Évaluation quantitative de la circulation de l'information sur les sites de presse en ligne

Noé Faure

2019 - 2020

Contents

List of Figures

1 Introduction

1.1 Présentation du sujet

Les recherches qui ont conduit à l'aboutissement de ce mémoire commencent par un premier constat : la recherche d'informations concernant un sujet précis sur les sites de presse en ligne est très souvent obstruée par une répétition conséquente de l'information qui s'y trouve. Deux articles issus de journaux différents peuvent apporter les mêmes informations sans qu'une quelconque valeur supplémentaire vienne les différencier. Nous est alors venue l'idée d'essayer de mesurer la quantité d'information inédite se trouvant dans un article en fonction du contenu des autres. Cette tentative s'est rapidement confrontée à une impasse, comme nous le verrons, car elle aurait signifié une capacité à mesurer le "sens" des mots que les articles contiennent. Cette problématique est très liée à l'incapacité persistante de donner une définition satisfaisante du concept d'information. Cela constituera notre première partie.
Étant dans l'impossibilité de mesurer l'information "nouvelle", nous avons alors pris le problème à rebours en essayant de mesurer l'information "répétée". Pour ce faire, nous avons établi un certain nombre d'indicateurs dont nous expliquerons le fonctionnement, les modalités et les limites afin de mesurer cette répétition. Ce sera l'objet de notre deuxième partie. Ces indicateurs seront par la suite mis en application à travers des études de cas portant sur des sujets très concrets ayant été traités par la presse. Nous comparerons notamment des dépêches de l'AFP avec différents articles portant sur les mêmes sujets. Dans un dernier temps, ces résultats seront l'occasion d'une analyse puis d'une conclusion sur la circulation de l'information sur les sites de presse en ligne.

2 Qu'est-ce que l'information ?

2.1 Étymologie et champs d'utilisation

Pour nous permettre de déterminer si un article contient davantage d'informations qu'un autre il est nécessaire de définir proprement ce qu'est l'information. La définition du terme d'information est à elle seule l'objet de nombreuses études. Étymologiquement le mot « information » provient du verbe transitif latin informare. Le dictionnaire illustré latin-français (Gaffiot) en donne plusieurs définitions :

- Façonner, former
- Représenter idéalement, décrire
- Façonner, disposer, organiser
- Former dans l'esprit
- Se représenter par la pensée, se faire une idée de

Par la suite le terme "information" a été réutilisé et redéfini de façons différentes dans de nombreux domaines parmi lesquels : la neurologie, le droit, la sociologie, les sciences de l'information ou l'informatique. À noter que le terme même d'informatique est construit sur le mot "information" auquel est apposé le suffixe "-ique" signifiant « relatif à » ou « qui est propre à ». La multiplicité des domaines dans lesquels le terme "information" trouve un sens est d'ailleurs très certainement la raison principale pour laquelle il est difficile de lui trouver une définition qui fasse consensus. Pour répondre à notre problématique, seules les définitions de deux domaines d'étude semblent pertinentes : celle des sciences sociales (en lien avec le journalisme) et celle de l'informatique. Il est toutefois important de remarquer que ces définitions ne sont pas immuables et sont l'objet de redéfinitions fréquentes. Remarquons également que les domaines de la recherche sont perméables et s'influencent les uns les autres. Les récents progrès en lien avec l'intelligence artificielle par exemple entrainent un dialogue nécessaire entre les neurologues et les chercheurs en informatique autour de la définition du terme d'information.

2.2 Définition du terme d'information en informatique

Le dictionnaire Larousse définit une information dans le domaine de l'informatique comme : « (un) élément de connaissance susceptible d'être représenté à l'aide de conventions pour être conservé, traité ou communiqué. ». Il est très fréquent que des systèmes informatiques proposent des services qui suivent cette logique : importer, traiter, exporter. Notre démarche au sein même de ce mémoire se rapproche de cette méthode.

L'unité mixte de recherche ATILF (Analyse et traitement informatique de la

langue française) à travers son dictionnaire TLF (Trésor de la langue française) qui regroupe des définitions issues de dictionnaires français du XIXème et XXème siècle propose trois définitions notables du terme d'information :

- Faits, événements nouveaux, en tant qu'ils sont connus, devenus publics.
- Fait, événement d'intérêt général traité et rendu public par la presse, la radio, la télévision.
- Ensemble de connaissances réunies sur un sujet déterminé.

La rédaction de ce dictionnaire a été achevée en 1994, selon les propres mots de cette unité de recherche : « Cette ressource, qui ne fait pas l'objet d'une veille lexicographique, est donc close "en l'état". Il est donc tout à fait naturel que les définitions qui s'y trouvent ne rendent pas compte des évolutions de la société. ». Elle ne constitue donc pas une ressource suffisante mais permet un regard lexicologique important. Fait intéressant, les présentes définitions établissent un lien indissociable entre l'information et sa diffusion. Nous reviendrons sur ce point par la suite lorsque nous questionnerons la relation entre l'information et la communication.

C'est l'ouvrage : Une théorie de l'information rédigée par Claude Shannon en 1948 qui a défini le concept mathématique d'information. Initialement publié par Shannon seul, il fut réédité en 1949 et préfacé par Warren Weaver sous le nom de La théorie de l'information tant son retentissement au sein de la communauté scientifique fut important. Cet ouvrage est toujours aujourd'hui considéré comme une référence absolue par la presse scientifique (il a été cité plus de 100 000 fois à raison d'en moyenne 7 000 citations par an).

L'information selon Claude Shannon

Si Warren Weaver a tenté d'étendre les découvertes de C. Shannon à une dimension plus large et notamment à celle des sciences sociales, il n'en demeure pas moins que la théorie initiale de Shannon était éminemment mathématique. La définition de l'information selon Shannon peut paraitre surprenante de prime abord et il est indispensable de la comprendre avec clarté afin d'écarter tout quiproquo. Ainsi Shannon définit l'information comme « une mesure de la liberté de choix dont on dispose lorsque l'on sélectionne un message »[5]. Pour lui, les notions d'incertitudes, d'entropie[1], de degré de liberté et d'information sont synonymes. Permettons-nous un exemple afin d'éclaircir ce que C. Shannon a voulu signifier.

[1]Le concept d'entropie a été introduit en 1865 par Rudolf Clausius, il mesure le niveau de désorganisation d'un système. Le second principe de la thermodynamique établit que « Toute transformation d'un système thermodynamique s'effectue avec augmentation de l'entropie globale incluant l'entropie du système et du milieu extérieur ». Il est ici reprit par Claude Shannon pour mesurer le degré de liberté du choix des symboles dans une langue.

Dans la mythologie grecque, Thésée en rentrant d'Athènes avait promis à son père, Egée, que son équipage hisserait une voile blanche à son retour s'il triomphait du Minotaure et une voile noire s'il trouvait la mort.

Dans cet exemple, deux « symboles » sont en jeu. Le premier symbole possible étant la voile blanche et le second, la voile noire. Un message peut être constitué de plusieurs symboles, par exemple un mot est constitué de plusieurs lettres, une partition de plusieurs notes etc. Dans notre exemple, très simple, le message n'est constitué que d'un seul symbole (voile noire ou voile blanche). L'information est alors maximale (H = 100%)[2]. Chaque symbole peut être choisi librement, sans règle liée à une structure grammaticale, à une langue etc. Shannon estime par exemple que l'information au sein de la langue anglaise (ou entropie) est proche de 50%. De la sorte comme le remarque Warren Weaver : « la moitié des lettres ou des mots (bien qu'on n'en ait pas conscience d'ordinaire) est contrôlée par la structure statistique de la langue ». Elle ne porte pas le sens. Cette part de la langue est nommée redondance. Elle s'oppose à l'information. L'union entre la redondance et l'information constitue l'ensemble du message.

L'histoire voulut que Thésée, sans doute distrait par la joie de son succès, oublia de hisser ses voiles blanches. Égée, son père, apercevant les voiles noires, de désespoir, se jeta dans la mer qui porte aujourd'hui son nom.

Nous avons vu ici que le sens du mot « information » tel que décrit par Shannon (et toujours en application dans les sciences de la télécommunication) n'est pas celui que nous recherchions. Cette parenthèse était cependant nécessaire puisque notre étude porte en partie sur ce domaine. De plus, cette définition constitue une des bases fondamentales dans les travaux de recherche traitant de l'information. Dans *Qu'est-ce que l'information ?* la chercheuse Olimpia Lombardi remarque par exemple que : "La vision de la théorie sémantique de l'information de Fred Dretske, la perspective adoptée par Peter Kosso dans son compte rendu de l'observation scientifique sur l'interaction-information, et l'approche syntaxique de Thomas Cover et Joy Thomas [...] adoptent (la théorie de Shannon) comme base formelle". [7]

2.3 Autres définitions

L'information toujours sans consensus en SHS

Dans son ouvrage *La connaissance et la circulation de l'information*, Fred Dretske, professeur à l'université de Standford, remarque que « si la théorie de l'information doit nous apprendre quelque chose sur le contenu informationnel des signaux, elle doit abandonner sa préoccupation des moyennes et nous dire

[2]Ce résultat est établi par l'utilisation de la formule : $H = -\sum_{i=0}^{n} p_i \log_(p_i)$, avec p_i la probabilité d'occurence du symbole.

quelque chose sur l'information contenue dans ces messages et ces signaux. Seuls des messages et des signaux particuliers portent un contenu. » [4]. Dretske a alors tenté d'introduire le sens dans les équations de Shannon, qui rappelons le, ne se soucient que du degré de liberté d'un mode de communication. Pour Fred Dretske le concept d'information doit être « sémantique ». C'est à dire porteur de sens. Reste à savoir quelle forme nouvelle donner à cette information porteuse de sens.

Les chercheurs font globalement le constat que le terme "information" entraîne dans son sillon un ensemble de concepts qui lui sont corollaires et notamment celui de "communication". Ce n'est donc pas un hasard si le champ universitaire qui interroge ces concepts se soit regroupé sous l'appellation de "sciences de l'information et de la communication" (SIC). Il est intéressant de remarquer en ce sens que le terme "communication" possède le même caractère interdisciplinaire que celui d'information. On le retrouve dans des champs d'application aussi variés que ceux de l'ingénierie civile à travers les voies de communication (les route, canaux, voies de chemins de fer etc.), les outils de communication (téléphone, satellite, radio), la publicité, le marketing ou encore les interactions sociales comme à travers une discussion par exemple [3]. La proximité entre les concepts d'information et de communication pousse même Marshall McLuhan, professeur à Cambridge, à considérer en 1977 que ces termes sont synonymes à travers la formule restée célèbre : « Le médium, c'est le message » [8]. Une conception plus commune considère que l'information constituerait "l'objet" à transmettre quand la communication serait le "moyen" de la transmission de cet objet. C'est la vision que semble défendre le chercheur Claude Baltz dans sa conception du « monde hypertexte » [1]. Prenant à rebours les méthodologies classiques consistant à définir l'objet de son étude pour en expliquer les comportements, Claude Baltz s'appuie sur l'état actuel des systèmes d'informations pour dresser une métaphore propre à définir l'information et la communication. Ainsi il définit la base de la structure hypertexte comme consistant en :

- Objets : réels (des personnes, des lieux, des objets physiques) mais également imaginaires (des idées, des théories)
- Relations : entre ces objets (logiques, affectives, linguistiques etc.)

De cette structure il conclut : « *toute information peut s'appréhender comme une modification de configuration dans un hypertexte* ». En cela, il semble proche des conceptions de McLuhan puisqu'une modification de la communication entre objet ou sur l'objet est assimilé à une information.

Cette vision de l'information est toutefois contestée par des chercheurs comme Eric Dacheux qui en pointe le caractère tautologique, si : « l'information c'est ce qui circule dans la communication ; la communication c'est la circulation de l'information » ni l'un, ni l'autre des deux concepts n'a été défini de façon

satisfaisante[3] [3]. D'où la nécessité selon lui d'en rétablir l'asymétrie. Proche de sujets relatifs à la philosophie, il décrit la communication comme étant ce qui permet l'identité. Si nous étions tous identiques, nous n'aurions rien à nous communiquer. Dans le même temps, la communication permet la construction de notre identité. De cela, il distingue deux type de communication : une communication dite "*égocentriste*" relative au partage de son identité ou à sa recherche et une communication dite "*altruiste*" relative à la compréhension de l'identité de l'autre ou au don de soi. Tout en soulignant la non exclusivité de ces deux types communication, il en expose quatre modalités permettant de les décrire : le temps, l'espace, la technique et le contexte. Ces quatre variables étant pour lui nécessaires et suffisantes pour décrire la communication. Si cette définition est séduisante, elle paraît toutefois camoufler un peu trop les entités responsables de cette communication : peut-on étendre cette vision de l'identité à des objets amenés à communiquer entre eux ? Leur caractère "*égocentriste*" et "*altruiste*" a-t-il alors encore un sens ? Et quant est-il du concept d'information ?

Ce rapprochement entre le concept d'information porté par l'homme et celui d'information porté par la machine, s'était déjà opéré des années en arrière par la création d'une science, multidisciplinaire elle aussi, nommée cybernétique en 1947 sous l'impulsion de Norbert Wiener. Il nous paraît indispensable de l'évoquer avant de conclure. Ne considérant pas comme Shannon que l'information est l'affaire de symboles transmis mais plutôt qu'elle précède le langage nécessaire à l'utilisation de ces symboles, la cybernétique encore une fois, opère le rapprochement entre l'information et la communication. Elle porte dès lors pour responsable de l'information le signal transmis et non le symbole utilisé pour la transmettre. Dans L'ordinateur et le cerveau, John Von Neumann, figure emblématique de la cybernétique considère : « *Ainsi, la logique et les mathématiques du système nerveux central, considérées comme des langages, (comprendre des symboles au sens de Shannon) doivent être structurellement, fondamentalement différentes des langages de notre expérience courante* » [9]. Si l'usage du terme cybernétique connaîtra un déclin après les années 70[4], il n'en demeure pas moins que ces théories seront à l'origine d'avancées majeures dans le domaine des sciences cognitives et constitueront les fondements théoriques de ce qu'on nomme aujourd'hui l'intelligence artificielle.

Nous n'irons pas plus loin dans ce questionnement de l'information de peur de nous écarter trop de notre sujet initial. Toutefois, il nous paraissait incontournable de questionner l'essence même de ce qu'est l'information pour pouvoir prétendre la quantifier ou d'analyser sa "circulation". Nous retiendrons

[3]On peut considérer ici que l'on tombe sous le coup d'une forme de "circularité logique" au sens épistémologique redéfini ainsi par le philosophe allemand Hans Albert (1921-1991).

[4]Consulter Google Ngram Viewer, pour les recherches : cybernetic, cybernetics, cybernétique ...

donc qu'en dépit de travaux brillants et fouillés, la recherche universitaire n'est pas encore parvenue à fournir une définition de l'information qui satisfasse à la fois l'ensemble des sciences dans lesquelles le terme trouve un sens et la somme des contextes empiriques dans lesquels il est employé. En particulier, la question du sens qui semble inhérente au concept d'information semble difficile à caractériser. C'est pourquoi deux visions s'opposent, l'une que l'on peut qualifier de "symbolique", celle de Shannon, qui ne se soucie pas du sens, l'autre que l'on pourrait qualifier de "sémantique" mais qui reste vacante. En dépit de cela, et malgré les lacunes mises en avant de ces modèles, nous nous rapprocherons dans notre étude de la définition de l'information donné par Baltz, qui se trouve particulièrement opportune dans notre cas puisqu'elle est issue de l'espace que nous observons, en l'occurrence le web. Nous assimilerons alors au terme d'information ce que Baltz nomme "objet", considérant à sa différence que l'ajout d'un objet identique au sein d'un réseau hypertexte n'apporte pas d'information. La mesure de similarité entre ces "objets" devenus information sera décrite par la suite dans la partie "Nature des indicateurs", prenant pour base théorique la mesure de "l'entropie" de Claude Shannon.

Dans la partie suivante nous allons proposer un ensemble d'indicateurs permettant de quantifier "symboliquement" la similarité entre plusieurs textes. Par la suite, ces indicateurs seront utilisés pour évaluer la similarité entre des dépêches et des articles de presse afin de mesurer la circulation des informations qu'ils véhiculent.

3 Nature des indicateurs

3.1 Pourquoi développer un nouvel outil ?

Nous avons en réalité commencé par utiliser des outils SEO de détection de plagiat disponibles gratuitement en ligne pour effectuer notre étude. Cependant cette utilisation c'est vite révélée limitée. En effet la majorité des sites proposant ce genre de services ne sont pas en Open Source, c'est à dire qu'il n'est pas possible de lire le code qui génère le résultat que l'on obtient. Par conséquent, nous sommes contraints d'accorder notre confiance à la qualité des indicateurs proposés. Or, l'identité des créateurs de ces sites (lorsqu'elle est présentée) ne donne généralement pas davantage de légitimité aux résultats. De plus, la comparaison issue de mêmes textes donne des résultats différents en fonction des comparateurs, ce qui renforce notre méfiance quant à leur précision.

Pour toutes ces raisons nous avons fait le choix de développer notre propre outil. Il est hébergé sur la plateforme participative GitHub, ce qui pourrait permettre à n'importe quel développeur d'apporter sa contribution au projet. De plus, le traitement étant effectué en Javascript, n'importe quel utilisateur peut consulter les opérations effectuées en temps réel sur son terminal. Le code est donc accessible et consultable par tous. Le mode d'emploi de l'application

est décrit à travers un feuillet de documentation présent en annexe. Nous allons à présent expliquer, avec le plus de clarté possible, le fonctionnement des indicateurs de similarité utilisés.

3.2 Indice de Jaccard

L'indice de Jaccard tient son nom du botaniste suisse Paul Jaccard (1868 - 1944), spécialiste en physiologie végétale, qui en est à l'origine. Il est utilisé en statistique pour évaluer la diversité entre deux échantillons. Il est nommé "coefficient de communauté" dans sa publication d'origine[5]. Mathématiquement il est décrit comme le rapport entre le cardinal de l'intersection et le cardinal de l'union des ensembles étudiés. Le cardinal correspond au nombre d'éléments présents dans un ensemble. Plus simplement, on peut le décrire comme la division du nombre d'éléments communs entre les échantillons sur le nombre d'éléments total. Plus simplement, on peut le décrire comme la division du nombre d'éléments communs entre les échantillons sur le nombre d'éléments total.

