpus, s'y appliquent sans difficulté.

Il serait donc intéressant de tester les algorithmes proposés lors de la compétition sur des documents réimprimés puis re-scannés, ou bien bruités de façon aléatoire ou avec des filtres utilisés pour simuler le bruit lié à l'impression et à la numérisation, afin d'étudier la perte de qualité des algorithmes utilisant des indices liés à l'image. En effet, même s'il existe de nombreux travaux sur la détection des multiples impressions et numérisations (*cf.* 1.2), nous pensons que cela pourrait sensiblement faire diminuer les scores de détection et rendre la tâche plus intéressante. De plus, si cette manipulation rend la détection plus complexe d'un point de vue de l'image, elle ne change rien à la sémantique du document, et rend donc les approches textuelles plus utiles encore.

De même, nous pourrions diversifier la procédure de capture des tickets de caisse en prenant des clichés avec plusieurs appareils, selon des angles et des luminosités plus variés. Nous avons d'ailleurs initié ce travail, comme évoqué dans la section 2.2, sur 2000 autres documents collectés de juillet 2017 à mai 2018, que nous avons numérisé avec trois appareils différents, dont un smartphone. Il serait donc intéressant de voir si les algorithmes et indices proposés résisteraient bien à cette diversité.

Il serait également intéressant de ne pas faire frauder les documents après numérisation, comme nous l'avons fait, mais avant, avec des ciseaux, de la colle et des crayons, par exemple. Ainsi, il n'y aurait pas de traces d'altérations de l'image numérique, mais seulement une rupture dans ce que représente l'image.

Par ailleurs, le corpus est très difficile à traiter d'un point de vue sémantique compte-tenu de la diversité des altérations sémantiques apportées. En effet, nous aurions dû « cadrer » un peu plus les fraudeurs lors de la journée organisée pour les amener à n'envisager que des fraudes réalistes. Nous avons au contraire laissé libre courss à leur imagination afin d'obtenir un maximum de types de fraudes possibles et ainsi créer un corpus et un système suffisamment génériques. Cette diversité est donc un challenge intéressant, mais très vaste, qui nécessiterait du temps pour être approfondie.

De plus, notre approche est essentiellement heuristique ; elle repose sur des hypothèses liées à l'observation de notre corpus qui n'est pas représentatif de la réalité pour des raisons d'échelle. Il serait donc nécessaire de modéliser toutes ces observations et inférences sous forme d'ontologie par exemple pour pouvoir en tirer des statistiques sur la fréquence des fraudes observées.

5.4 Conclusion

Il est très difficile de constituer un corpus qui satisfasse toutes les méthodes que l'on pourrait imaginer, allant du bas-niveau de traitement de l'image, qui s'intéresse au pixel et au signal de l'image, jusqu'aux méthodes utilisant toute la sémantique exploitable automatiquement. En effet, des corpus comme celui de Bertrand et al. (2015) sont parfaits pour évaluer les méthodes bas-niveau sur l'image, mais sont pauvres en informations textuelles, alors que notre corpus semble présenter des lacunes pour évaluer les méthodes de l'image, certainement car nous l'avons surtout construit dans l'idée de traiter le texte qui en est issu que l'image.

Face aux résultats des approches proposées lors de la compétition, et plus généralement aux travaux sur la détection d'images falsifiées et sur la sécurisation des images de documents, qui sont très développés et riches, nous nous sommes rendu compte au cours de ce chapitre de la complexité de traiter des fraudes documentaires par une approche sémantique de vérification des informations. En effet, la sémantique du ticket de caisse est particulièrement riche et la créativité de nos fraudeurs en matière de modifications apportées aux informations du document a fait de ce corpus un défi sur lequel nous espérons que de nombreux chercheurs s'attaqueront.

Nous avons reçu environ une quinzaine de demandes de téléchargement par mois de notre corpus depuis sa mise en ligne en août 2018. Ces demandes proviennent de

chercheurs, d'étudiants, ou d'entrepreneurs qui cherchent essentiellement à travailler sur l'image de document, dont beaucoup sur l'amélioration des techniques de reconnaissance de caractères sur des documents tels que les tickets de caisse. Certaines demandes concernent la détection des fraudes, mais aucune demande ne précise pour l'instant que c'est pour travailler sur le texte ou sur la sémantique, alors que plusieurs précisent leur préférence pour l'image.

Il serait très intéressant, à l'instar de la méthode 3, de rajouter à nos indices sur le texte des indices sur l'image, afin d'améliorer la détection des faux documents. Par ailleurs, la deuxième tâche de la compétition qui portait sur la localisation des fraudes dans les documents, et que nous avons laissée de côté pour des raisons de temps et de complexité, présente un challenge non encore résolu, même par les techniques de traitement de l'image. En effet, les deux soumissions reçues ont montré des résultats qui peuvent encore largement être améliorés. Nous pensons que le traitement sémantique de l'information peut tirer son épingle du jeu, notamment pour localiser les manipulations de l'image de type copier-déplacer pour lesquelles ces dernières ne peuvent pas déterminer quel est le morceau original et quel est le morceau déplacé (Artaud et al. 2018).

Par ailleurs, une augmentation du corpus par d'autres documents, et l'extraction d'autres indices, à la fois sémantiques et graphiques, permettraient certainement de fournir à des algorithmes d'apprentissage profond de quoi détecter de façon plus efficace et précise les faux documents.

Conclusion

La détection des faux documents est un enjeu planétaire, qui coûte des milliards d'euros tous les ans aux entreprises, aux administrations, aux États et aux particuliers. Les faux documents qui nous ont particulièrement intéressés sont les documents de la vie de tous les jours. Ils peuvent être de toutes sortes : factures, bulletins de salaire, avis d'imposition, certificats médicaux, justificatifs de scolarité, diplômes, quittances de loyer, curriculum vitae... Nous avons toutefois écarté les documents d'identité et les billets de banque qui contiennent déjà des sécurités intrinsèques et pour lesquels de nombreux outils de détection existent. Quels qu'ils soient, ces documents peuvent servir à justifier quelque chose, pour obtenir un droit ou une compensation, pour prouver une transaction. Falsifier un document, c'est donc chercher à obtenir davantage, chercher à prouver quelque chose qui n'est pas forcément vrai, afin, généralement, de gagner du temps ou de l'argent.

Aujourd'hui, la transformation numérique accélère le partage de documents et donne les moyens à tout un chacun de créer, numériser, modifier, copier-coller, effacer, éditer ou imprimer toutes sortes de documents. Il y a donc moins d'obstacles matériels à la fraude documentaire.

Les entreprises et les administrations, comme nous l'avons vu dans l'introduction, prennent de plus en plus conscience de cet enjeu et cherchent donc à se prémunir de pertes en cherchant des solutions matérielles ou logicielles pour l'automatisation de la détection des faux documents. Notre thèse s'inscrit donc dans un enjeu économique et politique croissant et propose, à l'heure des grandes masses de données, de rendre les échanges plus transparents et plus sûrs.

Rappel des contributions

Nous avons montré dans le chapitre 1 que de nombreux travaux existent pour différencier le vrai du faux, l'authentique du falsifié, la vérité du mensonge. Cependant, si de nombreuses solutions, académiques et industrielles, existent pour détecter des images trafiquées et des images de documents contenant des anomalies, peu de travaux ont été réalisés au niveau sémantique du contenu des documents auxquels nous nous intéressons dans ce travail. En parallèle de ces travaux du domaine des *Document Forensics*, nous avons montré qu'il existe également des recherches pour détecter les infox, ou *fake news*, qui se basent sur la sémantique des informations contenues dans les articles de presse examinés. Notre travail vient donc tenter de combler l'espace entre ces deux volets

en cherchant à démêler les vraies informations des fausses dans les documents. Cette approche vise en effet à combler les limites des approches sur l'image : certains faux documents ne sont pas des documents modifiés, mais des faux documents natifs. L'image n'est donc pas fausse, mais le fond, lui, l'est.

Afin de répondre au mieux à notre problématique, il nous fallait donc trouver des échantillons à analyser afin d'orienter notre système de détection de fraudes vers des documents crédibles, c'est-à-dire originaux. Étant donné qu'il est très difficile pour les entreprises et administrations de fournir ce genre de documents, trop sensibles et incluant souvent des données personnelles, nous avons constitué nous-mêmes notre propre corpus de documents. Nous avons choisi comme cas d'étude le ticket de caisse, qui contient de nombreuses informations falsifiables, et que l'on peut collecter et diffuser à grande échelle. Le chapitre 2 explique comment nous avons collecté et numérisé des tickets de caisse, comment nous avons traité les images pour en extraire le texte grâce à un OCR, puis comment nous avons corrigé, automatiquement puis manuellement ces transcriptions. Nous avons ensuite fait frauder les tickets par vingt-cinq personnes différentes, afin d'obtenir le plus de cas de fraudes possible, aussi bien d'un point de vue manipulation de l'image que d'un point de vue modification sémantique.

Le chapitre 3 présente les différentes informations contenues dans les tickets de caisse, qui se retrouvent également, pour la plupart d'entre elles, dans d'autres documents. Les techniques d'extraction d'informations et de peuplement d'ontologie utilisent en temps normal des techniques de traitement automatique des langues, afin de reconnaitre les expressions à relever. Dans notre cas, ces techniques ne peuvent pas être utilisées, étant donné la nature de nos documents, qui ne contiennent ni phrases ni discours. Nous présentons donc notre propre système d'extraction à base d'expressions régulières et de conditions, ainsi que le modèle ontologique que nous avons choisi pour organiser et sauvegarder ces informations.

Parmi ces informations, nous devons faire face à une difficulté importante dans le traitement sémantique de l'information : les abréviations. Les rédacteurs de tickets de caisse, comme pour de nombreux autres documents du fait de leur mise en page et de la présence récurrente de tableaux et de zones délimitées pour chaque information, doivent réduire l'information pour qu'elle tienne dans de petits espaces. Il faut donc trouver des stratégies pour comprendre ce que signifie chaque information, c'est-à-dire pour lier un mot ou une expression abrégée à son expansion la plus probable. Cette tâche est complexe, et même les humains ont parfois du mal à comprendre certaines abréviations. Nous avons donc proposé une méthode dans le chapitre 4 pour répondre à cette problématique importante quand on traite de la sémantique de documents.

Une fois toutes ces informations extraites et étendues, nous avons proposé dans le chapitre 5 des approches pour vérifier la cohérence des informations au sein d'un document et entre les documents, ainsi que des propositions pour rechercher des informations externes à comparer aux informations extraites des documents. Nous présentons également dans ce chapitre les méthodes et résultats de la compétition internationale que nous avons organisée et qui donnent d'excellents résultats sur la détection des faux documents, en utilisant des techniques de traitement de l'image.

