Répétitions et variations des textes générés.

Une analyse linguistique basée sur un corpus d'articles financiers rédigés en français

Anna-Maria De Cesare

Technische Universität Dresden

The goal of this paper, devoted to automatically generated texts, is twofold. First, and more importantly, we aim at presenting a methodology allowing to describe and assess the 'quality' of these texts, as well as uncover some of their specificities compared to non-generated texts. Second, we want to identify a series of relevant features related to both the lexicon and the macro- and micro-structuring of these texts. Based on the application of our methodology to a small work corpus of 100 articles produced in French by the CAC40 software in the field of finance, we conclude that, taken individually, each generated text shows a sufficiently rich internal variation to be perceived as natural.

Keywords: automatically generated texts; descriptive methodology; French; lexicon; textual properties; information structure

1. Introduction

La génération automatique de textes, basée sur des systèmes auxquels on se réfère avec des termes comme 'data-to-text generation' (Van der Lee, Krahmer & Wudden 2018), connaît un essor important depuis au moins une quinzaine d'années, notamment parce qu'elle permet de nombreuses applications dans le domaine commercial et des médias (pour une reconstruction historique, cf. entre autres Ponton 1997 ; Dierickx 2019 et Thurman 2019). Aujourd'hui, les logiciels de rédaction qui s'appuient sur des textes dit « à trou », un système de règles d'écriture et une base de données structurées sont de plus en plus courants. Ces logiciels, que les médias appellent aussi « moteurs de rédaction » et « robots-rédacteurs », sont d'ailleurs en passe de révolutionner le mode de production et de diffusion des textes. Pour toutes ces raisons, il est indispensable de se pencher de plus près sur les textes générés par ces logiciels, pour en dégager les spécificités.

Dans cette contribution, nous nous intéressons aux textes générés en langue française, qui ont jusqu'ici suscité peu d'intérêt de la part des linguistes. La recherche sur les textes générés en français se concentre principalement sur leur

CHIMERA. Romance Corpora and Linguistic Studies 8 (2021), 79-108. ISSN 2386-2629
© 2021 The Author(s); distributed under the Creative Commons Attribution License

mode de production, décrit dans le cadre de la linguistique computationnelle (cf. les travaux pionniers de Danlos 1991, 2000). Les études relatives aux textes générés dans d'autres langues, en premier lieu en anglais, s'intéressent également à leur mode de réception et perception, notamment de la part des lecteurs et des journalistes (Dierickz 2020). Dans le cadre des sciences de la communication et des médias, les textes générés sont étudiés en les comparant à des textes similaires rédigés par des journalistes en chair et en os (cf. Jung *et al.* 2017; Graefe *et al.* 2018 ; Stalph, Thaesler-Kordonouri & Thurman 2021 ainsi que la bibliographie citée par Thurman 2019 : 185 et Dierickx 2021 : 13-15).

Une réflexion sur les propriétés linguistiques, pragmatiques et textuelles de l'écriture fonctionnelle (à savoir non littéraire[1]) générée en français est en revanche encore rare (ce constat vaut également pour les textes générés en anglais et allemand[2]). On trouve une série d'observations intéressantes dans la thèse de Master de Schevenels (2019-2020), qui analyse un petit échantillon de 30 résumés de matches de football amateur générés par la société LabSense sur le site internet du quotidien belge *SudPresse*. Les paramètres analysés par Schevenels sont multiples et relativement hétérogènes. Ils concernent la longueur des textes (mesurée en nombre de caractères), les unités du paratexte (notamment la forme et fonction du titre), la nature du lexique (en particulier le registre et la présence de phrasèmes), l'orthographe et la ponctuation. Ces paramètres fournissent de précieux indicateurs pour évaluer la 'qualité' des textes générés et comprendre ce qui les distinguent des textes rédigés par des journalistes.

Pour mieux cerner la nature des textes générés automatiquement (à partir de données structurées, d'un texte à trou et de règles[3]), ainsi que ce qui les distingue des textes écrits par des journalistes en chair et en os, il est toutefois fondamental d'élargir la recherche et d'analyser des paramètres pas ou peu considérés par Schevenels (2019-2020), comme les questions touchant aux propriétés textuelles

[1] Pour une description approfondie relative à la génération de récit, cf. Belen Baez (2018).