Soient deux échantillons A et B, l'indice de Jaccard s'exprime par :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Pour plus de deux échantillons, a on a logiquement :

$$J(S_1, S_2, ..., S_n) = \frac{|S_1 \cap S_2 \cap ...S_n|}{|S_1 \cup S_2 \cup ...S_n|}$$

Pour l'analyse de texte, il est possible d'appliquer cet indice aux lettres et aux mots. Par exemple la comparaison entre les mots « bulbe » et « bulle », pour une analyse par les lettres donnerait :

Ensemble I des lettres communes : b,u,l,e
Ensemble U total des lettres utilisés : b,u,l,e

Le rapport de I sur U donne 1, soit un "coefficient de communauté" (pour reprendre l'expression du botaniste) de 100%. On peut également faire jouer la redondance des lettres, ce qui est en général plus pertinent :

Ensemble I des lettres communes : b,b,u,u,l,l,e,e
Ensemble U total des lettres utilisées : b,u,l,l,e,b,u,l,b,e
Indice de jaccard (le produit de 8 par 10) soit : 80%

Lorsqu'on prend en compte la redondance, on double mécaniquement la taille de l'intersection pour conserver les proportions de l'échantillon. Pour étudier la

[5] Paul Jaccard, « Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines »

similarité entre des textes on peut appliquer ce procédé aux mots plutôt qu'au lettres. Soient les phrases suivantes :

A : « J'ai deux chiens et un canard. »
B : « J'ai deux chats et un canard. »

On a :
Ensemble des mots communs : j', ai, deux, et, un, canard
Ensemble des mots total : j', ai, deux, chiens, chats, et, un, canard.
Indice de Jaccard (le produit de 6 par 8) : 75%

On peut de même prendre en compte la redondance au cas où un mot se répèterait (ce qui est très probable dans un texte). C'est d'ailleurs la solution que nous avons adoptée dans notre outil : un indice de Jaccard par les mots avec une prise en compte de la redondance.

3.3 Indice de Levenstein

L'indice de Levenstein s'appuie sur une « distance », nommée « distance de Levenstein ». Le concept de distance, emprunté à la géométrie a été étendu à d'autres domaines des mathématiques comme l'analyse ou la théorie des nombres. La distance de Levenshtein est initialement égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une phrase à une autre.

Pour reprendre notre exemple précédent, la distance entre les mots « bulle » et « bulbe » est de un. Il faut effectuer une opération pour passer de l'un à l'autre. En l'occurrence, le remplacement de la lettre l par la lettre b, et vice versa.

Encore une fois cet indicateur est applicable aux mots, la distance entre la phrase « J'ai deux chiens et un canard. » et « J'ai deux chats et un canard. » est également de un. Il faut remplacer le mot "chiens" par le mot "chat" pour passer d'une phrase à l'autre, ce qui constitue une seule opération.

Pour passer d'une distance à un indice en pourcentage on soustrait la distance au cardinal du texte le plus long, puis on le divise par ce même cardinal.

3.4 Indice Similar Text

La fonction "similar text" est présente nativement dans certain langage de programmation comme PHP (Hypertext Preprocessor, initialement Personal Home Page Tools). PHP est le langage dans lequel a été développé, entre autres, Facebook et Wikipédia. L'indice "similar text" fonctionne avec les caractères. Elle

commence par chercher la sous-chaîne[6] commune la plus longue puis répète l'opération pour les préfixes et les suffixes de cette chaîne de manière récursive. La somme des longueurs de toutes les sous-chaînes communes est enregistrée. Elle est ensuite divisée par la moyenne des longueurs des chaînes données puis multipliées par 100 pour obtenir un pourcentage. Le fonctionnement détaillé de cette fonction est décrit dans l'ouvrage *Programming Classics: Implementing the World's Best Algorithms* de Ian Oliver (1994) [10].

3.5 Indice de Jaro-Winkler

L'indice de Jaro-Winkler est normalisé de façon à avoir une mesure entre 0 et 1. Zéro représente l'absence de similarité et 1, l'égalité des chaînes comparées. Il suffit donc de le multiplier par cent pour obtenir un pourcentage final. C'est l'indice le plus complexe que nous avons utilisé.

L'indice de Jaro-Winkler est une amélioration de l'indice de Jaro (décrit par Matthew A. Jaro en 1985 [6]) établi par le statisticien William E. Winkler en 1999 [14]. Il a été créé pour analyser la similarité des noms et des prénoms dans le recensement américain.

Indice de Jaro
L'indice de Jaro est défini comme suit :

$$sim_j = \begin{cases} 0, & \text{si m} = 0 \\ \frac{1}{3}(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}), & \text{sinon} \end{cases}$$

sim_j : désigne la valeur de l'indice de Jaro, celui-ci vaut 0 si il n'y a aucun caractère en commun entre les deux chaînes. De la même façon que pour les autres indices, on pourra utiliser des mots plutôt que des caractères. Si au moins un caractère (ou mot) est présent dans les deux chaînes de caractères, on applique la formule en bas de l'accolade.

$|s_1|$: est la taille de la première chaîne de caractères
$|s_2|$: la taille de la seconde chaîne de caractères
m : m est le nombre de caractères (ou mots) correspondant entre les deux chaînes (m pour *match* en anglais), il est défini de façon particulière comme nous allons le voir.
t : le nombre de transpositions, nous allons également le définir par la suite.

Pour mieux comprendre la démarche nous allons prendre un exemple. Si nous souhaitions comparer les mots "Winkler" et "Welfare":

[6] En informatique on nomme "chaîne de caractère" une variable constituée d'un ensemble caractères (comme une phrase par exemple), une sous-chaîne est donc un sous ensemble d'une chaîne principale.

Il convient tout d'abord de calculer un premier seuil que l'on notera e.

$$e = \lfloor \frac{max(|s_1|, |s_2|)}{2} \rfloor - 1$$

On commence par chercher la taille du mot le plus long. En l'occurrence les deux mots ont la même taille (7 caractères) donc $max(|s1|, |s2|) = 7$. Ensuite, on divise le résultat par 2, auquel on applique la fonction partie entière inférieure (aussi appelé *floor* en anglais, plancher). Elle a pour but d'obtenir un nombre entier au cas où la division précédente ait donné un nombre à virgule. C'est ici le cas puisque la division de 7 par 2 donne 3,5. L'application de la fonction partie inférieure donnera 3. Ceci fait on retranche 1.

Finalement,
$e = \lfloor \frac{7}{2} \rfloor - 1 = 2$

Ensuite on superpose les deux chaînes de caractères comme ceci :

W	I	N	K	L	E	R
W	E	L	F	A	R	E

On parcourt les caractères de la première chaîne (W puis I puis N etc...). On considère que le caractère est similaire si le groupe constitué : du caractère en dessous, des e caractère(s) à sa gauche et des e caractère(s) à sa droite contient le caractère que l'on est en train de parcourir.

Par exemple pour la lettre L :

W	I	N	K	**L**	E	R
W	E	**L**	**F**	**A**	**R**	**E**

La lettre "L" est présente dans le groupe des lettres en gras de la deuxième colonne, donc le caractère est considéré comme similaire.

La somme des caractères similaires est noté m.
Ici $m = 4$, les caractères évalués comme similaires sont : W, L, E, R

Pour obtenir le nombre de transposition on superpose les caractères similaires dans l'ordre de leur apparition dans leur chaîne.
Dans l'exemple on obtient :

Le nombre de transposition correspond au nombre de caractères n'étant pas identiques aux caractères en dessous d'eux divisé par deux. Ici c'est le cas pour E et R. Soit $t = \frac{2}{2} = 1$

W	L	E	R
W	L	R	E
0	0	+1	+1

Il ne nous reste plus qu'à appliquer la formule du début. On obtient un indice de Jaro : $sim_j = 63,0\%$

Adaptation de Winkler L'adaptation proposé par Winkler ne concerne que les cas dépassant 0,7[7] sur l'indice de Jaro. Le cas précédent ne serait par exemple pas réévalué (0,63).

L'adaptation de Winkler est définie comme suit :

$$sim_w = sim_j + lp(1 - sim_j)$$

Où :
sim_j : est l'indice de Jaro obtenu suivant la méthode précédemment décrite.
p : est un facteur d'échelle, Winkler préconise d'utiliser 0.1
l : est la longueur du préfixe commun entre les deux chaîne. Pour les mots sa valeur maximale est de 4

Par exemple pour les noms suivant :
A : « John »
B : « Johnson »

L'indice de Jaro donne $sim_j = 0.79$
La réévaluation de Winkler donne :
$sim_w = 0.79 + 0.1 \times 2 \times (1.0 - .790) = 0.832$

L'adaptation de Winkler semble adaptée pour des chaînes de petites taille. Dans les paramètres utilisés précédemment, elle pourrait sembler moins pertinente pour les chaînes de grandes tailles (comme des textes). Cependant, c'est bien l'indice de Jaro-Winkler qu'il est coutume d'utiliser dans la détection du plagiat[8]. Les chercheurs B. Leonardo et S. Hansun de l'Université de la Multimedia Nusantara University remarquent dans leur article « Détection du plagiat dans les documents textuels à l'aide des algorithmes de distance Rabin-Karp et Jaro-Winkler » : « Sur la base d'autres recherches qui ont été faites, nous pouvons conclure que les algorithmes RabinKarp et Jaro-Winkler Distance peuvent être utilisés pour détecter le plagiat. » [2]. Nous conserverons donc cet indicateur.

[7] Valeur de seuil décidée par Winkler

[8] Brinardi Leonardo, Seng Hansun dans *Text Documents Plagiarism Detection using Rabin-Karp and Jaro-Winkler Distance Algorithms* : « Based on other researches that had been done; we have a basic conclusion that RabinKarp and Jaro-Winkler Distance algorithms can be used to detect plagiarism. »

3.6 Indice % Tot. substrings

Nous avons également proposé un nouvel indice basé sur l'algorithme de recherche de la plus longue sous-chaîne commune. Grâce à une valeur seuil qu'il est possible de changer directement à travers l'interface (fixé à 30 par défaut) l'algorithme détecte les passages communs dont le nombre de caractères dépassent cette valeur seuil.

Exemple :
Phrase 1 : « La Cigale, ayant chanté tout l'été, se trouva fort dépourvue quand la bise fut venue ». (84 caractères)
Phrase 2 : « La Cigale, ayant chanté tout l'été, se trouva fort déconvenue quand la bise fut venue ». (85 caractères)

L'algorithme cherche dans un premier temps la plus longue séquence commune dans les deux phrase. Il s'agit de « La cigale, ayant chanté tout l'été, se trouva fort dé » (53 caractères). La sous-chaîne fait plus que 30 caractères, la valeur seuil, donc elle est comptabilisée comme plagiée. L'algorithme cherche alors la seconde séquence commune la plus longue. Il trouve « ue quand la bise fut venue ». Cette sous-chaîne ne fait que 26 caractère, elle n'est pas prise en compte et l'algorithme s'arrête de chercher. L'indice final donne le rapport entre le nombre total de caractères présents dans les sous-chaînes communes sur le nombre total de caractères de l'échantillon le plus court.

Ici l'indice vaut : $(\frac{53}{84}) \times 100 = 63,1\%$

Détermination de la valeur seuil par défaut Pour déterminer la valeur seuil par défaut nous avons utilisé un échantillon témoin de 17 textes parfaitement hétérogènes que nous avons comparés un à un à un texte de référence (voir étude témoin). La moyenne du nombre de caractères de la plus longue séquence commune entre ces textes est de 15 et la médiane de 16. La valeur maximale est de 19. Nous avons donc estimé que choisir un seuil de détection égal au double de cette moyenne (30) constituait une marge suffisante pour considérer un passage comme commun entre deux textes.

3.7 Conclusion sur les indicateurs

Comme le relève A. Hery Purba et Z. Situmorang dans leur étude "Analyse comparative de l'algorithme Rabin-Karp et de la distance de Levenshtein dans le calcul de la similitude d'un texte", il est important de mettre en application plusieurs indicateurs dans l'analyse de plagiat afin d'augmenter ses chances de détection.

Les choix d'indicateurs que nous avons fait figurent parmi les plus communément employés par les algorithmes de détection de plagiat. Cependant d'autres indicateurs auraient également pu être utilisés, en voici une liste non exhaustive

:

- Algorithme de Rabin-Karp
- Similarité cosinus
- Indice de Tanimoto
- Méthode de Pemanfaatan
- Score de Fuzzy
- Algorithme phonétique Soundex

Comme nous pouvons le constater, les indicateurs présentés ne prennent pas en compte des similarités de sens et pourrait être trompés par une réécriture qui utiliserait des synonymes ou des paraphrases. Les difficultés liées à la mise en place d'une détection de la synonymie sont discutées dans la partie *Limites générales de l'étude*.

4 Études

4.1 Étude Témoin

Avant de pouvoir rendre de compte des résultats de l'exploitation des indices que nous venons de décrire, il est impératif de tenter d'évaluer leur performances pour le sujet qui nous intéresse. Pour ce faire, nous avons commencé par réaliser une étude témoin comparant 17 textes n'ayant vraisemblablement aucun rapport entre eux. Ce corpus hétérogène est composé d'articles de presse en rapport avec des sujets internationaux, d'interviews, de tests de produits commerciaux en passant par des rédactions plus historiques telles que le célèbre article « J'accuse... ! » d'Émile Zola publié en 1898 dans le journal l'Aurore. Chacun des textes est comparé à un texte de référence que nous avons choisi arbitrairement et qui se trouve être un article d'Europe 1 publié le 13 novembre 2016 sur la chasse aux truffes. Dans l'idéal, nous attendrions de nos indicateurs qu'ils renvoient la valeur de 0% pour des textes semblant traités de sujet différents et une valeur de 100% pour des textes identiques. En réalité, la probabilité que ceux-ci établissent une similarité de 0% entre deux textes, même extrêmement différents, est très basse pour des raisons que nous allons expliquées. En revanche, la comparaison de deux textes parfaitement identiques restituera bien systématiquement la valeur de 100% pour l'ensemble des indicateurs choisis.

Performance des indicateurs
Indice de Jaccard

La mesure moyenne de l'indice de Jaccard sur l'échantillon témoin semble relativement élevée, elle est de : 9,39%. On le rappelle, l'indice de Jaccard donne le ratio des mots communs sur le total des mots. L'analyse de cette similarité

sur des textes n'ayant aucun rapport entre eux met en exergue un phénomène particulièrement intéressant : la redondance au sens de Shannon. En d'autres termes, les contraintes structurelles liées à la langue française. Voici par exemple la liste des mots conjoints entre le premier et le deuxième article de notre échantillon témoin :

la	de	la	en
plus	sur	ce	alors
nous	un	dans	la
qui	est	le	de
vous	les	on	est
du	et	nous	sur
du	une	de	avec
avec	c'	est	l'
c'	et	la	en
est	de	et	vous
et	ce	que	de
quand	à	les	de
plus	elle	une	à
et	vous	faire	un
de	à	et	à
de	un	de	plus
?	tout	simplement	le
une	très	car	le
font	les	j'	ai
une	de	asssez	temps
les	les	les	sentir
oui	ses	son	a
fait	que	?	dont
du	pour	le	la
et	en	que	l'
les	les	sans	y

Figure 1: Mots similaires entre les deux premiers articles de l'échantillon.

l'	d'	en	un
:	que	la	en
et	plus	dans	la
avec	de	la	le
le	de	la	et
trois	qui	est	un
des	des	et	qui
chaque	pour	les	les
de	se	pour	un
vous	par	l'	et
au	par	les	de
maison	de	et	du
côté	de	et	le
voyage	le	le	de
une	journée	à	de
une	qu'	est-ce	que
fait	de	les	il
donne	aussi	des	petits
une	belle	les	de
d'	les	les	de
du	du	ne	pas
a	quand	qu'	n'
produit	du	peu	vous
de	et	et	les
le	le	son	ses

Figure 2: Mots similaires entre les deux premiers articles de l'échantillon.

La présence nécessaire dans chaque texte quel qu'il soit de : déterminants, pronoms, prépositions etc. entraîne nécessairement une similarité que l'on pourrait qualifier de "résiduelle"[9] entre ce texte et n'importe quel autre pourvu qu'ils

[9]Ce terme n'est pas très heureux mais permet de bien comprendre la situation. En traitement du signal on parle de "bruit".

soient tous deux rédigés en français. La probabilité que cet indice tombe à 0%, ce qui signifierait que ces deux textes n'ont pas de mots en commun, est donc hautement improbable. Pour l'interpréter correctement il serait donc nécessaire d'établir la proportion moyenne de ce "résidu" au sein de la langue française. Au dessus de ce niveau résiduel viendrait s'ajouter une zone d'incertitude dans laquelle il serait difficile de déterminer si les textes présentent des similarités ou s'il s'agit d'un surgissement inhabituel du résidu. Cette zone d'incertitude serait elle même majorée par un seuil à partir duquel on estimerait que la probabilité de similarité entre les textes devienne significative. À ce moment seulement l'interprétation du pourcentage de Jaccard pourrait être considérée comme une mesure de l'analogie entre les textes.

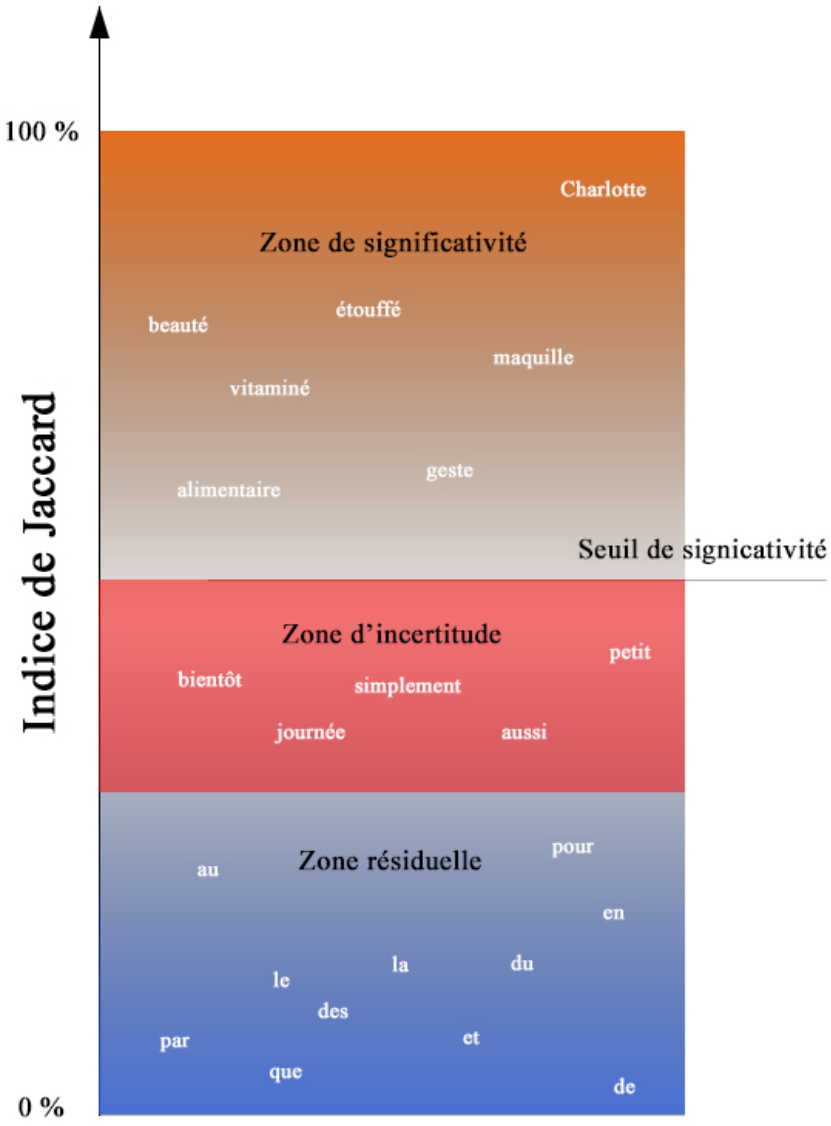


Figure 3: Représentation du seuil de significativité pour l'indice de Jaccard.