Le cas d'étude et l'implémentation que nous avons présentés ne sont pas, par définition, génériques : notre approche propose des règles établies sur l'observation de notre corpus, qui n'est probablement pas représentatif de l'ensemble des documents qui peuvent être échangés entre entreprise, administration et particulier. Néanmoins, l'approche que nous avons proposée – l'idée de vérifier le contenu du document en vérifiant sa cohérence et en le confrontant aux autres documents par l'extraction et la modélisation de ses informations – peut être appliquée à tous types de documents.

Cette chaîne de traitement, aussi innovante que complexe, apporte à chaque étape des difficultés et des erreurs qui impactent la suite de la chaîne. Ainsi, les erreurs de reconnaissance de caractères qui n'ont pas été corrigées provoquent des erreurs d'extraction d'information, et augmentent la difficulté pour trouver les bonnes expansions pour chaque nom d'article. De même, les erreurs d'extraction et les mauvaises expansions influent sur la qualité des indices utilisés pour la détection de faux documents : si tous les articles d'un ticket de caisse ne sont pas trouvés, le montant total ne risque pas d'être égal à la somme de leur prix. En revanche, si nous considérons que notre approche sert à détecter les anomalies dans un document, et que les erreurs OCR en font partie, alors notre approche peut s'avérer pertinente. Nous pourrions en effet utiliser notre approche de vérification des informations pour améliorer, ou corriger, la sortie OCR et lever des alertes chaque fois que les informations ne sont pas cohérentes ou vraisemblables.

Ce travail représente les premières esquisses d'un sujet peu exploré : la vérification des informations dans des documents par une approche qui mêle éléments de structure, éléments liés à l'image et éléments liés au sens du texte qu'ils contiennent. Cette problématique complexe décloisonne la notion de document, très souvent, à raison, analysé uniquement sur son aspect graphique. Il s'agit alors d'intégrer des techniques venant de nombreux domaines : traitement du signal, traitement des images, reconnaissance des formes, analyse de structure de documents, analyse lexicale, traitement syntaxique, analyse sémantique, ingénierie des connaissances, systèmes d'information, *web mining*...

L'approche textuelle que nous proposons n'obtient certes pas les meilleurs résultats dans la tâche de détection des faux documents et contient des limites dues à une approche très heuristique, focalisée sur les observations d'un seul corpus ; cependant, elle ouvre la voie à de nombreuses perspectives de traitement sémantique des documents.

Perspectives

Tout d'abord, il serait intéressant, puisque nous traitons de documents qui possèdent une mise en page très particulière et significative, de prendre en compte cette dernière dans l'extraction d'information. En effet, des outils comme DMOS (Description and MOdification of Segmentation) proposé par Couäsnon (2001), qui permettent de décrire et segmenter les documents possédant une structure fortement marquée grâce à l'utilisation de règles, pourraient aider à la fois à la reconnaissance du texte, en apprenant sur chaque partie du document, et à la fois à l'extraction d'informations. Il serait vraiment intéressant de combiner règles textuelles (expressions régulières) et règles de mise en page, par exemple en retrouvant les structures de tableau, très présentes dans de nombreux docu-

ments. De nombreux travaux, plus récents, s'attachent à extraire les structures complexes des documents, comme ceux de Alhéritière et al. (2017).

Par ailleurs, le fait d'utiliser une ontologie, dans un format XML/RDF, permettrait également de lier les informations que nous avons extraites avec d'autres données du web sémantique, et ainsi valider les informations, ou au contraire détecter les incohérences ou les anomalies. Le web sémantique se veut en effet une immense source de connaissances et permettrait de comparer nos informations avec des connaissances déjà validées. D'autres sources de données existent, notamment grâce à la politique d'ouverture des données publiques (*open data*). Ainsi, les données sur les entreprises, de type immatriculation, nom, adresse du siège social, etc. sont disponibles via des API et téléchargeables en ligne. Il serait donc souhaitable de prendre toutes ces ressources en compte, en sélectionnant un ensemble de sites et en créant les outils qui correspondent à chacun de ces sites, ou bien en créant un système de recherche d'information complet qui prendrait en compte toutes sortes de sources de données, de la page web traditionnelle au web sémantique, en passant par différents formats : CSV, JSON, XML...Par ailleurs, nous pourrions compléter cette ontologie avec des modèles de fraudes récurrentes, ce qui permettrait de formaliser notre approche et de lui donner ainsi un cadre plus théorique.

Nous pouvons également imaginer obtenir des informations d'un plus haut niveau sémantique et réussir ainsi à déterminer automatiquement pour chaque produit acheté sa catégorie (alimentation, hygiène...) voire des sous-catégories (plat cuisiné, fruits, légumes, boissons avec ou sans alcool...). Cela permettrait de produire des statistiques sur les types de produits achetés et sur les heures ou les dates de consommation, et ainsi déterminer des tendances sur des groupes de produits. Ces tendances pourraient être utilisées pour soulever des doutes sur les documents : peut-être qu'alors, acheter du vin chaud en plein mois de juillet, ou bien consommer des fraises au mois de décembre pourraient éveiller des soupçons...

Enfin, nous espérons que la mise en ligne et la diffusion de ce corpus de documents à la fois sous leur format image et leur format textuel permettra d'améliorer la reconnaissance de caractères sur ce type de documents, d'offrir la possibilité de lier l'extraction d'information à la segmentation de l'image, et enfin de proposer de futures approches plus génériques, utilisant à la fois la sémantiques et l'image, pour la détection des faux documents, et pour la localisation des fausses informations dans le texte et dans l'image. Ce corpus pourrait également être augmenté d'autres types de documents et permettre à l'avenir de traiter des indices encore plus haut niveau, grâce, peut-être, à l'intelligence artificielle. L'intérêt porté au corpus lors de la compétition et depuis sa mise en ligne, manifesté dans les commentaires déposés lors de la demande de téléchargement et par la quinzaine de téléchargements par mois avant même la publication de l'article, nous montre d'ailleurs qu'un tel corpus était attendu.

Annexe A

Procédure de fraude des tickets de caisse

Mode d'emploi

Prenez un Schoko-bons, une serviette, et venez nous voir :

1. Nous vous affectons un (premier) lot de 10 documents à frauder
2. Vous choisissez un ordinateur
   (a) Normal : paint, Gimp2, InkScape
   (b) Avec machine virtuelle installée : contenant Photoshop, Gimp
3. Vous téléchargez votre lot depuis l'adresse :
4. Vous modifiez une image comme bon vous semble et l'enregistrez en format .jpg
5. Vous modifiez le texte correspondant à l'image (même numéro que l'image) NB : vous pouvez commencer par modifier le texte puis l'image
6. Vous complétez le fichier `list_XXX.txt` avec les méthodes que vous avez utilisées
   (a) CPI (copier-coller à l'intérieur du document)
   (b) CPO (copier-coller à l'extérieur du document)
   (c) IMI (boite textuelle imitant la police)
   (d) CUT (suppression d'un ou plusieurs caractères)
   (e) Autres
7. Vous recommencez les étapes 4 à 6 pour chacun des documents !
8. Vous enregistrez tout et déposez le dossier complet fraudé sur `https://ao.univ-lr.fr/index.php/s/RCXzelu15axSmeH`

Allez vous servir à boire et à manger, discutez, souriez et recommencez !

FAQ

1. Pourquoi frauder des tickets de caisse ?
   — Remboursement de frais de mission (gagner un peu plus d'argent, produits non remboursés)
   — Preuve d'achat pour assurance (idem)
   — Preuve d'achat pour la garantie (date trop vieille...)
2. Qu'allez-vous faire de ces tickets ?

   Nous allons créer un dataset (corpus) de documents fraudés qui sera utilisé dans le cadre d'une compétition internationale de détection des fraudes. Les participants devront utiliser des techniques d'analyse d'image ou d'analyse de texte, ou les deux, pour trouver quels sont les documents fraudés parmi tous et où sont les fraudes dans le document.
3. Pourquoi remplir le tableau avec les méthodes utilisées ?

   Pour analyser les résultats de la détection des fraudes, il peut être intéressant d'analyser les types de fraude détectés selon les méthodes utilisées.

Annexe B

Procédure pour construire la Vérité Terrain des abréviations

Annotation à l'aveugle

Pour chaque annotateur·rice, créer un fichier JSON structuré comme suit (ou reprendre le fichier modèle fourni) :

```
{
"titre": "annotations PrénomAnnotateur·rice",
"correspondances": [
    {
       "produittc": "LA ROCHELOISE BLONDE",
      "produitannotateur ": "Bière blonde La Rocheloise"
    }, {
       "produittc": "160G BLC PLT 4TR.F",
       "produitannotateur ": "Blanc de poulet",
       "quantite": "4 tranches fines, 160g",
    }, {
       "produittc": "10CL CREME FRAICHE",
       "produitannotateur ": "Crème fraîche"
       "quantite": "10 cl"
    },]
}
```

Le titre est formé des mots « annotations » et du prénom de l'annotateur·rice. Les « correspondances » sont une liste de dictionnaires qui ont pour clés :

produittc (string) : le nom du produit inscrit sur le ticket de caisse ;

produitannotateur (string) : le nom de produit que l'annotateur·rice devine grâce à ses connaissances et ce qu'il/elle peut trouver sur Internet. Ce nom doit être le plus complet possible, avec la marque si indiquée ou devinée, le type du produit

(« bière » dans le premier exemple). Si l'annotateur·rice ne sait pas : mettre une chaîne de caractères vide ("") et la confiance à « nulle » ;

quantite (string, optionnel) : quand le produittc contient une quantité (ex : 75cl, 250g) ;

confiance (string, optionnel si) : 3 valeurs possibles :
- fort (par défaut, pas besoin de le mettre) : l'annotateur·rice est sûr que ce qu'il met correspond bien au produittc
- moyen : l'annotateur·rice a des doutes mais pense que c'est correct
- faible : l'annotateur·rice n'est pas sûr·e du tout mais a quand même écrit quelque chose
- nulle : l'annotateur·rice n'a rien écrit car il/elle ne sait pas