[2] Pour une analyse linguistique (basée, entre autres, sur la forme et fréquence des parties du discours) de textes générés en allemand dans le domaine du sport, cf. toutefois Meier-Vieracker (2021) (cf. aussi les observations, à caractère préliminaire, de Meier-Vieracker [2020]).

[3] Les textes générés par l'algorithme, à partir d'une séquence de mots ou d'une phrase (en anglais : d'un *prompt*), sont bien entendu complètement différents tant du point de vue de leur production que de leur application. Dans le domaine d'application relatif aux médias, les textes générés par le système GPT-3 sont rares, notamment parce qu'ils sont peu adaptés au journalisme d'investigation. Le fameux article intitulé « A robot wrote this entire article. Are you scared yet, human? » (publié par *The Guardian* le 08.09.2020) n'a en réalité pas été généré tel quel par l'algorithme. Les journalistes du quotidien britannique ont en effet sélectionné les meilleurs passages générés, qu'ils/elles ont ensuite assemblés pour créer l'article final.

et informationnelles de l'écriture générée par les logiciels de rédaction. C'est ce que nous proposons de faire dans la présente étude, dont le principal objectif est de nature descriptive. Notre intention – à plus long terme – est également de développer une méthodologie d'analyse adaptée pour décrire et expliquer les spécificités des textes générés.

Le présent travail s'articule en quatre parties. Dans un premier temps, nous décrivons le logiciel-rédacteur CAC40, ainsi que les propriétés du corpus de textes construit pour notre analyse, basé sur un échantillon de 100 commentaires générés par le logiciel à la clôture de la Bourse de Paris (§ 2). Nous présentons ensuite trois concepts descriptifs aptes à analyser les spécificités des textes générés – il s'agit des concepts que nous appelons *segments fixes, variables et variantes* – et les appliquons aux titres et aux corps des articles de notre corpus (§ 3). Après avoir décrit la macro- et microstructure des textes générés par le logiciel CAC40 (§ 4), nous dégageons comment se distribuent les segments fixes, les variables et les variantes dans les micro-unités textuelles des articles générés (§ 5). En guise de conclusion, nous proposons un bilan sur la qualité des textes générés par le logiciel-rédacteur CAC40 et présentons deux autres étapes à intégrer dans la recherche pour mieux comprendre les spécificités des textes générés (§ 6).

2. Base de données empiriques

2.1 Textes générés par le logiciel-rédacteur CAC40

La base de données sur laquelle s'appuie notre étude est constituée d'un échantillon de 100 textes générés par le « logiciel-rédacteur CAC40 » (aussi appelé « robot CAC40 » [4]) sur le site appia.dataecriture.fr[5], développé par la société demain.ai. Les textes sont générés comme « Démo » directement sur le site de la société. On les trouve en cliquant sur la phrase « J'y vais » pour « Découvrir des robots qui rédigent des contenus ». Il s'agit de textes à fonction principalement publicitaire, qui ont pour but de montrer à de potentiels clients ce qu'un tel logiciel est capable de produire et, au final, de les convaincre à utiliser la technologie dans leur domaine professionnel. Le terme *data-écriture* fait référence au fait que ce type d'écriture se base principalement sur du traitement de données.

[4] L'emploi du terme *robot* n'est pas approprié et génère confusion et inquiétude. Nous préférons donc utiliser le mot *logiciel* (pour une discussion plus approfondie, cf. Linden & Dierickx 2019).

[5] CAC40. https://appia.dataecriture.fr/category/cac40 (accédé le 6 octobre 2021).

Comme l'indique le nom du logiciel (à savoir « logiciel-rédacteur CAC40 »[6]), le cas d'usage proposé par la société demain.ai concerne le domaine boursier, en l'occurrence la rédaction automatisée d'articles sur la situation du marché parisien. Selon la société demain.ai, il s'agit du tout premier logiciel générateur de textes boursiers, qui ne se limite pas à décrire la journée écoulée mais est capable de commenter les cours et d'analyser la performance du CAC :

> Nous avons lancé récemment un robot commentateur boursier qui analyse chaque jour, en toute autonomie, la performance du CAC. Non seulement il commente les cours, mais également les volumes de transactions et les faits marquants de la séance. (Mégean, 9.9.2021)

Le logiciel-rédacteur CAC40 génère un nouvel article tous les jours ouvrés à 18h (aucun texte n'est généré le week-end, ni les jours fériés). Chaque semaine on trouve donc sur le site en question cinq nouveaux textes générés[7]. Ces textes sont disponibles sur le site de la société depuis mardi 11.05.2021. Les deux captures d'écran proposées dans les Figures 1 et 2 permettent de se faire une idée plus précise du type de texte que génère ce logiciel-rédacteur.