Seuil de significativité proposé pour l'indice de Jaccard

Pour établir ce seuil nous contenter de 17 articles aurait été insuffisant. Dans un soucis de temps et d'efficacité nous avons donc eu recours au site de génération de texte : enneagon.org. Ce site génère des textes de façon aléatoire en suivant la répartition statistique d'un référentiel textuel en langue française, selon le principe des chaînes de Markov. Plus simplement, il produit aléatoirement des textes dénuées de sens mais dont la structure des phrases est sémantiquement correcte. Nous avons généré 100 textes de 50 phrases chacun auxquels nous avons appliqué notre indice de Jaccard pour mesurer sa valeur moyenne au sein de la langue française. Voici les résultats :

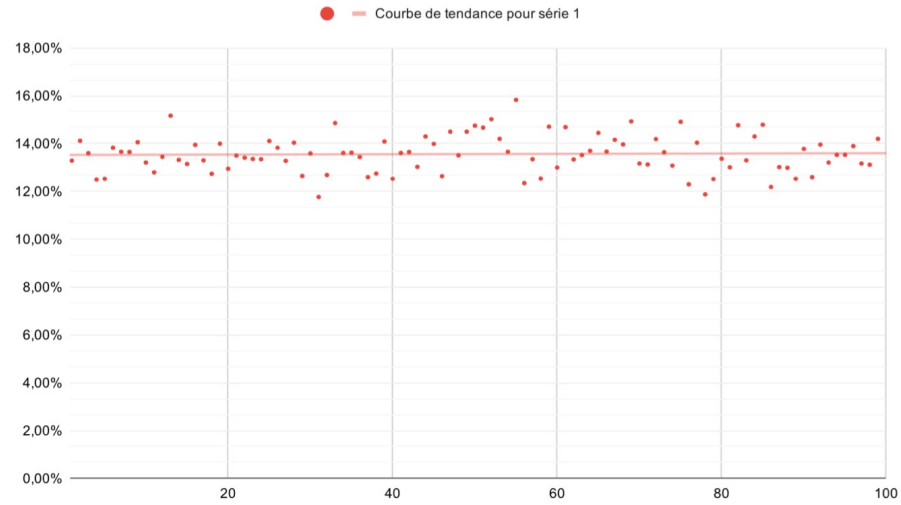


Figure 4: Répartition de l'indice de Jaccard sur un échantillon de 100 textes hétérogènes.

On obtient une moyenne de 13,56% avec un écart-type de 0,78%. La valeur minimale est de 11,7% et la valeur maximale de 15,83%. On peut donc considérer que notre "zone résiduelle" se situe en dessous de la valeur minimale, que la zone d'incertitude est centrée sur notre courbe de tendance à 13,56% et que le seuil de significativité se situe au abord de la valeur maximale. Afin de nous prémunir le plus possible des effets de bords, on choisira le seuil de significativité comme étant la valeur maximale de cet échantillon auquel on additionnera l'écart-type soit : 16,61%. Sera donc considéré comme similaires des textes dont l'indice de Jaccard dépasse ce seuil. En dessous, nous ne pourrons pas nous prononcer sur la ressemblance de ces textes.

Seuil de significativité proposé pour l'indice de Levenshtein

L'indice de Levenshtein est confronté au même problème que l'indice de Jaccard. Comptant les opérations nécessaires jusqu'à ce que les deux textes soit similaires, il est également affecté par le fait que les textes français détiennent une base indéfectible de mots en communs. Nous avons procédé de la même façon pour déterminer le seuil à partir duquel cet indice semble significatif. En voici les résultats :

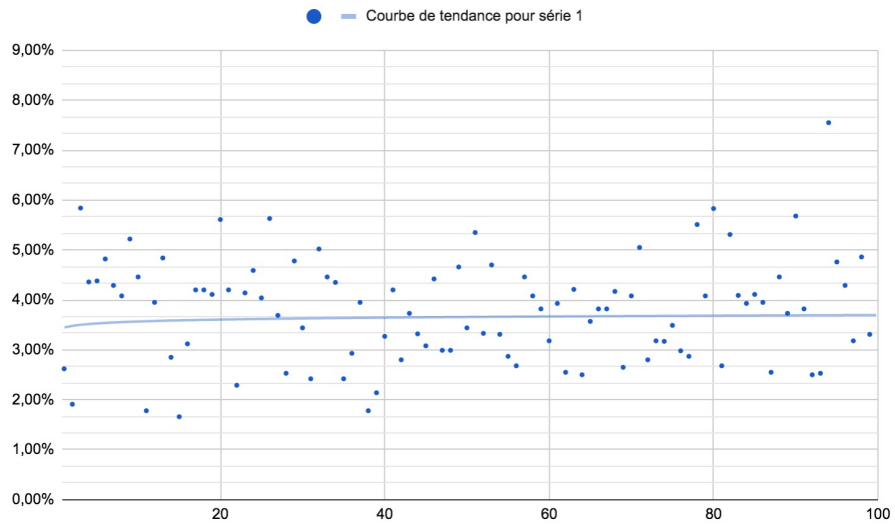


Figure 5: Répartition de l'indice de Levenshtein sur un échantillon de 100 textes hétérogènes.

La zone d'incertitude pour l'indice de Levenshtein est plus basse que celle de l'indice de Jaccard. La moyenne est de : 3,79%, l'écart-type de 1,05%, la valeur maximale de 7,55% et la valeur minimale de 1,66%. Le seuil de significativité, sera calculé de la même façon que pour l'indice de Jaccard, soit la valeur maximale plus l'écart-type, ce qui nous donne : 8,61%. Encore une fois, en dessous de ce seuil nous ne pourrons pas nous prononcer sur la ressemblance des textes.

Seuil de significativité pour les autres indices

Pour l'indice Similar Text, en suivant les mêmes opérations, on obtient un seuil de significativité à : 22,1%. En revanche, pour les deux autres indices (Jaro-Winkler et %Tot. Substring) cette méthodologie n'aurait pas de sens car leur fonctionnement n'est pas altéré par la présence de ce "résidu linguistique". On observe que l'indice %Tot. Substring est particulièrement efficace pour détecter le plagiat. Il obtient en moyenne une valeur de 0,46% pour les textes hétérogènes

et une valeur de 100% pour des textes similaires. C'est donc l'indice auquel nous accorderons le plus de crédit. En revanche, l'étude témoin semble démontrer que l'indice de Jaro-Winkler est, en l'état, inexploitable car il restitue des résultats très élevés et dispersés même pour des textes hétérogènes. Deux explications possibles à cela. D'une part, cet indice est davantage conçu pour l'évaluation de la similarité entre les mots et non entre des textes. D'autres part, les variables d'ajustements que nous avons testées (dont celles recommandées par Winkler) n'étaient peut être pas les plus adaptées en ce cas précis. Peut-être que d'autres variables auraient fourni des résultats plus satisfaisants.

En résumé, les contraintes grammaticales et syntaxiques de la langue française entraînent des similarités entre l'ensemble des textes de cette langue quels que soient les sujets qu'ils abordent. Il existe donc des seuils pour les indicateurs que nous avons utilisés en dessous desquels il est impossible d'établir une similarité avérée entre des textes. Les seuils retenus sont :

- Indice de Jaccard : 16,61%
- Indice de Levenshtein : 8,61%
- Indice Similar Text : 14,61%
- Indice de Jaro-Winkler : Indéfini
- Indice % Tot. Substring : 1%

Limites générales de l'étude

Au delà de ses limites intrinsèques à l'usage de chaque indicateur que l'on pourrait qualifier de "locales", il existe des contraintes plus "globales" qui pèsent aussi sur notre champ d'étude. Tout d'abord, nous avons tenté de comparer l'ensemble des articles issus de la "presse traditionnelle en ligne" (cf. Typologie, en cinquième partie) sur un événement donné. La recherche de ces articles, constitués en une sorte de revue de presse, s'est faite par l'intermédiaire de moteurs de recherche. Nous avons pour cela utilisé le moteur lié à la base de données d'Europresse et en second recours le moteur allemand Ecosia. Cet assemblage d'articles n'est peut-être pas exhaustif et ce pour une raison simple : les moteurs de recherche utilisent eux aussi des indices de similarité pour restituer les articles qu'ils jugent les plus pertinents. Pour prendre un exemple, dans notre troisième étude, nous avons étudié les articles relatifs au *bank run* grec de 2012. Pour cela, nous avons entré dans les moteurs un ensemble de mots-clefs tels que : *bank run*, Grèce, 2012 etc. L'ensemble des articles portés à notre connaissance sur le sujet est donc en amont conditionné par des algorithmes de recherche qui eux - mêmes sont susceptibles d'éluder des articles traitant du sujet souhaité. Usant de technologies de détection de similarité pour retourner un résultat, il est normal que nos objets de recherche (les articles) présentent dès le départ des similarités entre eux. Les points de rencontre entre ces objets se trouvant

autour des mots-clés choisis et de la date de l'événement dont il est question. En d'autres termes, il est probable qu'un article radicalement différent dans son traitement de l'événement, par l'usage de mot différents, d'un délai plus long pour traiter le sujet ou encore par l'évocation principale d'un fait corollaire à l'événement que nous recherchions (choix de l'angle), puisse avoir été oublié. Ceci constitue une première limite qu'il paraît difficile de dépasser. Pour affiner la recherche, il serait nécessaire d'analyser plus en profondeur les algorithmes de recherche des moteurs utilisés (ce dont nous n'avons ni la compétence, ni la permission) ou bien d'entamer des travaux inconsidérés de revue manuels de la presse aux abord des évènements étudiés. Cette dernière option est en pratique impossible à mettre en place et ne garantit pas moins l'absence d'un oubli : humain cette fois-ci. C'est donc une première limite globale de notre étude que l'on se doit de garder modestement à l'esprit.

Une seconde limite, très importante elle aussi, est la question du sens. Nous rapprochant dans cette analyse davantage de la conception "symbolique" de l'information au sens de Shannon, nous ne faisons pas état du "sens" équivalent que pourrait avoir deux textes de formes différentes. Une définition formelle du "sens", c'est à dire l'établissement d'un ensemble de règles propres à établir que deux messages possèdent la même signification est extrêmement difficile à mettre en place. Nous avons déjà abordé cette question à travers notre première partie lorsque nous nous sommes interrogés sur ce qu'est l'information. Nous n'y reviendrons pas. En dépit de cela, et même si l'établissement d'indices "sémantiques" irréprochables semblent illusoires, certaines pistes semblent pouvoir contribuer à l'amélioration de nos indices "symboliques". La synonymie notamment. On peut considérer par exemple que l'affirmation "cette surface est bleue" possède la même signification que "ce plan a la couleur de l'eau". Pourtant, il n'existe aucun mot en commun entre ces deux phrases. Elles passeront donc largement sous la détection des radars de nos indices de similarité. D'aucuns pourraient considérer que l'usage d'une table de hachage[10] permettrait de régler aisément le problème. Cependant cela serait considérer une synonymie absolument binaire entre les mots. Deux mots serait alors synonymes ou non. Or chacun sait que certains mots sont plus proches que d'autres, en fonction du contexte notamment et de ce que l'on souhaite exprimer. Par exemple "de fines attaches", nous paraît être plus proche de: "de minces attaches" que: "de maigres attaches". Ces trois adjectifs sont pourtant synonymes dans le Larousse[11]. D'où la nécessité apparente d'établir une métrique de la synonymie. Cette métrique se confronte immédiatement à la difficulté de la subjectivité individuelle de la "valeur" des mots et à son inscription dans une langue non pas figée mais vivante où les définitions changent et évoluent. La détection de métaphores équivalentes est également un problème d'apparence insoluble pour les bases de

[10] Une table de hachage en informatique est une structure de donnée reposant sur le modèle clé-valeur. Par exemple, le mot "sublime" pourrait être associé au mot "superbe" signifiant par leur association au sein de cette table qu'ils sont synonymes. Les tables de hachage sont nommées "dictionnaire" en langage python, Hash en ruby.

[11] Larousse.fr : onglet synonyme pour le mot "mince" (8 juin 2020)

données. Dans notre exemple précédent nous avons choisi “la couleur de l’eau” pour exprimer la couleur bleu, le stockage en base de l’ensemble des métaphores associées au bleu est par exemple irréalisable puisqu’il reviendrait à faire la liste exhaustive de tous les objets bleus. Liste inépuisable, mouvante et changeante dont il est impossible de tenir les comptes. La paraphrase se confronte au même problème.

Ces considérations, bien qu’elles soit responsables de notre restriction à une analyse “symbolique” et non “sémantique”, sont tout de mêmes rassurantes puisque leur abolition aurait signifié une dépossession de l’homme dans son usage exclusif de la poésie ou de la littérature. Nous considérerons donc qu’une reformulation même grossière constitue un apport en information. De ce fait, ne sera pas considéré comme similaire des faits identiques rapportés par des mots différents. À noter toutefois qu’un simple changement dans la structure d’une phrase sera lui, pris en considération (ex: “la rue était vide et noire” / “noire et vide était la rue”).

L’absence de la prise en compte du sens n’est finalement pas si décisive pour deux raisons. D’une part, comme nous le verrons dans nos études, le phénomène de réécriture ne se constate que très peu. Plusieurs raisons à cela : la faible valeur ajoutée d’un texte reformulé, son coût prohibitif par rapport à un texte reproduit à l’identique, l’ingratitude de la tâche de reformulation etc. D’autre part, comme le souligne Adeline Wrona, professeure au CELSA en sciences de l’information et de la communication dans *Circulation circulaire de l’information - Lexique des SIC*, le journalisme depuis le XIXème siècle a toujours eu un caractère métadiscursif, de rapporteur de faits : « c’est un discours, dans une société de discours, fabriqué par des professionnels du discours ». Il en découle que reformuler des faits est partie intégrante du métier de journaliste et qu’il n’y a pas nécessairement lieu de dénoncer une telle pratique.

4.2 Étude 1 - Premier décès d’un médecin du coronavirus en France

Les résultats de chaque étude sont présents en annexe, nous vous invitons à les consulter afin de mieux comprendre l’objet des commentaires qui vont suivre. Ces résultats sont également disponibles sur le site de l’application. Il peut être plus opportun de les consulter sur le site car il permet l’accès aux articles étudiés via les liens qui y sont associés. En tout, 132 articles issus de la presse en ligne auront été comparés au sein de ce travail.

Notre première étude concerne le décès du premier médecin français des suites du Covid-19. La dépêche AFP est parue à 9h51, heure de Paris. C’est une information “nationale”, elle concerne des faits ayant eu lieu sur le territoire français. Le médecin semble avoir été infecté dans le département de l’Oise et est décédé à Lille après avoir été transféré pour recevoir des soins. L’information

a été beaucoup reprise par les organes de presse français.

La plupart des journaux à l'exception du journal Le Monde ont traité l'information dans la journée. Le lendemain, date de parution de l'article du Monde, un deuxième médecin décédait du même virus. La plupart des journaux ont alors refait un sujet. Le temps de réaction a ici une importance cruciale dans le taux de similarité entre la dépêche et les articles. Le Monde par exemple a un taux de similarité tellement bas (1,70 %) qu'on peut estimer que la dépêche et l'article n'ont rien en commun.

Cet événement est survenu dans des conditions particulières de l'exercice de la profession. Suite à l'évolution de la pandémie de Covid-19, le gouvernement français a mis en place des mesures de confinement à partir du 17 mars. Ces mesures ont très probablement restreint les journalistes dans leur capacité à pouvoir interroger des témoins ou à mener une enquête. C'est ce qui fait également que le taux de similarité de l'article du journal Le Monde est très bas. L'article présente un témoignage original d'un délégué hospitalier de Compiègne et de Corinne Delys, secrétaire générale de la CGT à l'hôpital de Creil. Encore une fois c'est probablement le temps de rédaction de cet article qui a permis un apport en information plus conséquent.

Remarquons également que le journal Gala a un taux de similarité très bas (1,90%). Cela s'explique par l'angle journalistique adopté par le journal. Si la dépêche et de nombreux journaux reprennent les propos d'Olivier Véran, ministre de la santé, Gala s'attarde davantage sur des témoignages de la famille du médecin décédé publiés sur les réseaux sociaux. Ceux-ci n'apparaissent pas dans la dépêche. Les journaux qui ont traité à la fois les témoignages de la famille et les déclaration d'Olivier Véran ont généralement un taux de similarité qui avoisine les 40% (20 Minutes - 44,8%, Le Point 41,60%, La Voix du Nord 43,90%).

Le choix du recours à la périphrase a également une incidence sur les résultats. En particulier, la citation suivante : « Puisque vous m'en donnez l'occasion, je voudrais vraiment m'associer à la douleur, à la peine de la famille, à la douleur et à la peine de l'ensemble des soignants » prononcée par Olivier Véran sur RTL est présente sous forme de périphrase dans la dépêche de l'AFP : « M. Véran, qui s'est "associé à la douleur de la famille" ». Sur les 20 articles analysés, 12 ont repris telle quelle la périphrase de l'AFP, 4 l'ont citée comme suit : « Je tiens à m'associer à la douleur et à la peine des soignants. Ils payent un très lourd tribut ». Cette dernière diffère légèrement de la citation d'origine. Quant à la seconde partie concernant le "tribut" que payent les soignants, elle vient plus tard dans le discours et en ces termes : « C'est un très lourd tribut qui est payé par la grande famille des médecins aujourd'hui dans notre pays ». Le sens n'est pas déformé dans la citation que font ces quatre journaux. En revanche le fait qu'ils utilisent une même citation qui ne soit pas exactement celle du discours d'origine porte à croire que l'information a été reprise et non directement extraite de son contexte d'origine.