Annotation avec références

Pour chaque annotateur·rice, créer un fichier JSON structuré comme suit :

```
{
"titre": "annotations Chloé",
"correspondances": [
    {
       "produittc": "LA ROCHELOISE BLONDE",
       "produitannotateur ": "Bière blonde La Rocheloise"
    }, {
       "produittc": "10CL CREME FRAICHE",
       "produitweb": [
            "Crème fraîche épaisse",
            "Crème fraîche légère"],
       "confiance": "moyen"
    }, {
       "produittc": "160G BLC PLT 4TR.F",
       "produitweb": "Blanc de poulet Carrefour",
       "quantite": "la barquette de 4 tranches - 160g",
       "confiance": "fort"
    }, ]
}
```

Le titre est formé des mots « annotations » et du prénom de l'annotateur·rice. Les « correspondances » sont une liste de dictionnaires qui ont pour clés :

produittc (string) : le nom du produit inscrit sur le ticket de caisse

produitweb — (string) : le nom du produit issu du web correspondant au produit du ticket de caisse que l'annotateur·rice trouve dans le fichier de produits issus du web
- (liste) : s'il y a plusieurs produits possibles, liste de tous les produits correspondants issus de web

confiance (string) : 3 valeurs possibles :
- — fort : l'annotateur·rice est sûr·e que le produit web correspond bien au produittc
- — moyen : l'annotateur·rice a des doutes mais pense que c'est correct
- — faible : l'annotateur·rice n'est pas sûr du tout mais ne sait pas ce que ça pourrait être d'autre

produitannotateur·rice (string, à la place de produitweb) : quand l'annotateur·rice sait à quel produit ça correspond mais qu'il/elle ne le trouve pas dans le fichier issu du web. NB : pas de « confiance » quand « produitannotateur »

quantite (string, optionnel) : quand le produittc contient une quantité (ex : 75cl, 250g) et que l'annotateur·rice retrouve cette quantité dans le fichier issu du web associée au bon nom de produit. Cela permet d'augmenter la confiance dans l'annotation. NB : ce n'est pas parce que la quantité n'est pas la même que le produit n'est pas le bon (drive vend des portions plus familiales que city).

Pour chaque nom de produit du ticket de caisse, l'annotateur·rice cherche dans le fichier de noms de produits issus du web (fichier excel) le ou les produit(s) qui correspond(ent) le(s) plus. Pour cela, il/elle peut chercher un mot saillant du nom de produit grâce à la fonction « rechercher » d'Excel (ctrl + f). Il faut parfois chercher avec plusieurs mots avant de trouver un produit intéressant.

La quantité associée au nom de produit peut parfois permettre de choisir entre plusieurs possibilités. Dans le cas où la quantité est la même que sur le ticket de caisse, on peut l'indiquer dans le JSON.

L'annotateur·rice choisit un niveau de confiance selon les critères indiqués plus haut.

Par défaut, c'est un « produitweb » qui est indiqué. Cependant, si l'annotateur·rice ne trouve pas le produit correspondant dans le fichier des produits issus du web, il/elle peut créer un string « produitannotateur » en suivant la trame suivante : « [type de produit] [nom du produit] [marque du produit] ». Exemples : « Vin rosé Sauvignon Soif d'évasion », « Barre chocolatée Sundy Quinto Nestlé ».

Annexe C

Données extraites du corpus

L'extraction d'informations nous permet de tirer quelques informations sur notre corpus, qui viennent compléter la description que l'on en fait dans les chapitres 2 et 3. Nous présentons donc dans cette annexe quelques statistiques sur notre corpus.

Dates

Pour commencer, nous pouvons étudier la période de collecte de tickets de caisse sur la figure C.1. Nous devons préciser que 2 mails ont été envoyés en décembre (un premier pour expliquer puis un rappel), puis 2 en janvier (un rappel, et un mail sur les soldes d'hiver), un en février qui fait un premier point sur la collecte, un en avril, puis un en juin. Nous avons scanné cette première tranche de tickets courant juillet 2017. Certains tickets datent d'avant le début de la collecte (et certains même avant avril 2016) car plusieurs personnes avaient conservé de vieux tickets.

Provenances géographiques

Nous pouvons également visualiser et situer la diversité des adresses extraites et la répartition sur le territoire français sur la figure C.2, réalisée avec Google My Maps. Nous ne le voyons pas sur la carte pour des raisons d'échelle, mais l'un de nos tickets vient également de Guadeloupe. 82% de nos tickets ayant une adresse dans notre corpus viennent de Charente-Maritime (602), 7,5% du Loiret (55), 3,3% du Gard (24), 2,3% de Paris (17) et les 4,9% restants viennent de 17 autres départements. Les tickets ne viennent pas que de La Rochelle et de ses environs, même si la majorité d'entre eux restent locaux. La figure C.3 montre la localisation des entreprises ayant émis un ou plusieurs tickets de caisse sur La Rochelle et sa périphérie proche.

Montants des tickets

La figure C.4 indique les montants moyens et médians des tickets de caisse pour chaque mois. On peut constater que la médiane de novembre et décembre est plus élevée, ce qui est peut-être révélateur de la période d'achats de Noël. Les statistiques étant

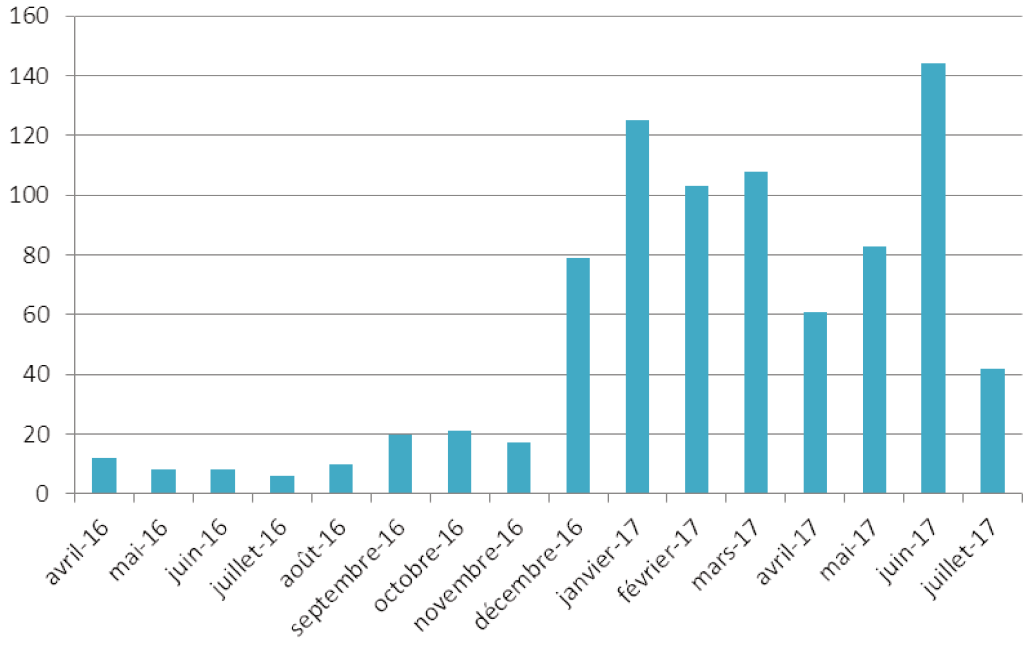


FIGURE C.1 – Nombre de tickets émis par mois.

effectuées sur le corpus utilisé pour la détection des faux documents, il est possible que certains montants soient faux et augmentent ainsi artificiellement les moyennes.

Entreprises

Nous pouvons également déterminer que les 1000 tickets de notre corpus viennent de seulement **302 entreprises différentes** : 11 Carrefour de tailles diverses (contact, city, express, market...) et 291 d'autres commerces de tous types (hyper marchés, commerces de proximité, restaurants, stands de marché...) Cela dit, ces chiffres peuvent n'être pas tout à fait exacts, car plusieurs problèmes apparaissent dans le décompte. En effet, certains magasins regroupent différents services qui émettent des tickets séparément les uns des autres. C'est le cas par exemple de Leclerc qui regroupe, à la même adresse, une poissonnerie, une station service, un espace culturel, une parapharmacie, un garage, une jardinerie... Étant donné que certains de ces services sont à l'intérieur du magasin lui-même, et d'autres seulement dans la galerie marchande, donc avec un accès séparés, la question se pose de savoir s'il faut recenser une seule entreprise ou plusieurs. Un autre problème concerne les noms au fil du temps : nous avons vu dans le chapitre 5 que les magasin Carrefour city avait changé de nom pour devenir seulement « city », c'est également le cas pour les magasins de l'enseigne U, dont l'un de ses magasins parisiens, situé rue de la Tombe Issoire dans le 14e arrondissement de Paris, est passé de « Super U » à « U express ».

FIGURE C.2 – Localisation des entreprises dont l'adresse a été extraite de notre corpus en France.

FIGURE C.3 – Localisation des entreprises dont l'adresse a été extraite de notre corpus à La Rochelle et ses environs.

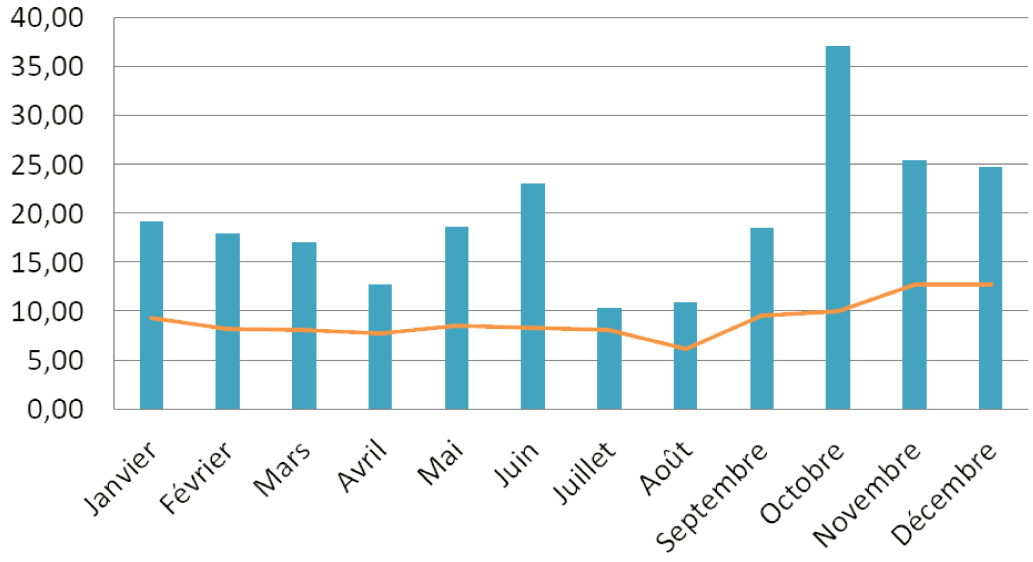


FIGURE C.4 – Moyenne (histogramme) et médiane (courbe) des montants totaux en euros des tickets de caisse selon les mois.