[6] Au niveau terminologique, « CAC signifie Cotation Assistée en Continu, cela veut dire que sa valeur varie en permanence tous les jours ouvrés de 9h00 à 17h30. Il est mis à jour toutes les 15 secondes. 40 parce qu'il se compose de 40 valeurs parmi les 100 premières capitalisations françaises. » (https://www.economie.gouv.fr/facileco/cac-40)

[7] On relève toutefois une anomalie : deux textes décrivant des résultats différents ont été générés le 25 mai 2021. Le premier s'intitule *Le CAC40 termine la séance en hausse modérée* ; le deuxième *Le CAC40 conclut la séance en baisse modérée*.

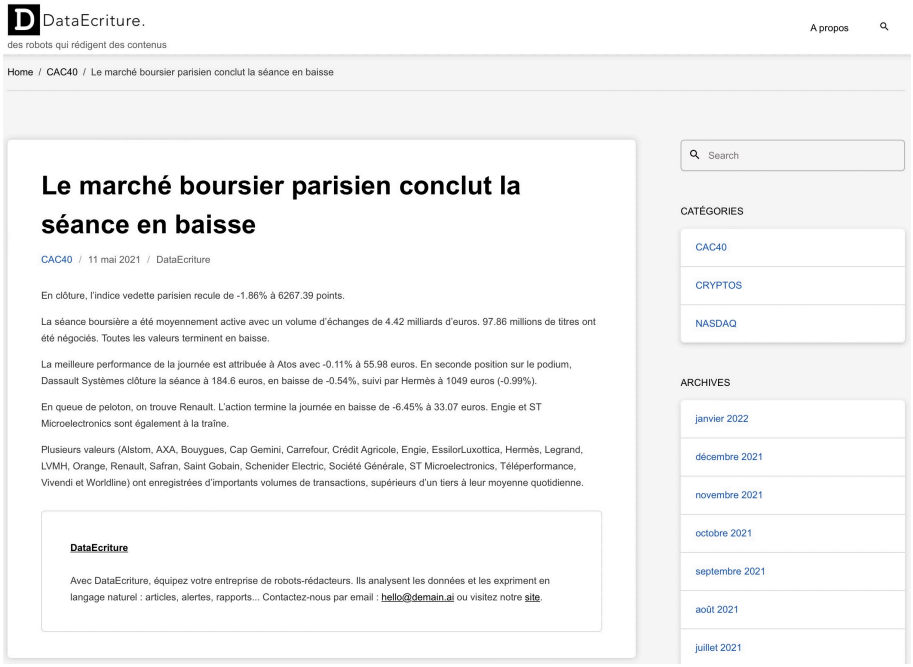

Figure 1. Premier texte généré automatiquement par le logiciel-rédacteur CAC40

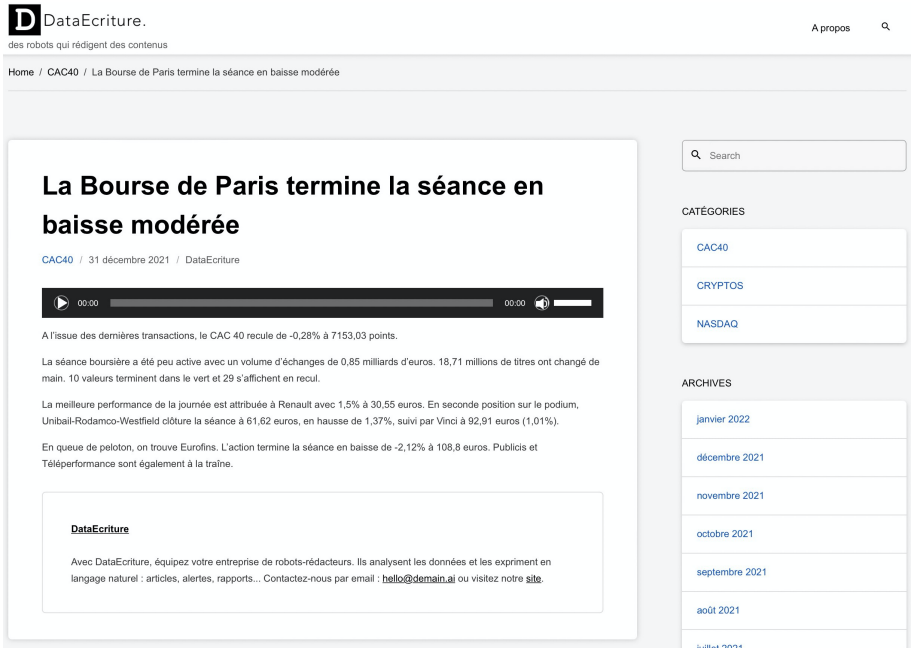

Figure 2. Dernier texte de 2021 généré automatiquement par le robot CAC40[8]

Comme on peut facilement le constater en observant les deux captures d'écran, les textes générés par le robot CAC40 se ressemblent beaucoup, notamment en ce qui concerne le choix du lexique (cf. par exemple la formulation du titre), leur format et structure générale. Il s'agit d'articles relativement courts, formés d'un titre et de quelques blocs de textes. Vu leur longueur et structure, ces textes peuvent être assimilés à la typologie des dépêches d'agence (pour une description, cf. De Cesare & Baranzini 2011), typologie à laquelle ils sont d'ailleurs reconduits par l'entreprise qui les produit :

> La DataEcriture s'effectue à partir de données qui, une fois collectées, sont interprétées par une série de calculs (règles). Ces données viennent alimenter notre plateforme partenaire de génération de contenu (Arria NLG) pour produire un texte (dépêche) et un post twitter (@DemainAi). (https://www.demain.ai/nos_publications/decouvrez-notre-robot-cac40-il-ecrit-tous-les-jours-et-en-plus-il-parle/; visité le 22.1.2022)

[8] Depuis le 6 juillet 2021, les textes sont également synthétisés automatiquement, ce qui permet de les écouter en mobilité.

2.2 Description du corpus de textes analysés : propriétés quantitatives