Six journaux sur les vingt ont un taux de similarité particulièrement élevé (plus de 93%). Corse Matin atteint même un taux de similarité de 100%, l'article ne contient que des phrases qui sont présentes dans la dépêche de l'AFP. Pour ces six journaux l'information a été extrêmement peu retravaillée et on peut considérer qu'aucun supplément d'information n'a été apporté. Souvent même, la dépêche contient davantage d'informations que les articles.

L'écart-type du taux de similarité relatif est très important au sein de cette étude : 39,2%. À titre indicatif, la valeur minimum est de 1,20% (France 3) et la valeur maximale atteint 100% pour Corse Matin. Cela signifie que bien que la majorité des journaux ont fait le choix de traiter cette information, certains ont apporté un contenu supplémentaire ou une réécriture des faits décrits par l'AFP et d'autres non.

En résumé, les facteurs qui poussent les taux de similarité à la hausse sont :

- Une reprise parfois totale et sans apport de la dépêche
- Le contexte sanitaire rendant plus difficile l'exercice du journalisme

Les facteurs qui poussent les taux de similarité à la baisse sont :

- L'utilisation des témoignages de la famille externe à la dépêche
- L'interview de témoins (journal Le Monde uniquement)

4.3 Étude 2 - Donald Trump suspend la contribution américaine à l'OMS

Notre deuxième étude porte sur un sujet qui a trait à l'actualité internationale : l'annonce de la suspension des contributions américaines à l'OMS par le président Donald Trump. La dépêche AFP est parue le 14 avril 2020 à 23h17. Le temps de réaction de certain journaux a été très court. Le Figaro par exemple a publié son article le 15 avril à 0h41, soit seulement 1h24 après la sortie de la dépêche et à une heure très tardive. Donald Trump a annoncé son retrait de l'organisation mondiale de la santé lors d'une conférence de presse qui s'est tenue à Washington vers 12h30 heure locale (soit 18h30 heure de Paris). La plupart des journaux français ont publié leur article le lendemain, au matin.

La quasi-totalité de la presse française a fait le choix de reprendre l'information, y compris les journaux régionaux. Le taux de reprise moyen selon notre dernier indicateur (seuil = 30) est de 46,5%. L'angle journalistique joue ici un rôle important dans ce résultat. En effet, lors de son discours, Donald Trump a également abordé la question du déconfinement, ce que rapporte la dépêche de l'AFP. Beaucoup de journaux n'ont pas traité cette partie du discours et se sont focalisés sur l'annonce du retrait des États-Unis de l'organisation mondiale de

la santé (Le Figaro, Le Point, Paris Match ...). Ce qui l'ont fait ont des taux de reprise généralement plus importants (Europe 1 - 77,2 %, 20 Minutes - 57,1 %).

Il s'agit d'une information majeure, la quantité de journaux ayant repris cette information en est d'ailleurs un indicateur. Les États-Unis étaient jusqu'alors le principal contributeur de l'organisation. De plus cette décision a été prise en pleine pandémie (Covid-19) ce qui est évidemment un facteur aggravant aux yeux de la communauté internationale. Pour des sujets de cette importance, les moyens humains déployés par les journaux sont généralement plus conséquents. La richesse du contenu s'en trouve généralement grandie par un recoupement de sources, une analyse, parfois un questionnement ou le développement de certaines conséquences possibles. En bref, un travail journalistique d'une plus grande qualité.

La source principale de cette information étant le discours de Donald Trump, il est naturel par l'effet de citation, qu'un nombre significatif de caractères soit identiques entre la dépêche et les articles. Ceci affecte évidemment le taux de similarité de chaque article à la hausse. Par exemple, sur les 22 articles analysés dans cette étude, 15 reprennent la citation suivante prononcée par Donald Trump : "Le monde a reçu plein de fausses informations sur la transmission et la mortalité". Il est important de noter que cette portion de citation est exactement la même dans les 15 articles. Aucun journal n'a replacé la citation dans un extrait plus long du discours ni employé des termes différents, ce qui aurait pu être le cas puisqu'il s'agit d'une traduction d'un discours prononcé en anglais. Une traduction complète de la phrase dont cette citation est extraite aurait pu être : "Les nouvelles données qui apparaissent chaque jour dans le monde entier montrent le manque de fiabilité des premiers rapports et le monde a reçu toutes sortes de fausses informations sur la transmission et la mortalité." (à 5min53 du discours). Le discours a duré une heure et huit minutes, beaucoup de reproches ont été formulés à l'encontre de l'OMS mais ce sont presque systématiquement les mêmes que nous retrouvons cités dans la presse. Cette similarité semble démontrer que la majeure partie des journaux n'ont sans doute pas formulé un commentaire sur la source première de l'information (le discours) mais sur une information déjà transformée (probablement traduite et résumée).

L'écart-type du nombre de caractère communs à la dépêche AFP entre les articles est assez important : 764 caractères. Cela signifie qu'en moyenne l'écart du nombre de caractères repris entre deux articles est de 764. L'écart-type relatif (proportionnel à la taille totale de l'article) lui est de 6,4 %. Cela permet de relativiser, les articles ayant repris davantage d'informations ont également davantage de contenu original.

Cette information concerne un sujet étranger, en l'occurrence américain, il paraît donc évident que la presse française ait plus de difficultés à posséder des sources originales comme un correspondant ou une interview par exemple. De ce fait, l'obligation d'avoir recours à des intermédiaires est plus impérieuse.

En résumé, les facteurs qui poussent les taux de similarité à la hausse sont :

- Un recours important à la citation
- Une traduction et des résumés du discours probablement identiques entre les journaux
- L'urgence de l'information
- Le caractère étranger de la source

Les facteurs qui poussent les taux de similarité à la baisse sont :

- La qualité du travail mis en oeuvre pour cette information
- La proportion relativement constante entre le contenu original et non-original.
- L'angle pris par certains journaux qui se sont affranchis des informations concernant le déconfinement et ne traite que le retrait de l'OMS

4.4 Étude 3 - Retraits massifs dans les banques grecques en 2012

Pour notre troisième étude nous avons volontairement choisi un sujet plus ancien, datant du 15 mai 2012. Dans le sillon de la crise financière de 2008, la Grèce doit faire face à une importante crise de sa dette publique qui commence en 2011. En 2012, les rumeurs de "Bank Run" et de "Grexit" commencent à courir. On dit qu'il y a "Bank Run" en économie quand les clients d'une banque privée retirent massivement leur argent de cette banque (en général suite aux rumeurs de faillite de la banque). Ces prophéties sont dites "auto-réalisatrices". C'est à dire que même une banque ayant une santé économique stable fera faillite si l'ensemble de ses clients viennent retirer leur argent. Les banques se prêtant de l'argent mutuellement sur le marché "inter-bancaire", la faillite d'une banque peut entraîner la faillite des autres par "effet domino". L'annonce du retrait de 700 millions d'euros des banques grecques le lundi 14 mai 2012, fut donc une information particulièrement importante et inquiétante. La dépêche de l'AFP est parue le 15 mai 2012 à 21h35 heure de Paris.

La première comparaison des articles à la dépêche donnait des résultats très hauts pour certains journaux et très bas pour d'autres. Toutefois à la lecture, force est de constater que les journaux dont le contenu différait de celui de la dépêche AFP se ressemblaient singulièrement. En réalité, pour cette information, beaucoup de journaux avaient exploité l'information d'une autre agence de presse : Reuters. Ce choix s'explique par la nature économique et financière de cette information. Reuters est davantage spécialisé dans la publication de ce genre type de contenu, son principal concurrent est Bloomberg. Tous deux

proposent des abonnements à des informations financières en temps réel comme des comptes d'entreprises, le cours du change ou celui des matières premières. Nous avons ré-utilisé l'échantillon d'articles précédent pour le comparer cette fois à la dépêche de Reuters concernant le *bank run* grec publiée le 16 mai 2012 à 9h53 (un jour plus tard que celle de l'AFP). Nous obtenons, comme cela était prévisible, des taux de similarité importants pour la majorité des journaux qui n'avaient pas repris la dépêche de l'AFP.

Cette incise met en avant une évidence : l'AFP ne détient pas l'exclusivité de la création de l'information. Cependant, en réunissant les analyses de similarité issues des deux agences de presse (tableau 3) on obtient des taux de reprise singulièrement élevés. Cela indique que, sur cette information, seulement deux sources sont à l'oeuvre et que la réécriture des journaux a été faible.

Le site d'information en ligne bfmtv.com a un score relativement plus faible que ses congénères, précisément en raison d'une réécriture. Celle-ci est cependant relativement effacée, ainsi le site a remplacé « (les) élections qui doivent avoir lieu le 17 juin » par « (les) élections législatives qui devraient avoir lieu dans le courant du mois de juin ». On remarque ici que la réécriture amène curieusement à la perte d'une information, celle de la date prévue des élections législatives. Celles-ci ont d'ailleurs bien eu lieu le 17 juin. On peut éventuellement arguer que cette réécriture envisage une panique bancaire généralisée qui aurait pu remettre en cause les élections. Catastrophisme dont cette chaîne d'information en continu fait souvent la démonstration. Les mentions "NDLR" (Note de la rédaction) ont également été déplacées, déjà présentes dans la dépêche de Reuters, elles ont été placées au début des parenthèses plutôt qu'à la fin. De nombreux paragraphes de la dépêche ont été supprimés, mais un paragraphe a été ajouté : « Une nouvelle réunion est prévue ce mercredi à 13h00 (10h00 GMT) entre le président Papoulias et les dirigeants des formations politiques pour la mise en place d'un gouvernement chargé d'expédier les affaires courantes jusqu'à la tenue des nouvelles élections. ». Enfin, et ce sera notre dernière remarque concernant cette réécriture, la dépêche Reuters est signée ainsi : « Harry Papachristou, Ingrid Melander et Karolina Tagaris, Henri-Pierre André et Julien Dury pour le service français, édité par Gilles Trequesser » quand celle présentée par BFM TV, efface leurs contributions à trois journalistes et se contente d'une signature lacunaire : « Harry Papachristou et Ingrid Melander; Henri-Pierre André pour le service français ». Curieuse censure.

L'article du journal Le Monde est intéressant dans la mesure où il obtient des scores de reprises significatifs pour les deux dépêches (20,8 % pour Reuters et 51,2% pour l'AFP). L'article est en réalité un conglomérat des deux dépêches. Le journal indique qu'il repose sur le travail de l'AFP à travers la mention "Le Monde avec AFP" présente à la fin de l'article mais pas celui de Reuters. Pourtant un peu plus de 20% de l'article contient des passages qui lui sont emprunté : « il s'attend à des sorties totales de l'ordre de 800 millions d'euros, a ajouté le président grec. » ou encore « elles reconnaissent un sentiment de "peur qui

pourrait évoluer en panique". ». Ainsi si l'on compare l'article à la réunion des deux dépêches le taux de reprise monte à 67,7% (et non à 72%, la somme des deux taux de reprises car des passages sont également similaires au sein des deux dépêches). L'article a également eu recours à l'écriture. La phrase : « Devant l'incertitude politique et économique, les Grecs ont procédé à des retraits massifs sur leurs comptes en banque et si les autorités n'évoquent pas de "panique bancaire" » remplace par exemple une citation directe de Károlos Papoúlias.

En résumé, les facteurs qui poussent les taux de similarité à la hausse sont :

- Pour la plupart une reprise quasi intacte des deux dépêches

Les facteurs qui poussent les taux de similarité à la baisse sont :

- Le recours plus ou moins important à la réécriture
- L'exception d'un réel travail de fond pour le journal La Croix

4.5 Étude 4 - Mort de George Floyd

George Floyd est un homme noir de 46 ans et père de deux enfants. Il est mort des suites d'une interpellation policière le 25 mai 2020 à Minneapolis. Son interpellation a été filmée et retransmise en direct sur Facebook. On peut y voir l'homme à terre subissant une étreinte au niveau de la nuque et se plaignant de ne pas pouvoir respirer : « *I can't breathe* ». Les passants qui filment la scène s'indignent de la situation et leur colère s'amplifie lorsque George Floyd ne semble plus montrer de signe de vie. Il sera transféré quelques minutes plus tard à l'Hennepin County Medical Center où il sera déclaré mort. Le lendemain 26 mai, d'importantes manifestations seront organisées à Minneapolis pour protester contre l'impunité des violences policières à l'encontre des noirs américains. Ces mouvements de contestation se propageront à travers l'ensemble des États-Unis ainsi que d'autres pays dont la France.

La première dépêche de l'AFP relatant l'évènement a été publiée le 26 mai à 21h23 (heure de Paris). Très rapidement, une seconde dépêche est publiée, deux heures plus tard (23h22). Elle reprend des éléments de la première dépêche et fait état du limogeage des quatres policiers impliqués dans l'interpellation. Les manifestations importantes qui se tiendront dans la soirée n'y sont pas encore décrites, puisqu'elles auront lieu quelques heures plus tard. La presse a parfois fait le choix de reprendre les informations de la première dépêche de l'AFP, parfois de la seconde, parfois des deux. La première se nomme Indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation, la seconde : Etats-Unis: quatre policiers limogés après la mort d'un Noir lors de son interpellation. Pour chaque article, nous avons effectué les tests de comparaison avec la dépêche qui semblait la plus proche du contenu de l'article. Lorsque nous avons comparé

l'article à l'association des deux dépêches nous avons ajouté la mention "(+)" au nom de la dépêche (voir tableaux).

On remarque tout d'abord que la reprise des dépêches a été forte pour cet évènement avec un taux de similarité moyen de 74,9% et un taux médian de 87,3% (indice % Tot. Substrings). Très souvent, le contenu original des articles se trouve dans le chapeau et dans les quelques lignes introductives qui le succèdent. Par la suite, ce sont généralement des phrases des dépêches qui sont reproduites à l'identique, parfois dans un ordre différent. Généralement la vidéo de l'interpellation est intégrée sur la page de l'article. Elle peut éventuellement faire l'objet d'une description écrite des faits qui s'y trouvent.

La première dépêche de l'AFP contient un erratum. Le mouvement Black Lives Matter est traduit entre parenthèses par "Le vie des Noirs compte", il y a donc une erreur bénigne dans le déterminant choisi pour la traduction. Dix journaux parmi les articles analysés ont repris la phrase associée à cette traduction, deux n'ont pas corrigé l'erreur : The Huffington Post et le journal La Provence. On remarque donc que certains journaux, au delà du fait qu'ils reprennent très largement le contenu des dépêches de l'AFP (99,7% : La Provence, 94,2% : The Huffington Post), ont probablement des degrés de relecture assez superficiels.

Beaucoup de journaux indiquent qu'ils ont édité leur contenu. La mention "mise à jour le" n'est pas présente sur tous les sites de presse. Cependant lorsqu'elle l'est, la date initiale de publication de l'article est très souvent différente de la date à laquelle il a été mis à jour. Souvent l'information est mise à jour dans les jours qui suivent la publication (Le Monde, publié le 27, mis à jour le 28 - France Info, publié le 29 mis à jour le 30 - LCI, publié le 27, mis à jour le 29 etc.). Il se peut toutefois qu'elle le soit un peu plus tard (20 Minutes, publié le 27/05, mis à jour le 03/06). Il serait intéressant de savoir quelles informations sont modifiées lors de ces mises à jour. S'agit-il d'ajouts d'informations, de corrections orthographiques annodines ou de rétroactions sur des faits qui auraient été contestés par la suite ? L'historique des modifications des articles n'est consultable que par les journalistes et nous n'y avons pas eu accès. En revanche, nous pouvons supposer que l'intégration au sein des sites de presse de mécanismes participatifs favorise la correction et la mise à jour (même tardive) des articles qui s'y trouvent. Le site 20 Minutes permet par exemple l'envoi de corrections et possède une zone dédiée aux commentaires dans laquelle les lecteurs peuvent échanger sur le sujet en question. C'est peut être une des raisons qui permet d'expliquer cette mise à jour si tardive de l'article. Dans la zone de commentaires certains utilisateurs interpellent sur le choix de certains mots plutôt que d'autres, sur la véracité des faits, etc.

Enfin, certains journaux ont également fait le choix de traiter les réactions de certaines stars et hommes politiques qui ne sont pas présentes dans les dépêches. Deux réactions ont été particulièrement couvertes. D'une part le tweet de LeBron James, champion noir américain de basket considéré comme

l'un des meilleurs joueurs de sa génération, relayant la photo de Colin Kaepernick (joueur de football américain) qui s'était agenouillé durant l'hymne national pour protester contre les violences policières. D'autre part, la réaction de l'ancien vice-président des États-Unis, Joe Biden, en lice pour les élections américaines du 3 novembre 2020. Il a déclaré : "C'est un rappel tragique que ceci n'est pas un incident isolé, mais un drame qui fait partie d'un cycle d'injustice systématique qui continue d'exister dans notre pays"[12]. Trois journaux ont traité ces réactions : France Info, Le Journal de Saône-et-Loire, l'Est Républicain[13]. France Info qui a davantage pris le temps pour couvrir l'information a également pu traiter les débordements survenus lors des manifestations de la nuit du 28 mai.

Cette tragédie en appelle d'autres. Lors de notre travail, les moteurs de recherche nous ont également renvoyé des faits plus anciens rassemblés sous le mot clé "indignation" particulièrement employé pour ce genre d'événements. Ce fut le cas par exemple pour Alton Sterling, jeune noir américain abattu par la police américaine alors qu'il vendait des CD sur un parking en 2016 ou celui de Ahmaud Arbery, jeune ambulancière de 26 ans, abattue 12 jours avant la mort de George Floyd après que la police se soit trompée de porte dans un immeuble de Louisville[14].

En résumé, les facteurs qui poussent les taux de similarité à la hausse sont :

- Une très forte reprise des informations des dépêches de l'AFP probablement due à un événement de dimension internationale

Les facteurs qui poussent les taux de similarité à la baisse sont :

- Le commentaire direct de la vidéo de l'interpellation de George Floyd
- La rédaction courante d'un paragraphe introductif inédit
- La probable mise à jour des articles
- Le traitement des réactions politiques et médiatiques autour de l'événement
- Un délai qui permet de traiter également les émeutes des jours qui ont suivis

[12] Propos original (28 mai 2020) : "*Is a tragic reminder that this was not an isolated incident, but a part of an ingrained systemic cycle of injustice that still exists in this country*".