Achats

Concernant les produits, nous avons extrait 4556 produits (2685 venant de Carrefour et 1870 venant d'ailleurs) pour un total d'environ 39 700€. Compte-tenu des résultats de l'évaluation de l'extraction d'information, nous pouvons estimer qu'il y a environ 3% de produits en plus, soit **4692 produits**. Sur l'échantillon évalué, 3% des produits Carrefour ne sont pas trouvés ainsi que 25% des articles des autres commerces, et 22% des produits relevés pour les autres commerces ne sont pas des produits.

Pour les tickets Carrefour, la moyenne est de 5,6 produits achetés, et la médiane est de 3. Pour les tickets autres, la moyenne est de 4,6 et la médiane de 3 également, sans compter les tickets pour lesquels aucun produit n'est extrait. Ces chiffres montrent surtout que les tickets que nous avons récoltés proviennent de petites courses, souvent faites uniquement pour le déjeuner ou pour le dîner. Cela est corrélé aux informations fournies par la figure C.5 qui montre la répartition des émissions de tickets selon les heures de la journée. Nous pouvons nettement voir deux pics entre midi et une heure, puis entre dix-huit et dix-neuf heures.

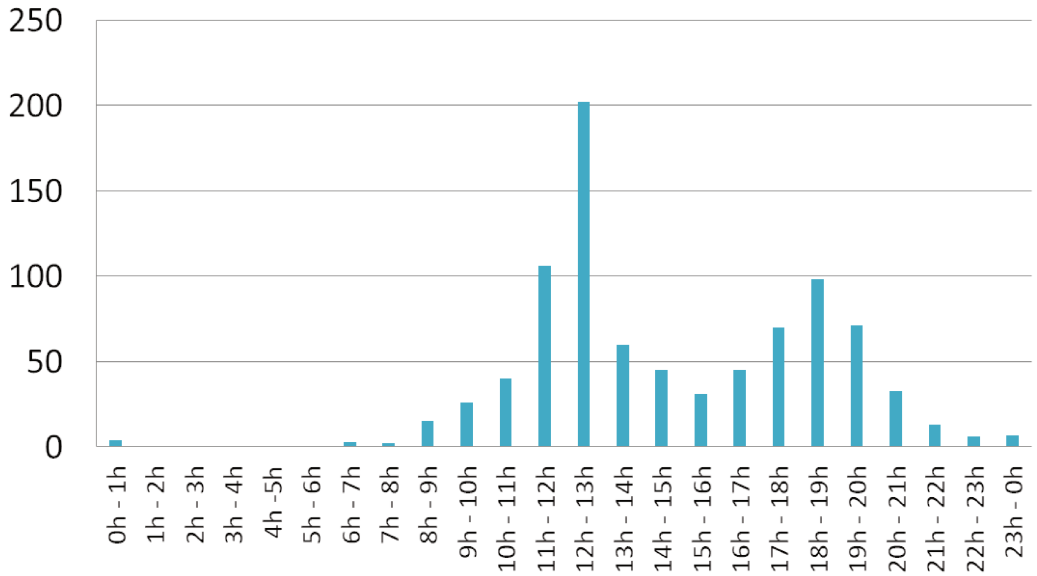


FIGURE C.5 – Nombre de tickets émis par tranche horaire de la journée.

Par ailleurs, comme nous pouvons le constater sur la figure C.6, la loi de Benford, qui déclare que la fréquence des chiffres de 1 à 9 décroît dans tout ensemble de nombres, est vérifiée sur l'ensemble du corpus si nous prenons le premier chiffre du prix. On compte par ailleurs 664 prix à moins de 1€, commençant donc par un zéro. Si l'on considère le premier chiffre significatif du prix, c'est-à-dire le premier chiffre non nul, la loi de Benford n'est pas vérifiée, comme nous pouvons le constater dans la figure C.7. En effet, les prix des produits sont souvent fixés de façon à ne pas dépasser le seuil psychologique de l'euro supplémentaire et finissent donc souvent par une somme élevée de centimes, commençant donc par 8 ou par 9.

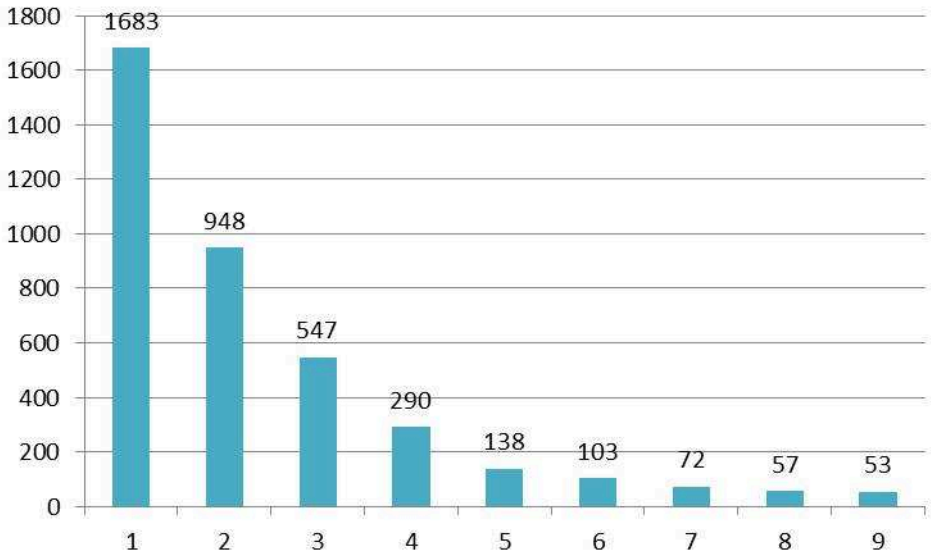


FIGURE C.6 – Graphique de la fréquence de chaque chiffre en initiale d'un prix dans l'ensemble du corpus.

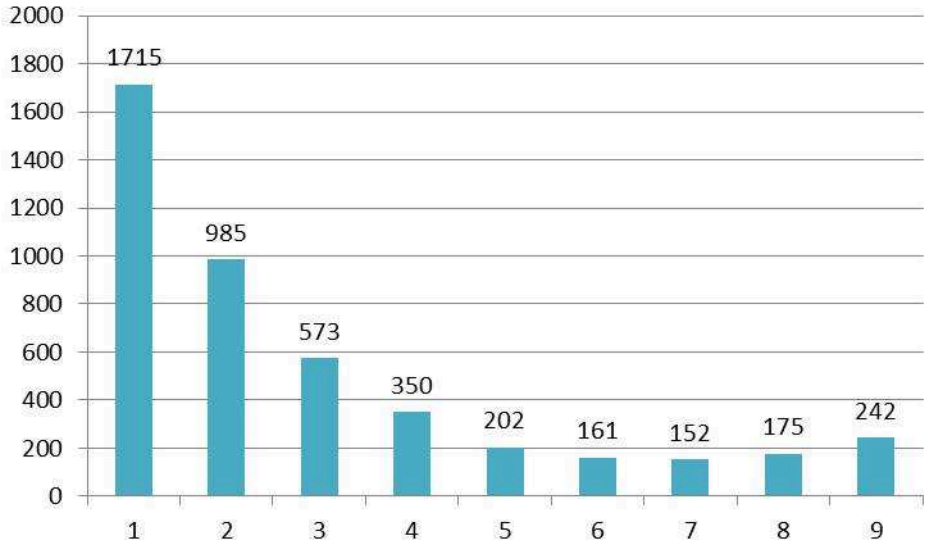


FIGURE C.7 – Graphique de la fréquence de chaque chiffre en premier chiffre non nul d'un prix dans l'ensemble du corpus.

Publications personnelles

Artaud, C. (2016), Document authentication by information checking, *in* 'Forum Jeune Chercheur INFORSID', Grenoble, France. Voir page 158.

Favre, C., Artaud, C., Duffau, C., Fraisier, O. & Kotto-Kombi, R. (2016), 'Forum jeunes chercheurs à inforsid 2016', *Revue des Sciences et Technologies de l'Information-Série ISI : Ingénierie des Systèmes d'Information* **22**(2), 121–147. Voir page 158.

Artaud, C., Doucet, A., Ogier, J.-M. & Poulain d'Andecy, V. (2017), Receipt dataset for fraud detection, *in* 'Computational Document Forensics (IWCDF), 2017 International Workshop on, in conjunction with ICDAR'. Voir page 65.

Artaud, C., Doucet, A., Ogier, J.-M. & Poulain d'Andecy, V. (2018), Automatic matching and expansion of abbreviated phrases without context, *in* '2018 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)', Hanoi, Vietnam. Voir page 125.

Artaud, C., Sidère, N., Doucet, A., Ogier, J.-M. & Poulain d'Andecy, V. (2018), Find it ! fraud detection contest report, *in* '2018 24th International Conference on Pattern Recognition (ICPR)', pp. 13–18. Voir pages 128 et 162.

Bibliographie

Abiteboul, S. (1997), Querying semi-structured data, *in* 'International Conference on Database Theory', Springer, pp. 1–18. Voir page 16.

Abramova, S. (2016), 'Detecting Copy – Move Forgeries in Scanned Text Documents', *Electronic Imaging* **2016**(8), 1–9. Voir page 30.

Achananuparp, P., Hu, X. & Shen, X. (2008), 'The evaluation of sentence similarity measures', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **5182 LNCS**, 305–316. Voir page 107.

Ackoff, R. L. (1989), 'From data to wisdom', *Journal of Applied Systems Analysis* **16**(1), 3–9. Voir pages 69 et 71.

Afli, H., Qiu, Z., Way, A. & Sheridan, P. (2016), 'Using smt for ocr error correction of historical texts'. Voir page 57.

Ahmed, A. G. H. & Shafait, F. (2014), Forgery detection based on intrinsic document contents, *in* 'Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on', IEEE, pp. 252–256. Voir page 33.

Ahmed, H., Traore, I. & Saad, S. (2017), Detection of online fake news using n-gram analysis and machine learning techniques, *in* 'International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments', Springer, pp. 127–138. Voir page 39.

Alaei, A. & Delalandre, M. (2014), A Complete Logo Detection/Recognition System for Document Images, *in* '2014 11th IAPR International Workshop on Document Analysis Systems', pp. 324–328. Voir page 34.

Alhéritière, H., Cloppet, F., Kurtz, C., Ogier, J.-M. & Vincent, N. (2017), A document straight line based segmentation for complex layout extraction, *in* 'Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on', Vol. 1, IEEE, pp. 1126–1131. Voir page 166.

Allcott, H. & Gentzkow, M. (2017), 'Social Media and Fake News in the 2016 Election', *Journal of Economic Perspectives* **31**(2), 211–236. Voir page 36.

Amerini, I., Caldelli, R., Bimbo, A. D., Fuccia, A. D., Saravo, L. & Rizzo, A. P. (2014), Copy-move forgery detection from printed images, *in* 'Media Watermarking, Security, and Forensics'. Voir page 29.

Amerini, I., Uricchio, T. & Caldelli, R. (2017), Tracing images back to their social network of origin : A CNN-based approach, *in* '2017 IEEE Workshop on Information Forensics and Security (WIFS)', pp. 1–6. Voir page 40.

Andro, M. & Saleh, I. (2015), 'Bibliothèques numériques et gamification : panorama et état de l'art', *I2D – Information, données & documents* **53**(4), 70–79. Voir page 61.

Baldwin, T., Kim, Y.-B., De Marneffe, M.-C., Ritter, A., Han, B. & Xu, W. (2015), Shared Tasks of the 2015 Workshop on Noisy User-generated Text : Twitter Lexical Normalization and Named Entity Recognition, *in* 'Proceedings of the ACL 2015 Workshop on Noisy User-generated Text', pp. 126–135. Voir page 106.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M. & Etzioni, O. (2007), Open information extraction from the web., *in* 'International Joint Conferences on Artificial Intelligence', Vol. 7, pp. 2670–2676. Voir page 72.

Bappy, J. H., Roy-Chowdhury, A. K., Bunk, J., Nataraj, L. & Manjunath, B. S. (2017), Exploiting Spatial Structure for Localizing Manipulated Image Regions, *in* 'Proceedings of the IEEE International Conference on Computer Vision', Vol. 2017-Octob, pp. 4980–4989. Voir page 30.

Bas, P., Furon, T., Cayre, F., Doërr, G. & Mathon, B. (2016), *Watermarking Security*, Springer Briefs, Springer. Voir page 26.

Batista-Navarro, R., Rak, R. & Ananiadou, S. (2015), 'Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics', *Journal of Cheminformatics* **7**(1), S6. Voir page 107.

Bayar, B. & Stamm, M. C. (2016), A deep learning approach to universal image manipulation detection using a new convolutional layer, *in* 'Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security', ACM, pp. 5–10. Voir page 30.

Berenguel, A., Terrades, O. R., Llados, J. & Canero, C. (2017*a*), E-Counterfeit : A Mobile-Server Platform for Document Counterfeit Detection, *in* 'Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)', Vol. 9, pp. 15–20. Voir page 31.

Berenguel, A., Terrades, O. R., Llados, J. & Canero, C. (2017*b*), Evaluation of Texture Descriptors for Validation of Counterfeit Documents, *in* 'Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)', Vol. 1, pp. 1237–1242. Voir page 31.

Bertrand, R., Gomez-Kramer, P., Terrades, O. R., Franco, P. & Ogier, J. M. (2013), A system based on intrinsic features for fraudulent document detection, *in* 'Proceedings of the International Conference on Document Analysis and Recognition, ICDAR', pp. 106–110. Voir pages 33 et 44.

Bertrand, R., Terrades, O. R., Gomez-Kramer, P., Franco, P. & Ogier, J.-M. (2015), A Conditional Random Field model for font forgery detection, *in* '2015 13th International Conference on Document Analysis and Recognition (ICDAR)', IEEE, pp. 576–580. Voir pages 33, 45 et 161.

Bick, E. (2003), 'Named entity recognition for danish', *I : Årbog for Nordisk Sprogteknologisk Forskningsprogram* **2004**. Voir page 70.

Birajdar, G. K. & Mankar, V. H. (2013), 'Digital image forgery detection using passive techniques : A survey', *Digital Investigation* **10**(3), 226–245. Voir pages 5, 24, 25, 28 et 30.

Böhme, R., Freiling, F. C., Gloe, T. & Kirchner, M. (2009), 'Multimedia forensics is not computer forensics', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **5718 LNCS**, 90–103. Voir pages 22 et 23.

Bondi, L., Lameri, S., Guera, D., Bestagini, P., Delp, E. J. & Tubaro, S. (2017), 'Tampering Detection and Localization Through Clustering of Camera-Based CNN Features', *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* **2017-July**, 1855–1864. Voir page 30.

Borden, S. L. & Tew, C. (2007), 'The Role of Journalist and the Performance of Journalism : Ethical Lessons From "Fake" News (Seriously)', *Journal of Mass Media Ethics* **22**(4), 300–314. Voir page 36.

Born, K. & Edgington, N. (2017), 'Analysis of philanthropic opportunities to mitigate the disinformation/propaganda problem', *Hewlett Foundation* . Voir page 36.

Brandtzaeg, P. B., Lüders, M., Spangenberg, J., Rath-Wiggins, L. & Følstad, A. (2015), 'Emerging journalistic verification practices concerning social media', *Journalism Practice* **10**(3), 323–342. Voir page 38.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140. Voir page 150.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32. Voir page 150.

Bui, Q. A., Mollard, D. & Tabbone, S. (2017), Selecting automatically pre-processing methods to improve ocr performances, *in* 'Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on', Vol. 1, IEEE, pp. 169–174. Voir page 54.

Bunk, J., Bappy, J. H., Mohammed, T. M., Nataraj, L., Flenner, A., Manjunath, B. S., Chandrasekaran, S., Roy-Chowdhury, A. K. & Peterson, L. (2017), 'Detection and Localization of Image Forgeries Using Resampling Features and Deep Learning', *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* **2017-July**, 1881–1889. Voir page 30.

Burie, J.-C., Chazalon, J., Coustaty, M., Eskenazi, S., Luqman, M. M., Mehri, M., Nayef, N., Ogier, J.-M., Prum, S. & Rusiñol, M. (2015), Icdar2015 competition on smartphone document capture and ocr (smartdoc), *in* 'Document Analysis and Recognition (ICDAR), 2015 13th International Conference on', IEEE, pp. 1161–1165. Voir page 52.

Chandrasekaran, B., Josephson, J. R. & Benjamins, V. R. (1999), 'What are ontologies, and why do we need them ?', *IEEE Intelligent Systems and their applications* **14**(1), 20–26. Voir page 73.

Chazalon, J., Gomez-Krämer, P., Burie, J.-C., Coustaty, M., Eskenazi, S., Luqman, M., Nayef, N., Rusiñol, M., Sidere, N. & Ogier, J.-M. (2017), Smartdoc 2017 video capture : Mobile document acquisition in video mode, *in* '2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)', Vol. 4, IEEE, pp. 11–16. Voir page 52.

Chen, C., McCloskey, S. & Yu, J. (2017), Image splicing detection via camera response function analysis, *in* '2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 1876–1885. Voir page 30.

Chen, N. & Blostein, D. (2007), 'A survey of document image classification : problem statement, classifier architecture and performance evaluation', *International Journal of Document Analysis and Recognition (IJDAR)* **10**(1), 1–16. Voir pages 25 et 26.

Chen, Y., Conroy, N. J. & Rubin, V. L. (2015), News in an online world : The need for an "automatic crap detector", *in* 'Proceedings of the Association for Information Science and Technology', Vol. 52, pp. 1–4. Voir page 38.

Chiron, G., Doucet, A., Coustaty, M. & Moreux, J.-P. (2017), Icdar2017 competition on post-ocr text correction, *in* 'Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on', Vol. 1, IEEE, pp. 1423–1428. Voir page 58.

Choo, C. W., Detlor, B. & Turnbull, D. (2000), *Web Work : Information Seeking and Knowledge Work on the World Wide Web*, Vol. 1 of *Information Science and Knowledge Management*, Springer Netherlands, Dordrecht. Voir page 69.

Christlein, V., Riess, C., Jordan, J., Riess, C. & Angelopoulou, E. (2012), 'An evaluation of popular copy-move forgery detection approaches', *IEEE Transactions on Information Forensics and Security* **7**(6), 1841–1854. Voir page 29.

Cimiano, P. & Völker, J. (2005), 'Text2Onto : A Framework for Ontology Learning and Data-Driven Change Discovery', *Natural Language Processing and Information Systems* pp. 227–238. Voir page 73.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P. (2011), 'Natural Language Processing (almost) from Scratch', *Journal of Machine Learning Research* **12**(Aug), 2493–2537. Voir page 71.

Conlon, S. J., Abrahams, A. S. & Simmons, L. L. (2015), 'Terrorism information extraction from online reports', *Journal of Computer Information Systems* **55**(3), 20–28. Voir page 72.

Conroy, N. J., Rubin, V. L. & Chen, Y. (2015), Automatic Deception Detection : Methods for Finding Fake News, *in* 'Proceedings of the 78th ASIST Annual Meeting : Information Science with Impact : Research in and for the Community', number October, p. 82. Voir page 39.

Conti, M., Lain, D., Lazzeretti, R., Lovisotto, G. & Quattrociocchi, W. (2017), It's always april fools' day ! : On the difficulty of social network misinformation classification via propagation features, *in* 'Information Forensics and Security (WIFS), 2017 IEEE Workshop on', IEEE, pp. 1–6. Voir page 39.

Couäsnon, B. (2001), Dmos : a generic document recognition method, application to an automatic generator of musical scores, mathematical formulae and table structures recognition systems, *in* 'Proceedings of Sixth International Conference on Document Analysis and Recognition', pp. 215–220. Voir page 165.

Cox, I., Miller, M., Bloom, J., Fridrich, J. & Kalker, T. (2008), *Digital Watermarking and Steganography*, 2 edn, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. Voir pages 23, 26 et 27.

Cozzolino, D., Gragnaniello, D. & Verdoliva, L. (2014), Image forgery detection through residual-based local descriptors and block-matching, *in* 'International Conference on Image Processing (ICIP)', IEEE, pp. 5297–5301. Voir page 132.

Cozzolino, D., Poggi, G. & Verdoliva, L. (2015), 'Efficient dense-field copy–move forgery detection', *IEEE Transactions on Information Forensics and Security* **10**(11), 2284–2297. Voir pages 29 et 131.

Cozzolino, D. & Verdoliva, L. (2018*a*), 'Camera-based image forgery localization using convolutional neural networks', *submitted* . Voir page 131.

Cozzolino, D. & Verdoliva, L. (2018*b*), 'Noiseprint : a cnn-based camera model fingerprint', *submitted* . Voir page 131.