[13] Le contenu des articles du journal de Saône-et-Loire et de L'Est Républicain sont rigoureusement les mêmes. Il existe un article des Dernières Nouvelles d'Alsace identique également. Nous ne l'avons pas intégré dans nos analyses.

[14] Paris Match (7 juillet 2016) : *Indignation après la mort d'un Afro-Américain lors de son interpellation* et Le Monde (13 mai 2020) : *Aux Etats-Unis, indignation après la mort d'une jeune ambulancière noire, tuée par la police.*

4.6 Étude 5 - Allocution d'Emmanuel Macron du 15 juin 2020

Le 15 juin 2020, Emmanuel Macron prenait la parole dans la continuité des interventions télévisées auxquelles il s'est livré concernant la crise du coronavirus. Il s'agissait de la quatrième intervention du président depuis le début de la crise. Il y a notamment évoqué le passage de la région parisienne en "zone verte", c'est à dire la levée de la majeure partie des restrictions prises pour endiguer l'épidémie, la réouverture des frontières européennes et la reprise des cours dans les écoles. Plus généralement, le discours a été l'occasion pour le président de la république de saluer le succès, selon lui, de nombreux aspects de la gestion de la crise. Ces éloges ont suscité de nombreuses et vives réactions dans les rangs de l'opposition qui ont été largement citées et reprises par la presse. L'usage des dépêches pour le traitement de cet évènement semblent avoir été largement délaissé. La majorité des résultats obtenus ne permettent pas d'établir un lien de similarité tangible entre les dépêches et les articles. Au mieux nous pouvons avancer qu'ils traitent d'un sujet commun. La circulation de l'information est ici très difficile à mesurer, étant donné les nombreuses dépêches qui sont parues au sujet de ce discours. Nous en avons utilisé deux : Macron, louant sa gestion de la crise, accélère le déconfinement et trace un "nouveau chemin" et Le satisfecit de Macron sur la gestion de crise fait bondir l'opposition. Il est difficile de savoir lesquelles ont été exploitées par les journaux, si tant est qu'elles l'aient été. De plus, il n'est pas rare que les sites de presse aient publié plusieurs articles sur ce discours en adoptant des angles de traitement différents. C'est le cas de France Info par exemple[15].

Plusieurs raisons peuvent expliquer un faible recours aux dépêches. Tout d'abord, le discours d'Emmanuel Macron a été diffusé par les grandes chaînes de télévision françaises. La source principale de l'événement n'a donc pas été rapportée mais est à la disposition directe de quiconque souhaitant la consulter. L'AFP peut apparaître, de ce fait, comme un intermédiaire non-essentiel au traitement de cette information. D'autre part, c'est une information nationale jugée importante par les médias. Les moyens humains pour la traiter ont donc été plus importants. Par exemple, l'article du journal Le Monde est le produit du travail de deux journalistes, Alexandre Lemarié et Olivier Faye. La ligne éditoriale des différents journaux prend également une part importante dans le traitement de l'information, en particulier dans le choix des personnes interrogées pour réagir à l'intervention. Le journal l'Humanité a donc très logiquement fait appel à un porte parole du PCF pour opérer le commentaire du discours d'Emmanuel Macron quand le Figaro choisissait de rapporter les propos d'Olivier Marleix, député Les Républicains de la deuxième circonscription d'Eure-Et-Loir. Dans une approche plus locale, France Bleu choisissait le té-

[15] France Info : Retrouvez l'allocution d'Emmanuel Macron pour détailler la phase trois du déconfinement / Le discours d'Emmanuel Macron a "sonné comme un renforcement du modèle libéral productiviste", dénonce l'écologiste Marie Toussaint / Discours d'Emmanuel Macron : le retour en classe le 22 juin est "une bonne nouvelle", réagit le Snes FSU etc.

moignage d'Elodie Lagarde, membre de la Fédération Syndicale Unitaire (FSU) de Dordogne. L'angle adopté, le choix des témoignages et le véritable travail de rédactions journalistiques mise en place pour cet évènement sont donc autant de facteurs qui ont participés à la diversité des contenus.

Cependant, l'addition de tous ces paramètres et l'usage sans doute très délaissé des dépêches[16] ne signifient pas que l'information n'ait pas circulé. En témoigne notamment l'omniprésence du terme "satisfecit" dans les articles pour caractériser le discours. L'usage communément peu répandu du mot et sa répétition dans les articles du Monde, de France Info, du Figaro, du Courrier Picard etc. font penser à un phénomène d'emprunt entre journaux plutôt qu'à un hasard. Le processus de citation entraîne comme souvent une similarité entre les articles : la réaction de Jean-Luc Mélenchon est par exemple souvent citée.

En résumé
À l'exception du journal Courrier International, l'usage des dépêches ne peut être avéré.

Les facteurs qui poussent les taux de similarité à la hausse sont :

- La présence de citation commune tirée du discours entre les dépêches et les articles

Les facteurs qui poussent les taux de similarité à la baisse sont :

- Un travail de rédaction
- L'usage de témoignages parfois inédits
- Une prise de position éditoriale
- L'angle adopté pour traiter le sujet
- La rédaction de plusieurs articles au sein d'un même journal pour traiter des aspects différents du discours

4.7 Étude 6 - Alunissage chinois sur la face cachée de la lune

Le 3 janvier 2019, l'administration spatiale nationale chinoise (CNSA) parvenait à alunir un module d'exploration sur la face cachée de la lune. Il s'agit d'une première mondiale, ni les américains, ni les soviétiques ne s'y étaient risqués jusqu'alors. De fait, cette opération présente des difficultés supplémentaires par rapport à un alunissage, déjà difficile, sur la face visible. Tout d'abord, s'il est

[16] En réalité, étant donné les indicateurs, seul Courrier International semble avoir eu recours aux dépêches comme en témoigne la présence de passages communs entre leur articles et les dépêches en dehors des effets de citations. Ceci s'explique probablement par le fonctionnement particulier de ce journal, très ancré dans la traduction de contenus tiers.

possible d'établir une communication "directe" entre la face visible et la terre, la mise en place d'une communication pour joindre la face cachée nécessite un relai. D'autre part, la face cachée est soumise à des conditions bien plus difficiles. La géographie des sols est plus vallonnée (ce qui rend l'alunissage des modules plus délicat) et les températures moyennes sont beaucoup plus basses. La réussite chinoise est donc un exploit important et a été couverte par l'ensemble de la presse française, nationale comme régionale. Cette couverture semble s'inscrire dans la poursuite d'une fascination populaire pour l'exploration spatiale amorcée en 1957 par le lancement du premier satellite artificiel Spoutnik 1. La dépêche de l'AFP est parue le 3 janvier 2019 à 9h25 : *La Chine, première à poser un engin sur la face cachée de la Lune*. L'AFP a également fourni des infographies du module chinois qui ont été amplement reprises par les sites de presse.

Une première remarque à ce sujet, la Chine possède ses propres agences de presse. La plus ancienne et la plus importante est la Xinhua, communément appelée en France "Agence Chine nouvelle". La seconde est le China News Service. L'agence Chine nouvelle délivre des informations en français et a ouvert une antenne en 2013 rue du Faubourg-Saint-Honoré[17]. Leurs informations sont souvent prises avec précautions par les médias occidentaux puisqu'elles apparaissent comme le relai officiel de la propagande d'état chinoise. Pour la journaliste Pascale Nivelle, correspondante du journal Libération à Pékin de 2006 à 2009 : « L'agence Xinhua expose l'information officielle policée qu'elle voudrait étendre au reste du monde. ». Elle n'a par exemple pas couvert la résistance Tibétaine ou le décès de Liu Xiaobo, prix nobel de la paix, mort dans en prison en 2017. C'est sans doute en partie pour ces raisons que les journaux ont tout de même fait le choix de reprendre en premier lieu la dépêche de l'AFP et non pas celle de Xinhua qui aurait pu paraître plus appropriée en cette situation. Cependant, certains articles viennent ajouter des informations de l'agence chinoise pour renforcer une structure généralement construite autour de la dépêche de l'AFP.

Certains journaux ont fait le choix de publier la dépêche telle quelle. On est d'ailleurs surpris de constater que c'est le choix adopté par le magazine de vulgarisation scientifique Sciences et Avenir alors même que ce sujet relève pleinement de son domaine de compétence. Dans la même lignée, le journal régional La Voix du Nord, reprend intégralement la dépêche sans citer l'AFP et appose même la signature de l'un de ses journalistes à la fin de l'article. Le Nouvel Obs quant à lui, s'il ne change rien au contenu de la dépêche, précisera tout de même en signature "L'Obs avec AFP".

Pour ce sujet, on constate étonnamment un important phénomène de réécriture. Bien qu'on ait pu le constater dans les études précédentes, celui-ci était resté très superficiel et réservé à un petit nombre de journaux. Ici, la réécriture a été pratiquée par beaucoup de journaux (Paris Match, France 24, France Bleu, Libération, l'Express etc.). Elle demeure souvent superficielle, avec une réorgan-

[17] Libération (12 juillet 2013) : *À Paris, Chine nouvelle veut épater la galerie.*

isation de l'ordre des phrases de la dépêche et des changements de mots épars. Le mot "engin" est échangé pour "robot" chez Paris Match, "l'alunissage inédit" est préféré aux termes de "premier alunissage" chez France 24 et France Bleu fait le choix de taire le nom de la télévision chinoise en changeant "la télévision d'Etat CCTV" par "la télévision publique chinoise". Ces reformulations étant bien sûr non exhaustives. Parfois on traduit aussi : la dépêche donne notamment l'heure d'alunissage du module au fuseau horaire du méridien de Greenwich (GMT[18]), certains journaux comme France Bleu ont fait le choix de transposer cet horaire à l'heure de Paris, sans doute plus évident pour les lecteurs. Cette récriture et cette réorganisation donnent lieu à des répétitions sans doute fortuites. France 24 par exemple répète par deux fois que la face visible de la lune : "offre de nombreuses surfaces planes pour se poser". France Bleu évoque le lancement à venir du Chang'e-5 pour récolter des échantillons, puis une ligne plus tard le lancement d'un "robot" dans le cadre d'une nouvelle mission. Il s'agit en réalité de la même mission.

Les journaux ajoutant de l'information en conservant la structure de la dépêche de l'AFP ajoutent très souvent les mêmes informations et dans les mêmes termes. Les ajouts concernent souvent le lancement de la future station spatiale chinoise "Palais céleste" qui devrait être assemblée pour 2024 et les températures mesurées sur la face cachée de la lune qui ne sont pas présentées dans la dépêches de l'AFP. Les Echos et France Info rapportent ces informations en utilisant les mêmes mots.

Viennent en dernier lieu les journaux qui ont réalisé un travail complet de traitement. Si on peut reconnaître les informations de la dépêche dans l'article de l'Express par exemple, la reformulation est tellement importante que l'article n'a plus grand chose à voir avec les formulations de l'AFP. De plus, le journal appose à son article une interview exclusive de Francis Rocard, astrophysicien et responsable du programme d'exploration du système solaire à l'agence spatiale française (CNES), ce qui lui confère une réelle valeur ajoutée. LCI également a rédigé un article complet sur le sujet appuyé par une interview du spationaute français Jean-François Clervoy. L'article le plus approfondi revenant très probablement à France Culture qui a fait l'effort d'un article de plusieurs pages contenant photographies, témoignages, infographies et montages vidéo. Ce développement a sans doute été permis par le "recyclage" d'un article plus ancien, la date de publication initiale indiquée étant le 28 septembre 2018.

En résumé, les facteurs qui poussent les taux de similarité à la hausse sont :

- La réorganisation des phrases qui n'affecte pas les taux de similarité

[18]L'appellation GMT utilisée par l'AFP est en réalité ici un abus de langage. GMT (Greenwich Mean Time) est une référence temporelle historique utilisée au XXème siècle et désignant l'heure solaire moyenne au méridien de Greenwich. Aujourd'hui on utilise le temps universel coordonné (UTC) établi sur des horloges atomiques. GMT est parfois utilisé à tort pour désigner l'UTC+0.

- L'usage très utilisé (même comme trame) de la dépêche de l'AFP
- Une réécriture sporadique qui ne change que quelques mots
- La bouderie des dépêches de Xinhua

Les facteurs qui poussent les taux de similarité à la baisse sont :

- Un travail de rédaction complet pour certains journaux
- L'ajout d'informations supplémentaires mais souvent dans les mêmes termes
- Les interviews de spécialistes en la matière

4.8 Conclusion sur les études

La reprise partielle ou totale des dépêches de l'AFP par les sites de presse est indéniable. Il nous paraît important de rappeler que nos taux de similarité ne reflètent que l'emprunt d'informations à ces dépêches. Or, les dépêches ne constituent pas la seule source d'information des journaux en ligne. L'analyse des différents articles a montré que certaines informations étaient également parfaitement similaires entre les articles mais elles ne sont pas prises en compte par les indices car elles ne proviennent pas des dépêches. Beaucoup d'articles doivent leur existence à une agrégation d'information recopiée à l'identique, davantage qu'à un travail de rédaction. La nature de l'information semble jouer son rôle dans le traitement de l'information. Les informations internationales sont souvent laissées à la grande discrétion des dépêches mais il arrive que des sujets soient plus traités que d'autres (l'alunissage du module chinois a été plus travaillé que l'annonce du retrait américain de l'OMS). Les informations de politique nationale, comme l'allocution d'Emmanuel Macron du 15 juin 2020, sont beaucoup plus travaillées et examinées. On constate également que l'usage de la réécriture est très rarement le produit d'une reformulation totale des phrases des dépêches mais bien plus souvent une réorganisation de l'ordre des informations parfois accompagnée du changement de quelques mots. L'agrégation de l'ensemble des taux calculés donne un indice %Tot. Substring moyen de 47,7% (médian : 49,5%) ce qui signifie qu'en moyenne la moitié des articles sont identiques au contenu des dépêches[19].

5 Analyse

Nous avons donc constaté, à travers ces études, que la reprise directe et sans réécriture des dépêches des agences de presse par les journaux en ligne est un phénomène courant et prononcé. Il est donc légitime de nous questionner sur les raisons de ces reprises massives. Nous nous appuierons pour ce faire sur

[19] L'indice de Jaccard moyen, lui, est de 50,7% (médian : 46,6%) cela signifie qu'environ la moitié des mots présents dans les dépêches sont les mêmes que ceux des articles.

les travaux de Pierre Bourdieu, qui avait déjà décrit ce processus de reprise à travers son concept de *circulation circulaire de l'information* au sujet de la presse papier et de la télévision. Nous nous questionnerons sur son applicabilité sur le champ du web et nous livrerons d'autres éléments de réponse avant de conclure.

5.1 Un point sur les agences de presse

Les agences de presses sont nées dans les années 1880. En France, les plus renommées d'entre elles sont : l'Agence France Presse (AFP), Reuters et l'Associated Press (AP). Elles vendent à des journaux des photographies, des infographies, des enregistrements et des dépêches sur des thématiques diverses issues du monde entier. Elles sont très largement exploitées par la presse française qui n'a pas toujours les moyens matériels, temporels et financiers d'envoyer des correspondants sur place pour traiter un sujet. Elles permettent également plus de liberté au comités de rédactions qui peuvent sélectionner les sujets sur lesquelles ils souhaitent travailler tout en continuant d'en couvrir d'autres en reprenant les informations des agenciers.

5.2 Qu'est-ce que la *circulation circulaire* de l'information ?

Le terme "circulation circulaire de l'information" a été popularisé par le sociologue Pierre Bourdieu dans son essai sur la télévision parue aux éditions Raison d'agir publié en 1996. Cet essai est une retranscription de vidéos initialement réalisées pour le Collège de France. Même si dans celles-ci Pierre Bourdieu entamera une gestuelle circulaire au moment du développement de son concept, le terme "circulation circulaire" ne sera présent que dans l'essai. Il y décrit un phénomène de clôture de l'information dans lequel les médias, qui regroupent dans son analyse la télévision et les journaux, seraient amenées à traiter des sujets semblables. Il commence par questionner les raisons de ce qui constitue l'événement médiatique et en cela il rejoint les travaux de Patrick Champagne sur le sujet[20]. Pour appuyer son propos il prend l'exemple saisissant des revues de presse. Chaque journal, qu'il soit papier ou télévisé, a pour coutume professionnelle d'effectuer un panorama des sujets traités par les autres médias pour en dégager ce qu'il s'apprête à traiter lui même. Pour reprendre ses propres mots : « Pour savoir ce que l'on va dire, il faut savoir ce que les autres ont dit ». Il note du même fait, que : « personne ne lit autant les journaux que les journalistes ». De cela il tire une conclusion qui peut paraître surprenante de prime abord : selon lui, c'est la mise en concurrence des différentes entités médiatiques et la course à l'audimat qui conduit à l'homogénéisation des hiérarchies d'importance accordées aux contenus qu'elles diffusent. À l'inverse, une collaboration entre les médias pourrait permettre une répartition des sujets dont la couverture est

[20] Consulter notamment *La misère du Monde* et *Les journalistes font-ils l'évènement ?*, Éditions de la Sorbonne

jugée nécessaire et entraînerait une diversification des thématiques abordées au sein de chaque média.

Il y a un second niveau de lecture lorsqu'on évoque "la circulation circulaire de l'information". On pourrait également y voir une critique des procédés "métadiscursif" utilisés par les médias en particulier télévisées. Il est fréquent que des médias établissent un dialogue sur le traitement de l'information au sein d'autres médias (Quotidien, Acrimed etc.). Dans une perspective encore plus égocentrée certaines émissions ont coutume de créer l'événement au coeur même de leur émission puis d'en opérer le commentaire lors de l'émission suivante, ce procédé pouvant être prolongé et répété à dessein. Ceci est particulièrement symptomatique de l'usage devenu boulimique du terme de "polémique". La polémique pouvant, entre autre, naître sur une émission des propos d'un invité ou d'un animateur. Celle-ci pourra par la suite subir le commentaire d'autres médias et/ou du média duquel elle est issue.

5.3 Typologie des organes de publications sur internet

Pour corroborer notre analyse, il nous parait également primordial de connaître les différents vecteurs qui participent de cette circulation de l'information en ligne. Permettons-nous donc un rappel des acteurs publiant sur le web en nous appuyant sur les travaux de Franck Rebillard, professeur à l'Université Lyon 2. Dans son ouvrage *Du traitement de l'information à son retraitement* [11] puis dans une publication ultérieure [12], Franck Rebillard propose une typologie des acteurs journalistiques sur internet. Cette typologie permet de caractériser le champ d'activité dans lequel s'inscrit notre étude. Elle décompose les acteurs responsables de la publication en ligne de la façon suivante :

- **La presse en ligne** : Traitant généralement une information généraliste, elle représente les organes séculaires de l'information tels que : *Le Monde*, *TF1* ou *Radio France*

- **Les agences de presse en ligne** : Elles fournissent aussi bien des informations généralistes que des informations spécialisées (notamment financières) et sont depuis longtemps implantées sur internet. *Reuters* et l'*AFP* en sont les exemples les plus célèbres.

- **Les webzines** : Contrairement à la presse en ligne, internet est leur support d'origine. Ils véhiculent de façon privilégiée des informations dites d'opinion ou spécialisées en cherchant généralement des « niches » leur permettant de se différencier et de subsister.

- **Les blogs** : Parangons de la publication dite "*autoritative*" [12], ils avaient connu un fort engouement au début du siècle (en octobre 2005, le site MySpace était le quatrième site le plus consulté au monde derrière ceux de Yahoo!, AOL,

et MSN et devant eBay et Facebook). L'émergence des réseaux sociaux semble toutefois leur avoir ravi une audience importante.

- **Les portails** : Historiquement utilisés par de nombreux utilisateurs comme "porte d'entrée" du web, les portails continuent de rencontrer un public important. *Yahoo.com* était ainsi le cinquième site le plus consulté en France en 2020[21]. *Yahoo!* a été rachetée en juillet 2016 par la société *Verizon* qui avait déjà procédée au rachat d'AOL (autre portail historique du web), en mai 2015[22].

- **Les agrégateurs** : Considérés comme méta-éditoriaux, ils rassemblent des articles publiés par d'autres sites d'information généralement à l'aide de moteurs de recherche (exemple : *Google News*).

- **Les bases d'archives** : Constituées en bases de données, elles regroupent de façon plus ou moins ouvertes du contenu publié en ligne et quelques fois par d'autres moyens de distribution (radio, presse écrite). Exemple : *Europresse*.

Il parait également judicieux de rajouter les réseaux sociaux dans cette typologie. Devenus de véritables "infomédiaires" comme les nomme lui-même Franck Rebillard dans un article ultérieur [13], ceux-ci se sont imposés comme des relais incontournables de l'information. À travers la multiplication des onglets permettant le "partage" et la facilitation de la transmission de contenu entre les sites, ils sont à l'origine de nouvelles façons de consommer et de faire circuler l'information. En ce sens et en s'inspirant de recommandations issues du Dublin Core[23], Facebook a déployé en 2010 l'Open Graph Protocol, permettant de faciliter la perméabilité des contenus sur le web[24]. C'est notamment ce protocole qui permet l'affichage de l'image, du titre et de la légende d'un article par la seule transmission du lien auquel il est associé sur la plupart des applications de messagerie modernes (Et en tout premier lieu celles que possède Facebook : Messenger, WhatsApp, Instagram).

D'une façon générale, on remarque que les moyens technologiques sont très souvent mis au service de la facilitation de la circulation de l'information. C'est le cas de l'Open Graph Protocol, c'est aussi celui de l'HTML5 qui s'impose aujourd'hui comme une norme incontournable du web. Proposé en 2014 par le W3C[25], il introduit de nouvelles balises telles que "article" ou "main" qui

[21] Alexa.com, le 10 mai 2020

[22] Le Monde (12 mai 2015) : *Verizon rachète AOL pour plus de 4 milliards de dollars*

[23] Dublin Core est à l'origine un consortium d'experts travaillant sur le sujet du web sémantique s'étant réunis pour la première fois en 1995 dans la ville de Dublin (Ohio). Il a donné naissance à de nombreuses recommandations en la matière.

[24] Facebook.com : « The Open Graph protocol enables any web page to become a rich object in a social graph. For instance, this is used on Facebook to allow any web page to have the same functionality as any other object on Facebook. »

[25] World Wide Web Consortium, organisme de standardisation à but non lucratif, fondé en

permettent, entre autre, de récupérer les informations de publication automatiquement et plus facilement sur les sites qui les hébergent (scrapping). Ces processus participent de la facilitation d'une forme d'interdiscursivité entre les différents acteurs du web. Interdiscursivité que Patrick Charaudeau en 2006 définissait comme étant « une notion générique de mise en relation de ce qui a été déjà dit quelle que soit la forme textuelle sous laquelle apparaît ce déjà dit ». Cette interdiscursivité semble donc également encouragée par le progrès technique.

On pourrait arguer que cette facilitation de la circulation de l'information est le résultat de la mise en application de la puissance des *infomédiaires* (Google, Facebook) en matière de développement technologique et que cette simplification intéresse. Étant des entreprises qui tirent leurs revenus du traitement, de l'analyse puis de la revente des données, il est évident que l'homogénéisation généralisée des contenus qu'ils exploitent facilite leur travail et ainsi, participe de leur rentabilité. On le rappelle, l'Open Graph Protocol, sous une désignation d'apparence spécieuse qui pourrait évoquer les initiatives des projets Open Source est une technologie développée par Facebook. À l'inverse, on comprendrait que les acteurs qui sont à l'origine de la création du contenu, notamment les organes de presse, aient au contraire tendance à réprimer cette circulation pour conserver la primauté des revenus qui y sont associés.

Il nous paraît également nécessaire de distinguer deux types de partage d'information en ligne. Le premier constitue une reproduction complète ou partielle d'un contenu initial transmis d'un site à un autre. C'est par exemple le cas des dépêches de l'AFP qui sont reproduites, souvent à l'identique, sur de nombreux sites de presse. Le second, qui s'apparente davantage à un effet de citation, est celui qui correspond à l'utilisation des liens hypertexte. Par cet usage, qui permet la navigation, l'utilisateur peut consulter ou transmettre une source d'information sans l'altérer ou la reproduire. C'est, cette fois-ci, les méthodes utilisées par Google News.

Dans nos études précédentes nous nous sommes concentrés essentiellement sur deux acteurs : les agences de presse et la presse en ligne. Il est toutefois essentiel de garder à l'esprit que ces deux acteurs s'inscrivent dans un réseau de transmission beaucoup plus large d'échange de l'information. On remarque d'ailleurs à ce propos que cette circulation s'impose quasi naturellement et qu'elle paraît très difficile à réprimer. En 2005, l'AFP avait porté plainte contre Google News lui réclamant 17,5 millions de dollars pour avoir exploité son contenu sans son consentement. Au terme du procès, les deux partis étaient parvenus à un accord dans lequel Google s'engageait à rémunérer L'AFP pour

octobre 1994

exploiter ses dépêches. Cet exemple constitue l'exception, la plupart des sites de presse en ligne consentent en réalité tacitement à l'exploitation de leurs articles par Google News (ou d'autre agrégateurs) en contrepartie du trafic qu'il génèrent sur leur site.

6 Conclusion

L'information est cette petite brique par laquelle nous construisons la connaissance. Si une définition commune n'a pas encore réussi à faire consensus, une chose est certaine, l'information ne peut s'empêcher de circuler. À travers ce travail nous avons analysé un comportement particulier de cette circulation sur un champ d'étude restreint : la presse en ligne. Plusieurs observations découlent de ces recherches. D'une part, les dépêches fournies par les agences de presse ne constituent pas uniquement la matière première nécessaire à la construction d'articles plus fournis. Elles sont souvent reprises telles quelles et publiées directement sur les sites de différents journaux. En cela, internet est un espace où la circulation de l'information est particulièrement favorisée voir contrainte. La presse traditionnelle (exception faite de la presse gratuite) n'a jamais eu à l'idée de publier sans réécriture des dépêches d'agence de presse sur leurs journaux papier. Pourtant, sur le web, elle s'y livre abondamment. Elle semble souscrire à une règle tacite d'internet où la rapidité du relai d'une information paraît prévaloir sur la qualité de son traitement. En réalité, bien plus qu'à un paradigme intrinsèque du web, la presse en ligne se heurte à un modèle économique qu'elle subit. Sur internet, ce ne sont pas les créateurs de contenus les plus appréciés qui sont les mieux rémunérés mais ce sont ceux qui les transmettent à l'audience la plus importante. La course au "trafic" sur les sites de presse en ligne semble ainsi plus que jamais participer à l'homogénéisation des contenus, comme Pierre Bourdieu l'avait déjà décrit au sujet de l'audimat pour la télévision. On retrouve d'ailleurs en cela une absolue synonymie de sens entre les mots "trafic" et "audimat". Seul le contexte (la télévision ou internet) vient légitimer l'emploi d'un des termes plutôt que l'autre. Cette homogénéisation ainsi entraînée apparaît d'autant plus préoccupante que les supports numériques sont de plus en plus privilégiés pour la consultation d'informations d'actualité. Il nous semble alors qu'une prise de distance face à l'instantanéité de l'actualité, une documentation plus fournie et plus en profondeur de sujets choisis et l'acceptation avouée de l'impossibilité de "tout traiter", pourraient constituer un début de réponse propre à délier le carcan délétère dans lequel la presse semble s'être enfermée sur le web.

References

[1] Claude Baltz. Le concept d'information : essai de définition. *Persée*, 1995.

[2] Seng Hansun Brinardi Leonardo. Text documents plagiarism detection

using rabin-karp and jaro-winkler distance algorithms. *Indonesian Journal of Electrical Engineering and Computer Science*, 2017.

[3] Eric Dacheux. La communication : éléments de synthèse. *Persée*, 2004.

[4] Fred I. Dretske. *Knowledge and the Flow of Information*. Center for the Study of Language and Inf, 1981.

[5] Claude Shannon et Warren Weaver. *La théorie mathématique de la communication*. The Board of Trustees of the University of Illinois, 1949.

[6] Matthew A. Jaro. Advances in record linking methodology as applied to the 1985 census of tampa florida. *Journal of the American Statistical Society*, 1989.

[7] Olimpia Lombardi. What is information? *Foundations of Science*, 9, 2004.

[8] Marshall McLuhan. *Pour comprendre les médias*. Le Seuil, 1977.

[9] John Von Neumann. *L'ordinateur et le cerveau*. Flammarion, 1956.

[10] Ian Oliver. *Programming Classics: Implementing the World's Best Algorithms*. Prentice Hall PTR, 1994.

[11] Franck Rebillard. Du traitement de l'information à son retraitement. *Lavoisier*, 2006.

[12] Franck Rebillard. L'information journalistique sur l'internet, entre diffusion mass-médiatique et circulation réticulaire de l'actualité. *HAL*, 2007.

[13] Franck Rebillard. Les intermédiaires de l'information en ligne. *INA, la revue des médias*, 2020.

[14] W. E Winkler. The state of record linkage and current research problems. *Statistics of Income Division, Internal Revenue Service Publication*, 1999.

7 Annexes et Résultats

ÉTUDE 1 - Premier décès d'un médecin du coronavirus en France

NOM DE LA DÉPÊCHE DE L'AFP	NOM DE L'ARTICLE	LIEN	JOURNAL	DATE	JACCARD	LENVENSHTEIN	SIMILAR TEXT	JARO WINKLER	% TOT. SUBSTRINGS	TAILLE TOT. SUBSTRINGS
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Ce que l'on sait sur le premier médecin mort en France à cause du coronavirus	Clic	L'Express	22/03/2020	38,5%	17,9%	21,2%	80,7%	32,6%	659
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus : premier décès d'un médecin hospitalier en France	Clic	Le Figaro	22/03/2020	49,2%	47,7%	51,1%	85,7%	95,9%	1010
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus : après la mort du premier médecin, « je suis prête à tous les sacrifices mais pas à jouer avec ma vie »	Clic	Le Monde	23/03/2020	16,1%	8,1%	8,7%	77,6%	1,7%	35
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus : premier décès d'un médecin hospitalier en France	Clic	France 24	22/03/2020	84,9%	82,6%	85,8%	94,9%	93,2%	1881

Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus : premier décès d'un médecin hospitalier en France	Clic	Paris Match	22/03/2020	81,6%	88,2%	93,5%	90,7%	98,2%	1980
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus : premier décès d'un médecin en France	Clic	France Info	22/03/2020	19,3%	10,5%	19,3%	87,2%	2,6%	29
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus : Un médecin hospitalier décède dans l'Oise, le premier annoncé en France	Clic	20 minutes	22/03/2020	42,5%	40,4%	66,7%	96,2%	44,8%	903
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus : le médecin mort était "un confrère qui a donné sa vie pour sauver celle des autres"	Clic	Europe 1	22/03/2020	29,6%	14,5%	29,2%	87,1%	25,1%	419
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus: la mort d'un médecin accroît l'inquiétude des soignants	Clic	Le Point	22/03/2020	36,8%	24,7%	34,1%	81,3%	41,6%	837
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus : contaminé, un médecin hospitalier décède à Lille	Clic	La voix du nord	22/03/2020	46,2%	44,5%	66,3%	93,3%	43,9%	888

Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Médecin de Compiègne mort du coronavirus : “C'était un urgentiste exceptionnel”	Clic	France 3	22/03/2020	19,1%	8,1%	15,1%	89,6%	1,2%	25
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus: Olivier Véran annonce la mort d'un médecin hospitalier, une première en France	Clic	BFM TV	22/03/2020	25,6%	5,5%	27,5%	90,3%	3,8%	71
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Qui était le docteur Razafindranazy , premier médecin mort du coronavirus en France ?	Clic	Gala	22/03/2020	20,2%	8,3%	19,1%	89,1%	1,9%	38
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Médecin décédé du coronavirus : l’ inquiétude des soignants s’ accroît, "on est tous bouleversés"	Clic	Sud Ouest	22/03/2020	26,8%	13,5%	26,9%	84,0%	20,3%	409
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus : 1er décès d’un médecin hospitalier annoncé en France	Clic	L'Obs	22/03/2020	72,0%	78,6%	91,8%	92,8%	83,7%	1690

Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Premier décès d'un médecin hospitalier infecté par le nouveau coronavirus annoncé en France	Clic	Nice Matin	22/03/2020	98,5%	96,9%	97,9%	93,2%	97,8%	1973
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus. Premier décès d'un médecin hospitalier annoncé en France	Clic	Paris Normandie	22/03/2020	95,9%	96,6%	88,0%	95,5%	95,8%	1930
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus : premier décès d'un médecin hospitalier annoncé en France	Clic	LCI	22/03/2020	72,1%	74,9%	60,8%	89,3%	68,9%	1386
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Un premier médecin décède du coronavirus, annonce Olivier Véran	Clic	The Huffington Post	22/03/2020	54,2%	51,8%	35,7%	91,8%	14,7%	296
Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Coronavirus: premier décès d'un médecin hospitalier annoncé en France	Clic	Corse Matin	22/03/2020	38,6%	31,1%	30,6%	69,8%	100,0%	617
				Moyennes	48,38%	42,21%	48,47%	88,01%	48,39%	854
				Écart-type	26,75%	32,97%	30,12%	6,66%	39,21%	720
				Médiane	40,58%	35,76%	34,90%	89,45%	42,75%	748
				Maximum	98,5%	96,9%	97,9%	96,2%	100,0%	1980
				Minimum	16,1%	5,5%	8,7%	69,8%	1,2%	25

ÉTUDE 2 - Donald Trump suspend la contribution américaine à l'OMS

NOM DE LA DÉPÊCHE DE L'AFP	NOM DE L'ARTICLE	LIEN	JOURNAL	DATE	JACCARD	LENVENSHTEIN	SIMILAR TEXT	JARO WINKLER	% TOT. SUBSTRINGS	TAILLE TOT. SUBSTRINGS
Donald Trump suspend la contribution américaine à l'OMS	Trump suspend la contribution américaine à l'OMS	Clic	Le Figaro	15/04/2020	53,3%	36,8%	31,8%	82,2%	42,0%	1302
Donald Trump suspend la contribution américaine à l'OMS	Coronavirus : Donald Trump suspend la contribution américaine à l' OMS	Clic	Le Monde	15/04/2020	43,1%	32,5%	17,1%	81,6%	14,4%	393
Donald Trump suspend la contribution américaine à l'OMS	Donald Trump suspend la contribution américaine à l' OMS	Clic	La Voix Du Nord	15/04/2020	52,8%	36,9%	30,9%	76,8%	51,0%	1085
Donald Trump suspend la contribution américaine à l'OMS	En pleine pandémie, Donald Trump suspend la contribution américaine à l'OMS	Clic	Paris Match	15/04/2020	41,0%	27,0%	1,6%	93,5%	3,5%	158
Donald Trump suspend la contribution américaine à l'OMS	Coronavirus : Donald Trump suspend la contribution américaine à l'OMS	Clic	L'Express	15/04/2020	44,1%	38,9%	32,9%	90,8%	30,5%	1307

Donald Trump suspend la contribution américaine à l'OMS	OMS : Donald Trump met sa menace à exécution et suspend la contribution américaine	Clic	La Croix	15/04/2020	51,7%	30,0%	39,7%	82,8%	45,4%	1281
Donald Trump suspend la contribution américaine à l'OMS	En pleine pandémie, Donald Trump suspend la contribution américaine à l'OMS	Clic	LCI	15/04/2020	49,8%	32,6%	31,5%	84,6%	30,6%	992
Donald Trump suspend la contribution américaine à l'OMS	Coronavirus : Donald Trump suspend le financement américain de l'OMS	Clic	20 minutes	15/04/2020	60,1%	39,5%	51,8%	86,6%	57,1%	1915
Donald Trump suspend la contribution américaine à l'OMS	Coronavirus : Donald Trump suspend la contribution américaine à l'OMS	Clic	Europe 1	15/04/2020	85,7%	64,5%	79,3%	95,8%	77,2%	3429
Donald Trump suspend la contribution américaine à l'OMS	Coronavirus : Trump suspend la contribution américaine à l'OMS	Clic	Le Point	15/04/2020	42,0%	38,0%	26,6%	91,0%	24,6%	856
Donald Trump suspend la contribution américaine à l'OMS	Coronavirus : Donald Trump suspend la contribution américaine à l'OMS	Clic	Les Echos	15/04/2020	43,9%	31,0%	26,5%	91,6%	16,2%	566