Cruz, F., Sidere, N., Coustaty, M., D'Andecy, V. P. & Ogier, J.-M. (2017), 'Local Binary Patterns for Document Forgery Detection', *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* pp. 1223–1228. Voir page 31.

Cunningham, H., Maynard, D., Bontcheva, K. & Tablan, V. (2002), GATE : A framework and graphical development environment for robust NLP tools and applications., *in*

'Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)', Vol. 19, pp. 168–175. Voir page 73.

de Marneffe, M.-C., Padó, S. & Manning, C. D. (2009), Multi-word expressions in textual inference : Much ado about nothing ?, *in* 'Proceedings of the 2009 Workshop on Applied Textual Inference', Association for Computational Linguistics, pp. 1–9. Voir page 106.

Dhiman, S. & Singh, O. (2016), 'Analysis of Visible and Invisible Image Watermarking – A Review', *International Journal of Computer Applications* **147**(3), 975–8887. Voir page 27.

Di Giovanni, J.-L. (2016), 'La fraude explose en France', *Global Economic Crime Survey 2016* . Voir page 10.

Durtschi, C., Hillison, W. & Pacini, C. (2004), 'The effective use of benford's law to assist in detecting fraud in accounting data', *Journal of forensic accounting* **5**(1), 17–34. Voir page 137.

Eibe, F., Hall, M. & Witten, I. (2016), 'The weka workbench. online appendix for" data mining : Practical machine learning tools and techniques', *Morgan Kaufmann* . Voir page 148.

Elkasrawi, S., Dengel, A., Abdelsamad, A. & Bukhari, S. S. (2016), What you see is what you get ? automatic image verification for online news content, *in* 'Document Analysis Systems (DAS), 2016 12th IAPR Workshop on', IEEE, pp. 114–119. Voir page 40.

Elkasrawi, S. & Shafait, F. (2014), Printer identification using supervised learning for document forgery detection, *in* 'Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on', IEEE, pp. 146–150. Voir page 32.

Eskenazi, S. (2017), On the stability of document analysis algorithms : application to hybrid document hashing technologies, PhD thesis, Université de La Rochelle. Voir pages 27 et 34.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S. & Yates, A. (2004), Web-scale information extraction in KnowItAll, *in* 'Proceedings of the 13th International Conference on World Wide Web', WWW '04, pp. 100–110. Voir page 72.

Fang, Z., Wang, S. & Zhang, X. (2009), Image splicing detection using camera characteristic inconsistency, *in* 'Proceedings of International Conference on Multimedia Information Networking and Security,', Vol. 1, pp. 20–24. Voir page 30.

Farid, H. (2009), 'Exposing digital forgeries from jpeg ghosts', *IEEE Transactions on Information Forensics and Security* **4**(1), 154–160. Voir page 30.

Feng, S., Banerjee, R. & Choi, Y. (2012), Syntactic stylometry for deception detection, *in* 'Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Short Papers-Volume 2', Association for Computational Linguistics, pp. 171–175. Voir page 39.

Fotsoh, A. (2018), Recherche d'entités nommées complexes sur le Web - propositions pour l'extraction et pour le calcul de similarité, PhD thesis, Université de Pau et des Pays de l'Adour. Voir pages 72 et 159.

Fridrich, J. & Kodovsky, J. (2012), 'Rich models for steganalysis of digital images', *IEEE Transactions on Information Forensics and Security* **7**(3), 868–882. Voir page 132.

Fridrich, J., Soukal, D. & Lukáš, J. (2003), Detection of Copy-Move Forgery in Digital Images, *in* 'Proceedings of Digital Forensic Research Workshop', Vol. 3, pp. 652–663. Voir pages 29 et 130.

Friedman, N., Geiger, D. & Goldszmidt, M. (1997), 'Bayesian network classifiers', *Machine learning* **29**(2-3), 131–163. Voir page 149.

Gaudan, S., Kirsch, H. & Rebholz-Schuhmann, D. (2005), 'Resolving abbreviations to their senses in medline', *Bioinformatics* **21**(18), 3658–3664. Voir page 107.

Goasdoué, F., Karanasos, K., Katsis, Y., Leblay, J., Manolescu, I. & Zampetakis, S. (2013), Fact checking and analyzing the web, *in* 'Proceedings of the 2013 international conference on Management of data - SIGMOD '13', p. 997. Voir pages 38 et 40.

Grevisse, M. & Lits, M. (2009), *Le petit Grevisse : Grammaire française*, Grevisse Langue Française, De Boeck Secondaire. Voir pages 105 et 106.

Grishman, R. & Sundheim, B. (1996), Message Understanding Conference-6, *in* 'Proceedings of the 16th conference on Computational linguistics -', Vol. 1, Association for Computational Linguistics, Morristown, NJ, USA, p. 466. Voir pages 70 et 72.

Gruber, T. R. (1993), 'A translation approach to portable ontology specifications', *Knowledge acquisition* **5**(2), 199–220. Voir page 73.

Hazman, M., R. El-Beltagy, S. & Rafea, A. (2011), 'A Survey of Ontology Learning Approaches', *International Journal of Computer Applications* **22**(9), 36–43. Voir page 73.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), 'Deep residual learning for image recognition'. Voir page 130.

Hearst, M. A. (1992), Automatic acquisition of hyponyms from large text corpora, *in* 'Proceedings of the 14th conference on Computational linguistics', Vol. 2, Association for Computational Linguistics, pp. 539–545. Voir page 74.

Ho, A. T. & Li, S. (2015), *Handbook of Digital Forensics of Multimedia Data and Devices*, Wiley. Voir page 23.

Hobbs, J. R. & Riloff, E. (2010), 'Information extraction.', *Handbook of natural language processing, Second Edition* **2**. Voir page 71.

Hogenboom, F., Frasincar, F., Kaymak, U., De Jong, F. & Caron, E. (2016), 'A Survey of event extraction methods from text for decision support systems', *Decision Support Systems* **85**, 12–22. Voir page 72.

Horne, B. D. & Adali, S. (2017), 'This Just In : Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News'. Voir page 40.

Jean-Caurant, A., Tamani, N., Courboulay, V. & Burie, J.-C. (2017), Lexicographical-based order for post-ocr correction of named entities, *in* 'Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on', Vol. 1, IEEE, pp. 1192–1197. Voir page 57.

John, G. H. & Langley, P. (1995), Estimating continuous distributions in bayesian classifiers, *in* 'Eleventh Conference on Uncertainty in Artificial Intelligence', Morgan Kaufmann, San Mateo, pp. 338–345. Voir page 149.

Joseph, A. (2008), 'Etude comparative et synchronique du langage SMS en italien, français et anglais', (January 2008). Voir page 106.

Kanungo, T., Haralick, R. M. & Phillips, I. (1993), Global and local document degradation models, *in* 'In Proceedings of the International Conference on Document Analysis and Recognition', pp. 730–734. Voir page 44.

Kaur, A. & Richa, S. (2013), 'Copy-Move Forgery Detection using DCT and SIFT', *International Journal of Computer Applications* **70**(7), 30–34. Voir page 29.

Khanna, N. & Delp, E. J. (2009), Source scanner identification for scanned documents, *in* 'Proceedings of the 2009 1st IEEE International Workshop on Information Forensics and Security, WIFS 2009', pp. 166–170. Voir page 32.

Khanna, N. & Delp, E. J. (2010), 'Intrinsic signatures for scanned documents forensics : Effect of font shape and size', *ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems : Nano-Bio Circuit Fabrics and Systems* pp. 3060–3063. Voir page 32.

Kirchner, M. (2008), Fast and reliable resampling detection by spectral analysis of fixed linear predictor residue, *in* 'Proceedings of the 10th ACM workshop on Multimedia and security - MM&Sec '08', p. 11. Voir page 30.

Kissos, I. & Dershowitz, N. (2016), Ocr error correction using character correction and feature-based word classification, *in* 'Document Analysis Systems (DAS), 2016 12th IAPR Workshop on', IEEE, pp. 198–203. Voir page 57.

Klein, B. (2016), 'Python Advanced : Recursive and Iterative Implementation of the Edit Distance'.
URL: *https ://www.python-course.eu/levenshtein_ distance.php* Voir page 120.

Korus, P. (2017), 'Digital image integrity – a survey of protection and verification techniques', *Digital Signal Processing : A Review Journal* **71**, 1–26. Voir pages 5, 24, 28 et 29.

Krautgartner, K. (2003), 'Techniques d'abréviation dans les webchats francophones', *Linguistik Online* **15**(3). Voir page 106.

Kripke, S. (1980), *Naming and Necessity*, Harvard University Press. Voir page 70.

Lampert, C. H., Mei, L. & Breuel, T. M. (2006), Printing technique classification for document counterfeit detection, *in* 'Computational Intelligence and Security, 2006 International Conference on', Vol. 1, IEEE, pp. 639–644. Voir pages 31 et 32.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. & Dyer, C. (2016), Neural architectures for named entity recognition, *in* 'Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)'. Voir page 71.

Lamy, J.-B. (2017), 'Owlready : Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies', *Artificial intelligence in medicine* **80**, 11–28. Voir page 84.

Langlotz, T. & Bimber, O. (2007), Unsynchronized 4d barcodes, *in* 'International Symposium on Visual Computing', Springer, pp. 363–374. Voir page 27.

Larkey, L. S., Ogilvie, P., Price, M. A. & Tamilio, B. (2000), Acrophile : An automated acronym extractor and server, *in* 'Proceedings of the ACM Fifth International Conference on Digital Libraries, DL'00, Dallas TX', ACM Press, pp. 205–214. Voir page 107.

Lavion, D. (2018), 'Pulling fraud out of the shadows', *2018 Global Economic Crime and Fraud Survey* . Voir page 10.

Lehmann, J. & Völker, J. (2014), An Introduction to Ontology Learning, *in* 'Perspectives on Ontology Learning', IOS Press, Amsterdam, The Netherlands, pp. 9–14. Voir page 73.

Lei, Y., Wang, Y. & Huang, J. (2011), 'Robust image hash in radon transform domain for authentication', *Signal Processing : Image Communication* **26**(6), 280 – 288. Voir page 27.

Lejeune, G., Brixtel, R., Doucet, A. & Lucas, N. (2015), 'Multilingual event extraction for epidemic detection', *Artificial Intelligence in Medicine* **65**(2), 131–143. Voir page 72.

Levenshtein, V. I. (1966), Binary codes capable of correcting deletions, insertions, and reversals, *in* 'Soviet physics doklady', Vol. 10, pp. 707–710. Voir page 107.

Liu, X., Zhou, M., Wei, F., Fu, Z. & Zhou, X. (2012), Joint inference of named entity recognition and normalization for tweets, *in* 'Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1', Association for Computational Linguistics, pp. 526–535. Voir page 106.