Donald Trump suspend la contribution américaine à l'OMS	Covid-19 : Donald Trump suspend la contribution des États-Unis à l'OMS	Clic	France 24	15/04/2020	64,4%	38,8%	52,4%	95,5%	53,3%	2371
Donald Trump suspend la contribution américaine à l'OMS	Donald Trump suspend la contribution américaine à l' OMS	Clic	Libération	15/04/0202	47,0%	22,1%	14,6%	72,3%	15,1%	213
Donald Trump suspend la contribution américaine à l'OMS	États-Unis. Déjà 26 000 morts du coronavirus, mais Trump suspend la contribution américaine à l' OMS	Clic	Ouest France	15/04/2020	42,1%	35,0%	26,7%	94,0%	25,1%	1065
Donald Trump suspend la contribution américaine à l'OMS	Etats-Unis : Donald Trump suspend la contribution américaine à l'OMS	Clic	Midi Libre	15/04/2020	44,2%	20,7%	10,6%	67,6%	14,1%	169
Donald Trump suspend la contribution américaine à l'OMS	Coronavirus : Donald Trump suspend la contribution américaine à l'OMS	Clic	La Provence	15/04/2020	64,0%	38,7%	31,6%	84,4%	64,1%	1327
Donald Trump suspend la contribution américaine à l'OMS	Coronavirus: Donald Trump suspend la contribution américaine à l' OMS	Clic	Courrier Picard	15/04/2020	54,5%	26,4%	29,4%	84,8%	54,4%	980

Donald Trump suspend la contribution américaine à l'OMS	Trump suspend la contribution américaine à l'OMS	Clic	Sciences et Avenir	15/04/2020	65,1%	38,7%	31,4%	84,2%	64,8%	1325
Donald Trump suspend la contribution américaine à l'OMS	Trump suspend les versements américains à l'OMS pour sa "mauvaise gestion" du coronavirus	Clic	The Huffington Post	15/04/2020	43,3%	32,7%	30,4%	96,0%	22,7%	1007
Donald Trump suspend la contribution américaine à l'OMS	Covid-19. Trump coupe les vivres à l' OMS, l' accusant d'être responsable de la crise	Clic	Courrier International	15/04/2020	34,6%	30,1%	17,1%	89,8%	4,7%	175
Donald Trump suspend la contribution américaine à l'OMS	Trump suspend la contribution à l'OMS	Clic	La Dépêche du midi	15/04/2020	49,1%	36,1%	27,8%	85,5%	48,9%	962
Donald Trump suspend la contribution américaine à l'OMS	Coronavirus: Trump suspend la contribution américaine à l'OMS pour "mauvaise gestion"	Clic	BFM TV	15/04/2020	57,6%	35,4%	33,6%	80,3%	47,4%	1324
				Moyennes	51,52%	34,65%	30,70%	85,99%	36,69%	1 100
				Écart-type	11,36%	8,60%	15,91%	7,55%	21,08%	764
				Médiane	49,45%	35,20%	30,65%	85,15%	36,30%	1 036
				Maximum	85,7%	64,5%	79,3%	96,0%	77,2%	3429
				Minimum	34,6%	20,7%	1,6%	67,6%	3,5%	158

ÉTUDE 3 - Retraits massifs dans les banques grecques en 2012

NOM DE LA DÉPÊCHE DE L'AFP	NOM DE L'ARTICLE	LIEN	JOURNAL	DATE	JACCARD	LENVENSHTEIN	SIMILAR TEXT	JARO WINKLER	% TOT. SUBSTRINGS	TAILLE TOT. SUBSTRINGS
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	700 millions d'euros retirés des banques grecques lundi	Clic	Les Echos	15/05/2012	96,1%	88,4%	89,5%	91,7%	96,4%	1540
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Grèce: 700M€ retirés des banques hier	Clic	Le Figaro	15/05/2012	94,7%	94,2%	96,3%	94,1%	92,7%	1473
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Grèce : 700M euros retirés des banques lundi	Clic	Europe 1	15/05/2012	86,0%	56,0%	56,0%	86,0%	95,1%	854
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Retraits massifs dans les banques grecques	Clic	BFM TV	16/05/2012	41,9%	29,5%	25,3%	86,8%	2,4%	39
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Les Grecs, qui redoutent le pire, ont déjà retiré 700 millions d'euros des banques	Clic	La Depeche	17/05/2012	52,3%	37,2%	43,7%	87,7%	38,3%	524

Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Vent de panique : les Grecs vident leurs comptes en banque	Clic	RTL	16/05/2012	51,6%	28,8%	22,9%	81,7%	2,4%	39
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Retraits massifs dans les banques grecques	Clic	Challenges	16/05/2012	49,1%	30,5%	19,5%	89,6%	2,4%	39
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Retraits massifs dans les banques grecques	Clic	20 Minutes	16/05/12	40,5%	27,3%	26,7%	88,8%	2,4%	39
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Un juge à la tête de la Grèce	Clic	Le Point	16/05/2012	53,7%	22,9%	10,6%	78,6%	2,6%	41
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	La hantise d'une panique bancaire pèse sur la zone euro	Clic	La Croix	21/05/2012	44,3%	19,2%	7,6%	83,4%	1,9%	31
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Retraits massifs dans les banques grecques	Clic	L'Express	16/05/2012	48,1%	30,5%	19,3%	89,4%	2,4%	39

Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Retraits massifs dans les banques grecques	Clic	Le Monde	16/05/2012	67,0%	53,0%	60,9%	94,3%	51,2%	817
				Moyennes	60,44%	43,13%	39,86%	87,68%	32,52%	456
				Écart-type	20,48%	25,04%	29,81%	4,76%	40,87%	582
				Médiane	51,95%	30,50%	26,00%	88,25%	2,50%	40
				Maximum	96,1%	94,2%	96,3%	94,3%	96,4%	1540
				Minimum	40,5%	19,2%	7,6%	78,6%	1,9%	31

NOM DE LA DÉPÊCHE DE REUTERS	NOM DE L'ARTICLE	LIEN	JOURNAL	DATE	JACCARD	LENVENSHTEIN	SIMILAR TEXT	JARO WINKLER	% TOT. SUBSTRINGS	TAILLE TOT. SUBSTRINGS
Retraits massifs dans les banques grecques	700 millions d'euros retirés des banques grecques lundi	Clic	Les Echos	15/05/2020	44,4%	31,6%	26,4%	81,0%	2,2%	39
Retraits massifs dans les banques grecques	Grèce: 700M € retirés des banques hier	Clic	Le Figaro	15/05/2012	44,3%	31,1%	19,1%	89,3%	2,5%	39
Retraits massifs dans les banques grecques	Grèce : 700M euros retirés des banques lundi	Clic	Europe 1	15/05/2012	44,4%	23,7%	12,3%	82,7%	3,1%	28
Retraits massifs dans les banques grecques	Retraits massifs dans les banques grecques	Clic	BFM TV	16/05/2012	73,0%	51,0%	59,0%	91,4%	67,9%	1344

Retraits massifs dans les banques grecques	Les Grecs, qui redoutent le pire, ont déjà retiré 700 millions d'euros des banques	Clic	La Dépêche	17/05/2012	50,4%	29,2%	6,9%	79,5%	2,8%	38
Retraits massifs dans les banques grecques	Vent de panique : les Grecs vident leurs comptes en banque	Clic	RTL	16/05/2012	97,6%	74,5%	85,5%	92,7%	90,2%	2313
Retraits massifs dans les banques grecques	Retraits massifs dans les banques grecques	Clic	Challenges	16/05/2020	85,2%	98,0%	98,0%	94,0%	95,4%	2429
Retraits massifs dans les banques grecques	Retraits massifs dans les banques grecques	Clic	20 Minutes	16/05/2012	74,5%	50,9%	57,1%	81,5%	74,0%	1339
Retraits massifs dans les banques grecques	Un juge à la tête de la Grèce	Clic	Le Point	16/05/2012	71,3%	28,4%	29,7%	83,6%	53,7%	1379
Retraits massifs dans les banques grecques	Sauver la Grèce, contre vents et marées. La hantise d'une panique bancaire pèse sur la zone euro	Clic	La Croix	21/05/2012	48,5%	26,2%	12,9%	78,8%	1,5%	38
Retraits massifs dans les banques grecques	Retraits massifs dans les banques grecques	Clic	L'Express	16/05/2012	96,4%	98,8%	99,3%	89,0%	99,5%	2545

Retraits massifs dans les banques grecques	Retraits massifs dans les banques grecques	Clic	Le Monde	16/05/2012	48,3%	31,6%	30,7%	81,5%	20,8%	364
				Moyennes	64,86%	47,92%	44,74%	85,42%	42,80%	991
				Écart-type	20,65%	27,68%	34,04%	5,48%	41,11%	1026
				Médiane	60,85%	31,60%	30,20%	83,15%	37,25%	852
				Maximum	97,6%	98,8%	99,3%	94,0%	99,5%	2545
				Minimum	44,3%	23,7%	6,9%	78,8%	1,5%	28

NOM DE LA DÉPÊCHE	NOM DE L'ARTICLE	LIEN	JOURNAL	DATE	JACCARD	LENVENSHTEIN	SIMILAR TEXT	JARO WINKLER	% TOT. SUBSTRINGS	TAILLE TOT. SUBSTRINGS
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	700 millions d'euros retirés des banques grecques lundi	Clic	Les Echos	15/05/2012	96,1%	88,4%	89,5%	91,7%	96,4%	1540
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Grèce: 700M€ retirés des banques hier	Clic	Le Figaro	15/05/2012	94,7%	94,2%	96,3%	94,1%	92,7%	1473
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Grèce : 700M euros retirés des banques lundi	Clic	Europe 1	15/05/2012	86,0%	56,0%	56,0%	86,0%	95,1%	854
Retraits massifs dans les banques grecques	Retraits massifs dans les banques grecques	Clic	BFM TV	16/05/2012	73,0%	51,0%	59,0%	91,4%	67,9%	1344

Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Les Grecs, qui redoutent le pire, ont déjà retiré 700 millions d'euros des banques	Clic	La Depeche	17/05/2012	52,3%	37,2%	43,7%	87,7%	38,3%	524
Retraits massifs dans les banques grecques	Vent de panique : les Grecs vident leurs comptes en banque	Clic	RTL	16/05/2012	97,6%	74,5%	85,5%	92,7%	90,2%	2313
Retraits massifs dans les banques grecques	Retraits massifs dans les banques grecques	Clic	Challenges	16/05/2020	85,2%	98,0%	98,0%	94,0%	95,4%	2429
Retraits massifs dans les banques grecques	Retraits massifs dans les banques grecques	Clic	20 Minutes	16/05/2012	74,5%	50,9%	57,1%	81,5%	74,0%	1339
Retraits massifs dans les banques grecques	Un juge à la tête de la Grèce	Clic	Le Point	16/05/2012	71,3%	28,4%	29,7%	83,6%	53,7%	1379
Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	La hantise d'une panique bancaire pèse sur la zone euro	Clic	La Croix	21/05/2012	44,3%	19,2%	7,6%	83,4%	1,9%	31
Retraits massifs dans les banques grecques	Retraits massifs dans les banques grecques	Clic	L'Express	16/05/2012	96,4%	98,8%	99,3%	89,0%	99,5%	2545

Grèce: 700 millions d'euros retirés des banques grecques lundi (président)	Retraits massifs dans les banques grecques	Clic	Le Monde	16/05/2012	67,0%	53,0%	60,9%	94,3%	51,2%	817
				Moyennes	78,20%	62,47%	65,22%	89,12%	71,36%	1 382
				Écart-type	17,69%	27,76%	29,32%	4,60%	30,19%	769
				Médiane	79,85%	54,50%	59,95%	90,20%	82,10%	1 362
				Maximum	97,6%	98,8%	99,3%	94,3%	99,5%	2545
				Minimum	44,3%	19,2%	7,6%	81,5%	1,9%	31

ÉTUDE 4 - Mort de George Floyd

NOM DE LA DÉPÊCHE DE L'AFP	NOM DE L'ARTICLE	LIEN	JOURNAL	DATE	JACCARD	LENVENSHTEIN	SIMILAR TEXT	JARO WINKLER	% TOT. SUBSTRINGS	TAILLE TOT. SUBSTRINGS
Indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation	Indignation aux États-Unis après la mort d'un homme noir lors de son interpellation	Clic	Le Figaro	26/05/2020	64,0%	41,4%	51,2%	63,9%	88,7%	2256
Indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation (+)	États-Unis : indignation après la mort d'un homme noir lors d'une interpellation	Clic	L'Express	27/05/2020	61,0%	22,5%	29,5%	65,8%	91,6%	4033
Indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation	États-Unis. Indignation après la mort d'un homme Noir lors de son interpellation à Minneapolis	Clic	Ouest France	26/05/2020	81,6%	85,3%	91,6%	69,3%	89,7%	2353
Indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation	États-Unis : indignation après la mort d'un homme noir étouffé par un policier	Clic	RTL	27/05/2020	83,7%	58,0%	69,7%	67,1%	94,9%	2409

Indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation	Aux États-Unis, la mort d'un homme noir lors d'une interpellation indigne	Clic	The Huffington Post	26/05/2020	82,4%	83,5%	85,0%	65,2%	94,2%	2466
Indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation	Indignation aux Etats-Unis après la mort d'un homme noir lors de son interpellation	Clic	La Provence	27/05/2020	88,9%	79,4%	80,7%	87,4%	99,7%	2256
Indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation	Vague d'indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation	Clic	L'Est Républicain	27/05/2020	41,3%	20,9%	26,5%	58,6%	62,8%	1594
Indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation (+)	Etats-Unis : Quatre policiers limogés après la mort d'un homme noir lors de son interpellation	Clic	20 Minutes	27/05/2020	58,9%	50,3%	62,3%	66,8%	83,5%	3869
Etats-Unis: quatre policiers limogés après la mort d'un Noir lors de son interpellation	Quatre policiers renvoyés après la mort d'un homme noir au cours d'une interpellation aux Etats-Unis	Clic	Le Nouvel Obs	27/05/2020	78,5%	81,9%	91,6%	68,1%	89,4%	3751

Etats-Unis: quatre policiers limogés après la mort d'un Noir lors de son interpellation	Aux Etats-Unis, quatre policiers limogés après la mort d'un homme noir lors de son interpellation	Clic	Libération	27/05/2020	71,7%	64,6%	71,1%	67,2%	91,2%	3819
Indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation	Vague d'indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation	Clic	Le journal de Saône-et-Loire	27/05/2020	41,1%	20,7%	26,4%	58,7%	62,8%	1593
Indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation (+)	Etats-Unis : un homme noir, George Floyd, meurt étouffé lors d'une interpellation par quatre policiers blancs	Clic	Sud Ouest	27/05/2020	55,6%	52,0%	56,2%	63,6%	85,8%	3559
Indignation aux Etats-Unis après la mort d'un Noir lors de son interpellation (+)	Comment la mort de George Floyd lors de son arrestation par la police embrase les Etats-Unis	Clic	France Info	29/05/2020	44,7%	13,5%	21,2%	69,1%	21,5%	1440
Etats-Unis: quatre policiers limogés après la mort d'un Noir lors de son interpellation	Quatre policiers limogés après la mort d'un homme noir lors de son interpellation aux Etats-Unis	Clic	Le Monde	27/05/2020	75,2%	76,0%	88,8%	67,5%	85,9%	3591

Etats-Unis: quatre policiers limogés après la mort d'un Noir lors de son interpellation	Etats-Unis : quatre policiers limogés après la mort d'un homme noir lors de son interpellation	Clic	Le Parisien	27/05/2020	29,1%	16,3%	21,6%	52,6%	31,6%	486
Etats-Unis: quatre policiers limogés après la mort d'un Noir lors de son interpellation	Etats-Unis : quatre policiers de Minneapolis licenciés après la mort d'un homme noir lors d'une interpellation violente	Clic	LCI	27/05/2020	41,9%	17,3%	23,1%	62,6%	24,4%	662
				Moyennes	62,48%	48,98%	56,03%	65,84%	74,86%	2 509
				Écart-type	18,69%	27,40%	27,65%	7,29%	26,38%	1160
				Médiane	62,50%	51,15%	59,25%	66,30%	87,30%	2 381
				Maximum	88,9%	85,3%	91,6%	87,4%	99,7%	4033
				Minimum	29,1%	13,5%	21,2%	52,6%	21,5%	486

ÉTUDE 5 - Allocution d'Emmanuel Macron du 15 juin 2020

NOM DE LA DÉPÊCHE DE L'AFP	NOM DE L'ARTICLE	LIEN	JOURNAL	DATE	JACCARD	LENVENSHTEIN	SIMILAR TEXT	JARO WINKLER	% TOT. SUBSTRINGS	TAILLE TOT. SUBSTRINGS
Le satisfecit de Macron sur la gestion de crise fait bondir l'opposition (+)	Allocution présidentielle : Macron reste confiné dans ses certitudes	Clic	L'Humanité	15/06/2020	14,8%	6,8%	5,2%	47,7%	1,3%	32
Le satisfecit de Macron sur la gestion de crise fait bondir l'opposition (+)	« L'Etat a tenu » : Emmanuel Macron s'offre un satisfecit sur sa gestion de crise	Clic	Le Monde	15/06/2020	38,0%	8,0%	20,9%	64,0%	10,0%	941
Le satisfecit de Macron sur la gestion de crise fait bondir l'opposition (+)	Un discours pour rien, "bavardage gluant" : les réactions politiques après le discours de Macron	Clic	La Dépêche	15/06/2020	16,6%	7,1%	9,2%	48,6%	8,4%	222
Le satisfecit de Macron sur la gestion de crise fait bondir l'opposition (+)	Déconfinement : Emmanuel Macron loue sa gestion de la crise lors d'une opération de communication	Clic	20 Minutes	15/06/2020	20,8%	8,0%	12,7%	52,9%	1,0%	41

Le satisfecit de Macron sur la gestion de crise fait bondir l'opposition (+)	Dans son allocution, Macron n'a pas répondu à toutes les interrogations	Clic	The Huffington Post	15/06/2020	28,8%	7,6%	8,8%	59,6%	1,3%	102
Le satisfecit de Macron sur la gestion de crise fait bondir l'opposition (+)	Allocution de Macron : l'opposition déplore un «discours pour rien» et un «satisfecit malvenu»	Clic	Le Figaro	15/05/2020	20,0%	7,4%	8,8%	51,3%	2,5%	85
Le satisfecit de Macron sur la gestion de crise fait bondir l'opposition (+)	France.Dans son allocution, Macron “tourne la page du coronavirus” et lance “l’acte III de son quinquennat”	Clic	Courrier International	15/06/2020	27,9%	9,9%	16,9%	54,6%	21,0%	919
Le satisfecit de Macron sur la gestion de crise fait bondir l'opposition (+)	Dordogne : les réactions des Périgourdins à l'allocution d'Emmanuel Macron	Clic	France Bleu	15/06/2020	14,6%	6,5%	6,7%	48,2%	1,0%	27
Le satisfecit de Macron sur la gestion de crise fait bondir l'opposition (+)	Zone verte partout dès lundi et écoles totalement rouvertes le 22 juin	Clic	Le Courrier Picard	15/06/2020	17,4%	7,9%	11,4%	48,2%	10,1%	306