Lowe, D. G. (2004), 'Distinctive image features from scale invariant keypoints', *International Journal of Computer Vision* **60**, 91–11020042. Voir page 30.

Lukasik, M., Cohn, T. & Bontcheva, K. (2015), Point process modelling of rumour dynamics in social media, *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)', Vol. 2, pp. 518–523. Voir page 39.

Ma, J., Gao, W., Wei, Z., Lu, Y. & Wong, K.-F. (2015), Detect rumors using time series of social context information on microblogging websites, *in* 'Proceedings of the 24th ACM International on Conference on Information and Knowledge Management', ACM, pp. 1751–1754. Voir page 39.

Madani, A., Boussaid, O. & Zegour, D. E. (2013), 'Semi-structured documents mining : A review and comparison', *Procedia Computer Science* **22**, 330 – 339. 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013. Voir page 15.

Maedche, A. & Staab, S. (2001), 'Ontology learning for the Semantic Web', *IEEE Intelligent Systems* **16**(2), 72–79. Voir page 73.

Maergner, P., Riesen, K., Ingold, R. & Fischer, A. (2017), A Structural Approach to Offline Signature Verification Using Graph Edit Distance, *in* '2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)', pp. 1216–1222. Voir page 34.

Magdy, A. & Wanas, N. (2010), Web-based statistical fact checking of textual documents, *in* 'Proceedings of the 2nd international workshop on Search and mining user-generated contents', ACM, pp. 103–110. Voir pages 40 et 159.

Mahdian, B. & Saic, S. (2010), 'A bibliography on blind methods for identifying image forgery', *Signal Processing : Image Communication* **25**(6), 389 – 399. Voir page 28.

Mahdian, B. & Stanislav, S. (2008), 'Blind authentication using periodic properties of interpolation', *IEEE Transactions on Information Forensics and Security* **3**(3), 529–538. Voir page 30.

Martinet, A. (1967), *Éléments de Linguistique Générale ...*, Collection Armand Colin, no. 340. Section de litérature, Librairie Armand Colin. Voir page 106.

Maynard, D., Funk, A. & Peters, W. (2009), SPRAT : a tool for automatic semantic pattern-based ontology population, *in* 'International conference for digital libraries and the semantic web (ICSD)'. Voir page 74.

Maynard, D., Peters, W. & Li, Y. (2006), 'Metrics for evaluation of ontology-based information extraction', *CEUR Workshop Proceedings* **179**. Voir pages 74 et 94.

Mei, J., Islam, A., Wu, Y., Moh'd, A. & Milios, E. E. (2016), 'Statistical Learning for OCR Text Correction'. Voir page 57.

Micenková, B., Van Beusekom, J. & Shafait, F. (2015), 'Stamp verification for automated document authentication', *Computational Forensics* **8915**, 117–129. Voir pages 33 et 34.

Mikkilineni, A. K., Ali, G. N., Chiang, P.-J., Chiu, G. T., Allebach, J. P. & Delp, E. J. (2004), Signature-embedding in printed documents for security and forensic applications, *in* 'Security, Steganography, and Watermarking of Multimedia Contents VI', Vol. 5306, International Society for Optics and Photonics, pp. 455–467. Voir page 27.

Misra, H., Cappé, O. & Yvon, F. (2008), Using lda to detect semantically incoherent documents, *in* 'Proceedings of the Twelfth Conference on Computational Natural Language Learning', Association for Computational Linguistics, pp. 41–48. Voir page 40.

Mokhtar, K., Bukhari, S. S. & Dengel, A. (2018), Ocr error correction : State-of-the-art vs an nmt-based approach, *in* '2018 13th IAPR International Workshop on Document Analysis Systems (DAS)', IEEE, pp. 429–434. Voir page 57.

Musen, M. A. (2015), 'The protégé project', *AI Matters* **1**(4), 4–12. Voir page 75.

Muslea, I. (1999), Extraction Patterns for Information Extraction Tasks : A Survey, *in* 'The AAAI99 Workshop on Machine Learning for Information Extraction', Vol. 2. Voir page 70.

Nadeau, D. & Sekine, S. (2007), 'A survey of named entity recognition and classification', *Linguisticae Investigationes* **30**(1), 3–26. Voir pages 70 et 71.

Nadeau, D. & Turney, P. D. (2005), A supervised learning approach to acronym identification, *in* 'Proceedings of the 18th Canadian Society Conference on Advances in Artificial Intelligence', AI'05, Springer-Verlag, Berlin, Heidelberg, pp. 319–329. Voir page 107.

Navarro, G. (2001), 'A guided tour to approximate string matching', *ACM Computing Surveys* **33**(1), 31–88. Voir page 108.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L. & Nielsen, R. K. (2018), 'Reuters institute digital news report 2018'. Voir pages 34 et 36.

Ng, T. & Chang, S. (2004), A model for image splicing, *in* '2004 International Conference on Image Processing, 2004. ICIP '04.', Vol. 2, pp. 1169–1172 Vol.2. Voir page 30.

Nigrini, M. J. (2012), *Benford's Law : Applications for forensic accounting, auditing, and fraud detection*, Vol. 586, John Wiley & Sons. Voir page 137.

Okawa, M. (2018), Offline Signature Verification with VLAD Using Fused KAZE Features from Foreground and Background Signature Images, *in* 'Proceedings of the International Conference on Document Analysis and Recognition, ICDAR', Vol. 1, pp. 1198–1203. Voir page 34.

OLAF (2014), 'Détection de faux documents dans le cadre des actions structurelles, guide pratique à l'intention des autorités de gestion'. Voir pages 14, 22 et 24.

Paumier, S. (2003), De la reconnaissance des formes linguistiques à l'analyse syntaxique, PhD thesis, Université Marne-la-Vallée. Voir page 71.

Piva, A. (2013), 'An Overview on Image Forensics', *ISRN Signal Processing* **2013**(2), 1–22. Voir pages 24 et 28.

Poibeau, T. (2003), *Extraction automatique d'information : Du texte brut au web sémantique*, Lavoisier. ISBN 2-7462-0610-2. Voir page 71.

Poibeau, T., Saggion, H., Piskorski, J. & Yangarber, R. (2012), *Multi-source, multilingual information extraction and summarization*, Springer Science & Business Media. Voir page 71.

Poisel, R. & Tjoa, S. (2011), Forensics investigations of multimedia data : A review of the state-of-the-art, *in* 'Proceedings - 6th International Conference on IT Security Incident Management and IT Forensics, IMF 2011', pp. 48–61. Voir pages 23 et 28.

Popescu, A. C. & Farid, H. (2004), 'Exposing digital forgeries by detecting duplicated image regions', *Dept. Comput. Sci., Dartmouth College, Tech. Rep. TR2004-515* pp. 1–11. Voir page 29.

Popescu, A. C. & Farid, H. (2005), 'Exposing digital forgeries by detecting traces of resampling', *IEEE Transactions on signal processing* **53**(2), 758–767. Voir page 30.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J. & Stein, B. (2017), 'A Stylometric Inquiry into Hyperpartisan and Fake News'. Voir page 39.

Quinlan, R. (1993), *C4.5 : Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA. Voir page 149.

Rajapaksha, P., Farahbakhsh, R. & Crespi, N. (2017), Identifying content originator in social networks, *in* 'GLOBECOM 2017-2017 IEEE Global Communications Conference', IEEE, pp. 1–6. Voir page 39.

Rao, Y. & Ni, J. (2016), A deep learning approach to detection of splicing and copy-move forgeries in images, *in* '2016 IEEE International Workshop on Information Forensics and Security (WIFS)', pp. 1–6. Voir page 30.

Rastier, F. (2005), 'Enjeux épistémologiques de la linguistique de corpus', *La linguistique de corpus* pp. 31–45. Voir page 46.

Reynaert, M. (2014), On ocr ground truths and ocr post-correction gold standards, tools and formats, *in* 'Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage', ACM, pp. 159–166. Voir page 57.

Riegel, M., Pellat, J.-C. & Rioul, R. (2016), *Grammaire méthodique du français*, Presses universitaires de France. Voir page 105.

Rigaud, C., Tsopze, N., Burie, J.-C. & Ogier, J.-M. (2013), Robust frame and text extraction from comic books, *in* Y.-B. Kwon & J.-M. Ogier, eds, 'Graphics Recognition. New Trends and Challenges', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 129–138. Voir page 52.

Roche, M. & Prince, V. (2007), 'AcroDef : A Quality Measure for Discriminating Expansions of Ambiguous Acronyms', *Modeling and Using Context* pp. 411–424. Voir page 106.

Rota, P., Sangineto, E., Conotter, V. & Pramerdorfer, C. (2017), 'Bad teacher or unruly student : Can deep learning say something in Image Forensics analysis ?', pp. 2503–2508. Voir page 30.

Rowley, J. (2007), 'The wisdom hierarchy : representations of the dikw hierarchy', *Journal of Information Science* **33**(2), 163–180. Voir pages 69 et 70.

Rubin, V., Conroy, N., Chen, Y. & Cornwell, S. (2016), Fake news or truth ? using satirical cues to detect potentially misleading news, *in* 'Proceedings of the Second Workshop on Computational Approaches to Deception Detection', pp. 7–17. Voir page 40.

Rubin, V. L., Chen, Y. & Conroy, N. J. (2015), Deception Detection for News : Three Types of Fake News, *in* 'Proceedings of the Association for Information Science and Technology', Vol. 52, pp. 1–4. Voir page 38.

Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E. & Suter, B. W. (1990), 'The multilayer perceptron as an approximation to a bayes optimal discriminant function', *IEEE Transactions on Neural Networks* **1**(4), 296–298. Voir page 150.

Sagot, B. & Gábor, K. (2014), Détection et correction automatique d'entités nommées dans des corpus OCRisés, *in* 'Traitement Automatique du Langage Naturel 2014'. Voir page 57.

Schetinger, V., Oliveira, M. M., da Silva, R. & Carvalho, T. J. (2017), 'Humans are easily fooled by digital images', *Computers and Graphics (Pergamon)* **68**, 142–151. Voir page 17.

Schifferes, S., Newman, N., Thurman, N., Corney, D., Göker, A. & Martin, C. (2014), 'Identifying and Verifying News through Social Media : Developing a user-centred tool for professional journalists', *Digital Journalism* **2**(3), 406–418. Voir page 38.

Schulze, C., Schreyer, M., Stahl, A. & Breuel, T. (2009), Using Dct Features for Printing Technique and Copy Detection, *in* 'IFIP International Conference on Digital Forensics', pp. 95–106. Voir pages 31 et 32.

Serrano, L. (2014), Vers une capitalisation des connaissances orientée utilisateur : extraction et structuration automatiques de l'information issue de sources ouvertes, PhD thesis, Université de Caen. Voir page 70.

Shah, R. & Jain, S. (2014), 'Ontology-based Information Extraction : An Overview and a Study of different Approaches', *International Journal of Computer Applications* **87**(4), 6–8. Voir page 74.

Shang, S., Memon, N. & Kong, X. (2014), 'Detecting documents forged by printing and copying', *EURASIP Journal on Advances in Signal Processing* **2014**(1), 140. Voir pages 5 et 32.

Sharnagat, R. (2014), 'Named entity recognition : A literature survey', *Center For Indian Language Technology* . Voir pages 70 et 71.

Shu, K., Sliva, A., Wang, S., Tang, J. & Liu, H. (2017), 'Fake news detection on social media : A data mining perspective', *ACM SIGKDD Explorations Newsletter* **19**(1), 22–36. Voir page 39.

Sidère, N., Cruz, F., Coustaty, M. & Ogier, J. M. (2017), 'A dataset for forgery detection and spotting in document images', *Proceedings - 2017 7th International Conference on Emerging Security Technologies, EST 2017* pp. 26–31. Voir pages 45 et 62.

Silberztein, M. (2007), 'Complex annotations with nooj', *Proceedings of the 2007 International NooJ Conference* . Voir page 71.

Simperl, E., Bürger, T., Hangl, S., Wörgl, S. & Popov, I. (2012), 'ONTOCOM : A reliable cost estimation method for ontology development projects', *Journal of Web Semantics* **16**, 1–16. Voir page 73.

Srikantaiah, K., Suraj, M., Venugopal, K. & Patnaik, L. M. (2013), 'Similarity based dynamic web data extraction and integration system from search engine result pages for web content mining', *ACEEE International Journal on Information Technology* **3**(1), 42–49. Voir page 159.

Stamatatos, E. (2009), 'A survey of modern authorship attribution methods', *Journal of the American Society for Information Science and Technology* **60**(3), 538–556. Voir page 39.

Stamm, M. C., Wu, M. & Liu, K. J. (2013), 'Information forensics : An overview of the first decade', *IEEE Access* **1**, 167–200. Voir pages 23 et 24.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S. & de Alfaro, L. (2017), Some like it Hoax : Automated fake news detection in social networks, *in* 'Proceedings of the Second Workshop on Data Science for Social Good (SoGood), Skopje, Macedonia, 2017. CEUR Workshop Proceedings', Vol. 1960, pp. 1–12. Voir page 39.

Tchan, J. (2004), The development of an image analysis system that can detect fraudulent alterations made to printed images, *in* R. L. van Renesse, ed., 'Electronic Imaging 2004', International Society for Optics and Photonics, pp. 151–159. Voir page 32.

Teissèdre, C. (2012), Analyse sémantique automatique des adverbiaux de localisation temporelle : application à la recherche d'information et à l'acquisition de connaissances., Theses, Université de Nanterre - Paris X. Voir page 70.

Teyssou, D., Leung, J.-M., Apostolidis, E., Apostolidis, K., Papadopoulos, S., Zampoglou, M., Papadopoulou, O. & Mezaris, V. (2017), 'The InVID Plug-in : Web Video Verification on the Browser'. Voir page 38.

Thompson, P., Mcnaught, J. & Ananiadou, S. (2015), Customised OCR Correction for Historical Medical Text, *in* G. Guidi, R. Scopigno & F. Remondino, eds, 'International Congress on Digital Heritage - Theme 1 - Digitization And Acquisition', IEEE. Voir page 57.

Tkachenko, I. (2015), Generation and analysis of graphical codes using textured patterns for printed document authentication, Theses, Université de Montpellier. Voir page 27.

Tucker, J. A., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D. & Nyhan, B. (2018), 'Social Media, Political Polarization, and Political Disinformation : A Review of the Scientific Literature', *Ssrn* (March), 1–95. Voir pages 35 et 64.

UNODC (2010), 'Guide for the development of forensic document examination capacity'. Voir pages 15, 22 et 26.

van Beusekom, J., Shafait, F. & Breuel, T. (2009), Automatic Line Orientation Measurement for Questioned Document Examination, *in* 'Computational Forensics, Proceedings', Vol. 5718, pp. 165–173. Voir pages 33 et 45.

van Beusekom, J., Shafait, F. & Breuel, T. M. (2010), Document inspection using textline alignment, *in* '{DAS} '10 : {P}roceedings of the 9th {IAPR} {I}nternational {W}orkshop on {D}ocument {A}nalysis {S}ystems', pp. 263–270. Voir pages 33 et 45.

van Beusekom, J., Stahl, A. & Shafait, F. (2015), 'Lessons learned from automatic forgery detection in over 100,000 invoices', *Computational Forensics* **8915**, 130–142. Voir page 45.

van Renesse, R. (1997), 'Paper based document security - a review', *European Conference on Security and Detection - ECOS97 Incorporating the One Day Symposium on Technology Used for Combatting Fraud* **1997**, 75–80. Voir pages 15 et 18.

Villán, R., Voloshynovskiy, S., Koval, O., Deguillaume, F. & Pun, T. (2007), Tamper-proofing of electronic and printed text documents via robust hashing and data-hiding, *in* 'Security, Steganography, and Watermarking of Multimedia Contents IX', Vol. 6505, International Society for Optics and Photonics, p. 65051T. Voir page 27.

Wang, W. Y. (2017), '"Liar, Liar Pants on Fire" : A New Benchmark Dataset for Fake News Detection'. Voir page 38.

Wimalasuriya, D. C. & Dou, D. (2010), 'Ontology-based information extraction : An introduction and a survey of current approaches', *Journal of Information Science* **36**(3), 306–323. Voir page 74.

Witten, I. H., Bray, Z., Mahoui, M. & Teahan, W. J. (1999), 'Using Language Models for Generic Entity Extraction', *Proceedings of the International Conference on Machine Learning Workshop on Text Mining* (August), 1–11. Voir page 70.

Wu, D., Zhou, X. & Niu, X. (2009), 'A novel image hash algorithm resistant to print–scan', *Signal processing* **89**(12), 2415–2424. Voir page 27.

Wu, Y., Denny, J. C., Rosenbloom, S. T., Miller, R. A., Giuse, D. A. & Xu, H. (2012), A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries, *in* 'AMIA annual symposium proceedings', Vol. 2012, American Medical Informatics Association, p. 997. Voir page 107.

Yeates, S. (1999), Automatic extraction of acronyms from text, *in* 'New Zealand Computer Science Research Students' Conference', pp. 117–124. Voir page 107.

Yu, H., Hripcsak, G. & Friedman, C. (2002), 'Mapping abbreviations to full forms in biomedical articles', *Journal of the American Medical Informatics Association* **9**(3), 262–272. Voir page 107.

Yu, M., Li, G., Deng, D. & Feng, J. (2016), 'String similarity search and join : a survey', *Frontiers of Computer Science* **10**(3), 399–417. Voir page 107.

Zaidan, B. B., Zaidan, A. A., Karim, H. A. & Ahmad, N. N. (2017), 'A new digital watermarking evaluation and benchmarking methodology using an external group of evaluators and multi-criteria analysis based on 'large-scale data'', *Software - Practice and Experience* **47**(10), 1365–1392. Voir page 27.

Zubiaga, A., Liakata, M., Procter, R., Bontcheva, K. & Tolmie, P. (2015), Towards detecting rumours in social media., *in* 'AAAI Workshop : AI for Cities'. Voir page 38.

Chloé ARTAUD

Détection des fraudes : de l'image à la sémantique du contenu

Résumé : Les entreprises, les administrations, et parfois les particuliers, doivent faire face à de nombreuses fraudes sur les documents qu'ils reçoivent de l'extérieur ou qu'ils traitent en interne. Les factures, les notes de frais, les justificatifs... tout document servant de preuve peut être falsifié dans le but de gagner plus d'argent ou de ne pas en perdre. En France, on estime les pertes dues aux fraudes à plusieurs milliards d'euros par an. Etant donné que le flux de documents échangés, numériques ou papiers, est très important, il serait extrêmement coûteux en temps et en argent de les faire tous vérifier par des experts de la détection des fraudes. C'est pourquoi nous proposons dans notre thèse un système de détection automatique des faux documents.
Si la plupart des travaux en détection automatique des faux documents se concentrent sur des indices graphiques, nous cherchons quant à nous à vérifier les informations textuelles du document afin de détecter des incohérences ou des invraisemblances. Pour cela, nous avons tout d'abord constitué un corpus de tickets de caisse que nous avons numérisés et dont nous avons extrait le texte. Après avoir corrigé les sorties de l'OCR et fait falsifier une partie des documents, nous en avons extrait les informations et nous les avons modélisées dans une ontologie, afin de garder les liens sémantiques entre elles. Les informations ainsi extraites, et augmentées de leurs possibles désambiguïsations, peuvent être vérifiées les unes par rapport aux autres au sein du document et à travers la base de connaissances constituée. Les liens sémantiques de l'ontologie permettent également de chercher l'information dans d'autres sources de connaissances, et notamment sur Internet.
Mots clés : détection des faux documents, corpus de documents, extraction d'information, ontologie, abréviations, fausses informations

Fraud detection : from image to semantics of content

Summary : Companies, administrations, and sometimes individuals, have to face many frauds on documents they receive from outside or process internally. Invoices, expense reports, receipts... any document used as proof can be falsified in order to earn more money or not to lose it. In France, losses due to fraud are estimated at several billion euros per year. Since the flow of documents exchanged, whether digital or paper, is very important, it would be extremely costly and time-consuming to have them all checked by fraud detection experts. That's why we propose in our thesis a system for automatic detection of false documents.
While most of the work in automatic document detection focuses on graphic clues, we seek to verify the textual information in the document in order to detect inconsistencies or implausibilities. To do this, we first compiled a corpus of documents that we digitized. After correcting the characters recognition outputs and falsifying part of the documents, we extracted the information and modelled them in an ontology, in order to keep the semantic links between them. The information thus extracted, and increased by its possible disambiguation, can be verified against each other within the document and through the knowledge base established. The semantic links of ontology also make it possible to search for information in other sources of knowledge, particularly on the Internet.
Keywords : false documents detection, dataset of documents, information extraction, ontology, abbreviation, fake information

RÉGION Nouvelle-Aquitaine
Informatique Image Interaction Univ. La Rochelle
DGA