Le satisfecit de Macron sur la gestion de crise fait bondir l'opposition (+)	Déconfinement : les principales réactions au discours d'Emmanuel Macron	Clic	CNews	14/06/2020	13,2%	7,0%	8,1%	43,2%	17,7%	349
Le satisfecit de Macron sur la gestion de crise fait bondir l'opposition (+)	Allocution d'Emmanuel Macron : "déni surréaliste", "auto-satisfecit", "discours un peu light", les réactions très critiques de l'opposition	Clic	France Info	14/06/2020	23,9%	8,3%	12,5%	56,3%	3,5%	159
Le satisfecit de Macron sur la gestion de crise fait bondir l'opposition (+)	Déconfinement . Emmanuel Macron veut accélérer la reprise du travail : « Je compte sur vous ! »	Clic	Ouest France	14/05/2020	18,1%	7,2%	9,1%	50,6%	0,6%	22
				Moyennes	21,18%	7,64%	10,86%	52,10%	6,53%	267
				Écart-type	7,31%	0,90%	4,42%	5,77%	7,00%	328
				Médiane	19,05%	7,50%	9,15%	50,95%	3,00%	131
				Maximum	38,0%	9,9%	20,9%	64,0%	21,0%	941
				Minimum	13,2%	6,5%	5,2%	43,2%	0,6%	22

ÉTUDE 6 - Alunissage chinois sur la face cachée de la lune

NOM DE LA DÉPÊCHE DE L'AFP	NOM DE L'ARTICLE	LIEN	JOURNAL	DATE	JACCARD	LENVENSHTEIN	SIMILAR TEXT	JARO WINKLER	% TOT. SUBSTRINGS	TAILLE TOT. SUBSTRINGS
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine réussit le premier alunissage sur la face cachée de la Lune	Clic	Le Monde	03/01/2019	32,1%	10,7%	15,4%	58,6%	4,4%	181
La Chine, première à poser un engin sur la face cachée de la Lune	Face cachée de la Lune: Qu'est-ce que mijote la Chine exactement?	Clic	20 Minutes	03/01/2019	29,1%	9,4%	13,5%	58,1%	4,1%	172
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine s'est posée sur la face cachée de la Lune, une première mondiale	Clic	Le Figaro	03/01/2019	49,3%	19,0%	47,2%	66,0%	33,3%	1271
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine, première à poser un engin sur la face cachée de la Lune	Clic	Sciences et Avenir	03/01/2019	97,8%	97,8%	98,0%	96,1%	100,0%	4022
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine réussit le premier alunissage jamais réalisé sur la face cachée de la Lune	Clic	Ouest France	03/01/2019	72,1%	67,2%	83,6%	77,4%	73,8%	2928

La Chine, première à poser un engin sur la face cachée de la Lune	La Chine devient le premier pays à poser un engin sur la face cachée de la Lune	Clic	Paris Match	03/01/2019	70,0%	45,5%	69,3%	69,0%	72,8%	2801
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine, première à poser un engin sur la face cachée de la Lune	Clic	La Voix Du Nord	03/01/2019	17,4%	17,5%	16,5%	68,7%	100,0%	662
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine, première à poser un engin sur la face cachée de la Lune	Clic	Le Nouvel Obs	03/01/2019	100,0%	100,0%	100,0%	100,0%	100,0%	4022
La Chine, première à poser un engin sur la face cachée de la Lune	Face cachée de la Lune: la Chine première venue	Clic	Libération	03/01/2019	37,0%	24,5%	29,6%	58,7%	52,9%	936
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine, premier pays à poser un engin spatial sur la face cachée de la Lune	Clic	France Bleu	03/01/2019	38,2%	29,2%	36,6%	58,5%	58,4%	1136
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine pose un engin spatial sur la face cachée de la Lune, une première	Clic	France 24	03/01/2019	38,3%	21,2%	31,0%	59,6%	50,0%	955

La Chine, première à poser un engin sur la face cachée de la Lune	La Chine, premier pays à se poser sur la face cachée de la Lune	Clic	Les Echos	03/01/2019	53,2%	37,8%	54,2%	65,2%	57,0%	1671
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine réussit le premier alunissage jamais réalisé sur la face cachée de la Lune	Clic	France Info	03/01/2019	54,4%	29,8%	50,4%	67,1%	41,5%	1355
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine, premier pays à poser un engin sur la face cachée de la Lune	Clic	Le Courrier Picard	03/02/2019	67,3%	46,5%	69,3%	68,4%	72,6%	2693
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine dépose Chang'e 4 sur la face cachée de la Lune	Clic	L'Express	03/01/2019	36,2%	13,5%	25,5%	58,9%	14,5%	589
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine pose une sonde spatiale sur la face cachée de la Lune	Clic	La Nouvelle République	03/01/2019	35,2%	12,0%	21,3%	61,2%	14,6%	386
La Chine, première à poser un engin sur la face cachée de la Lune	Une sonde chinoise alunit sur la face cachée de la Lune	Clic	Le Point	03/01/2019	54,7%	35,3%	56,6%	65,7%	51,8%	1672

La Chine, première à poser un engin sur la face cachée de la Lune	Une première : la Chine réussit à alunir sur la face cachée de la Lune	Clic	Le Progrès	03/01/2019	36,1%	26,2%	30,1%	58,2%	55,5%	948
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine en tête de la conquête spatiale après son exploit sur la Lune	Clic	Courrier International	03/01/2019	25,3%	10,0%	13,1%	53,7%	7,6%	119
La Chine, première à poser un engin sur la face cachée de la Lune	Les premières images de la face cachée de la Lune envoyées par la Chine	Clic	The Huffington Post	03/01/2020	22,4%	11,7%	18,6%	49,0%	44,3%	504
La Chine, première à poser un engin sur la face cachée de la Lune	Une première : la Chine réussit à alunir sur la face cachée de la Lune	Clic	L'Alsace	03/01/2019	24,1%	15,6%	20,8%	52,4%	52,0%	602
La Chine, première à poser un engin sur la face cachée de la Lune	Les Chinois, premiers à atterrir sur la face cachée de la lune	Clic	France Inter	03/01/2019	54,7%	36,6%	56,1%	67,3%	50,5%	1659
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine réussit l'alunissage d'un engin sur la face cachée de la Lune, une première mondiale	Clic	Nice Matin	03/01/2019	58,0%	44,5%	56,6%	67,3%	60,9%	1816

La Chine, première à poser un engin sur la face cachée de la Lune	La Chine, premier pays à alunir sur la face cachée de la Lune	Clic	Sud Ouest	03/01/2019	38,7%	29,9%	38,3%	62,8%	59,7%	1231
La Chine, première à poser un engin sur la face cachée de la Lune	La Chine démontre sa puissance spatiale en se posant sur la face cachée de la Lune	Clic	France Culture	03/01/2019	18,4%	7,4%	7,1%	51,8%	6,2%	254
La Chine, première à poser un engin sur la face cachée de la Lune	Sonde chinoise sur la face cachée de la Lune : et l'Homme, quand y reposera-t-il le pied ?	Clic	LCI	03/01/2019	27,0%	9,6%	7,0%	57,6%	2,7%	113
				Moyennes	45,65%	31,09%	40,99%	64,51%	47,73%	1 335
				Écart-type	22,08%	24,79%	26,96%	11,78%	29,92%	1131
				Médiane	38,25%	25,35%	33,80%	62,00%	51,90%	1 046
				Maximum	100,0%	100,0%	100,0%	100,0%	100,0%	4022
				Minimum	17,4%	7,4%	7,0%	49,0%	2,7%	113

ÉTUDE TÉMOIN

NOM DE L'ARTICLE AFP	NOM DE L'ARTICLE	LIEN	JOURNAL	DATE	JACCARD	LENVENSHTEIN	SIMILAR TEXT	JARO WINKLER	% TOT. SUBSTRINGS	TAILLE TOT. SUBSTRINGS
A la chasse aux truffes	A la chasse aux truffes	Clic	Europe 1	13/11/2016	100,0%	100,0%	100,0%	100,0%	100,0%	3581
A la chasse aux truffes	Le silence de l' ONU face aux bombardements russes en Syrie	Clic	Le Monde	10/04/2020	8,1%	8,4%	13,7%	94,6%	0,5%	17
A la chasse aux truffes	Stefan Zweig ou l'horreur de la politique	Clic	Le Monde Diplomatique	09/04/2020	10,3%	6,8%	16,6%	92,9%	0,6%	19
A la chasse aux truffes	Civilization VI : Une adaptation solide sur PS4 et Xbox One	Clic	Jeuxvideo.com	25/11/2019	9,0%	5,8%	8,1%	75,4%	0,5%	18
A la chasse aux truffes	Emma Roberts : "Vieillir est une belle chose qui ne me fait pas peur"	Clic	Marie Claire	30/03/2020	12,6%	6,1%	8,9%	91,7%	0,7%	19
A la chasse aux truffes	Ségolène Royal dit avoir compris ses erreurs	Clic	France Inter	26/06/2011	9,5%	6,1%	10,1%	90,8%	0,5%	16
A la chasse aux truffes	J'accuse... !	Clic	L'Aurore	13/01/1898	7,0%	3,9%	3,7%	71,2%	0,4%	15
A la chasse aux truffes	Le SUV à hydrogène qui purifie l'air Hyundai NEXO	Clic	Auto Moto	27/03/2020	10,7%	6,4%	6,5%	89,5%	0,4%	15

A la chasse aux truffes	L'art de faire des pâtes fraîches	Clic	Marmiton	12/02/2019	12,4%	7,8%	12,9%	87,0%	0,4%	16
A la chasse aux truffes	Arthur Chevallier – Napoléon, le plus grand des confinés	Clic	Le Point	13/04/2020	9,4%	7,5%	7,6%	82,5%	0,4%	13
A la chasse aux truffes	Le Cac 40 recule de plus de 1%, l'espoir de déconfinement s'estompe	Clic	Les Echos	08/04/2020	9,5%	6,6%	13,7%	94,6%	0,3%	11
A la chasse aux truffes	Film noir	Clic	Wikipédia	29/03/2020	8,6%	7,5%	7,0%	83,0%	0,4%	16
A la chasse aux truffes	"Un peuple prêt à sacrifier un peu de liberté pour un peu de sécurité..." : Benjamin Franklin a-t-il vraiment dit ça?	Clic	Les Inrocks	19/11/2015	8,9%	5,9%	14,6%	89,8%	0,3%	11
A la chasse aux truffes	La mission lunaire américaine a son budget	Clic	Science&vie	13/04/2020	7,4%	5,7%	9,2%	79,4%	0,3%	11
A la chasse aux truffes	Les déplacements autorisés pour l'adoption d'animaux domestiques	Clic	AuFéminin	12/04/2020	9,1%	8,8%	18,1%	85,6%	0,4%	11
A la chasse aux truffes	Doug Sanders est mort	Clic	L'Équipe	13/04/2020	10,2%	6,6%	11,5%	78,9%	0,9%	15

A la chasse aux truffes	Le recueil comme genre	Clic	Fabula	24/04/2001	7,5%	6,5%	7,6%	76,3%	0,5%	17
				Moyennes	9,39%	6,65%	10,61%	85,20%	0,47%	15
				Écart-type	1,61%	1,18%	4,00%	7,35%	0,16%	3
				Médiane	9,24%	6,56%	9,65%	86,30%	0,40%	16
				Maximum	12,6%	8,8%	18,1%	94,6%	0,9%	19
				Minimum	7,0%	3,9%	3,7%	71,2%	0,3%	11

AFP
Comparator

Documentation

Noé Faure

Juin 2020

Introduction

https://noefaure.github.io/AFP-comparator/index.html

L'application est disponible via le lien ci-dessus.
Elle est hébergée par GitHub et ouverte à la contribution.

Page d'accueil

1 : Il est possible de télécharger le mémoire et la documentation de l'application directement sur le site.

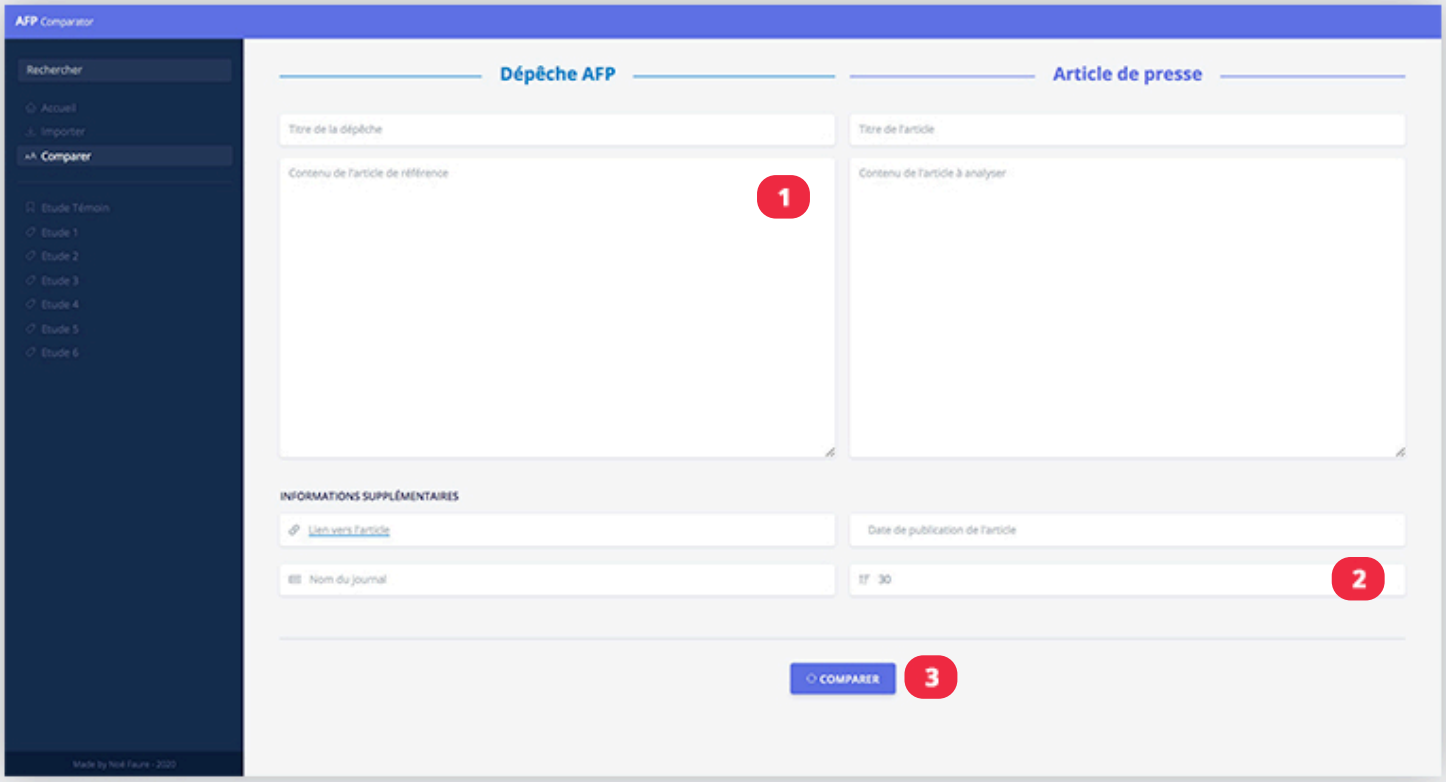
2 : La barre de recherche peut permettre de rechercher une étude en particulier si leur nombre venait à augmenter.

3 : Vous pouvez revenir à la page d'accueil à tout moment en cliquant sur ce lien.

4 : L'essentiel de l'application est disponible ici. Ce lien redirige vers la page qui permet de comparer des dépêches. Le lien « importer » permet l'affichage de fichiers au format TSV mais pas encore son enregistrement en base.

5 : L'étude témoin comporte un agrégat hétérogène d'articles, il permet de mesurer l'efficacité des indices de similarité utilisés. Consulter « Étude Témoin » dans le mémoire.

6 : La liste des différentes études est disponible ici.



Comparaison de textes

1 : Les champs sont assez explicites. Remplir à gauche le texte de référence, à droite le texte à comparer. Tous les champs sont facultatifs mais il nécessaire de remplir au minimum les deux champs de contenu pour espérer obtenir un résultat exploitable.

2 : Ce champ permet de varier le seuil de détection de l'indice « %Tot. substring ». Par défaut il est établi à 30. C'est à dire que les chaînes de caractères de plus de 30 symboles identiques entre les deux textes seront détectées. Elles apparaitront surlignées. Consulter le chapitre en question dans le mémoire pour plus d'information.

3 : Cliquer pour lancer la comparaison.

NOM DE LA DÉPÊCHE	NOM DE L'ARTICLE	LIEN	JOURNAL	DATE	JACCARD	LEVENSHTEIN	SIMILAR TEXT	JARO WINKLER	% TOT. SUBSTRINGS	TAILLE TOT. SUBSTRINGS
Violences policières : nouveaux rassemblements en France prévus ce 9 juin	Violences policières: Phillippe échange à Evry avec policiers et habitants	Clic	Le Monde	09/06/2020	30.3%	8.5%	16.3%	57.4%	19.5%	730

COPIER DANS LE PRESSE-PAPIER TÉLÉCHARGER AU FORMAT TSV

Les diagrammes circulaires en haut de la page illustrent les différents indices utilisés dans la comparaison. Les phrases surlignées en jaune représentent les passages en commun (dépassant le seuil choisi au préalable, par défaut 30 caractères) entre les deux textes. Le nombre sur fond violet indique le nombre de caractères présents dans ce passage. À sa droite le pourcentage représente la proportion du passage par rapport à l'ensemble du texte. En bas, l'ensemble des informations sont résumées dans un tableau. Il est possible de copier le tableau dans le presse-papier en cliquant sur le bouton « copier dans le presse-papier ». Il peut par la suite être collé dans un tableur (format optimisé pour les tableurs Google Sheet).