Le corpus que nous avons construit (le *Corpus CAC40*) contient un échantillon de 100 textes générés par le logiciel-rédacteur CAC40 sur le site de demain.ai du 11.05.2021 au 28.09.2021. Tous les textes sont librement accessibles dans les archives du site.

Une première remarque quantitative, qui s'impose dans le cadre de la linguistique des corpus, est que le Corpus CAC40 est très petit, surtout quand on le compare aux corpus de la dernière génération, qui contiennent des milliards de mots (cf., le « French Web 2017 », aussi dit « frTenTen17 », compte par exemple plus de 5,7 milliards de mots issus de plus de 14 millions de documents différents). Cela dit, quand on s'occupe de textes générés de manière automatique, à partir d'un texte à trou, le fait de travailler avec un échantillon de 100 textes ne constitue pas un défaut majeur parce que ces textes sont par nature relativement similaires. Comme on le verra de manière plus détaillée dans le § 3, les 100 textes décrivent en effet les mêmes événements (performance du CAC et faits marquant de la journée écoulée), dans le même ordre et en grande partie avec les mêmes mots. Les informations qui varient d'un texte à l'autre correspondent à leur tour généralement à des données chiffrées ou des noms de valeurs cotées en bourse.

Les 100 textes sélectionnés pour l'analyse ont ensuite été téléchargés sur la plateforme Sketch Engine (pour détails, cf. Kilgariff *et al.* 2014), ce qui permet de créer un corpus dans le sens technique du terme. La taille du corpus, mesurée sur la base de paramètres tels que le nombre de phrases, mots, lemmes et *token* (à savoir du nombre total d'occurrences de signes graphiques : mots, signes de ponctuation, abréviations etc.), est décrite dans le Tableau 1.

Tableau 1. Propriétés générales du Corpus CAC40

phrases	845
lemmes	156
formes uniques différentes (mots, signes de ponctuation etc.)	1'316
mots (nombre d'occurrences)	14'383
tokens	19'180

Le Corpus CAC40 est formé de 845 phrases (on a en moyenne 8.5 phrases par texte). Les textes contiennent en outre un nombre relativement exigu de lemmes (N = 156) par rapport au nombre d'occurrences des mots[9] qui le composent (N =

[9] Le concept de *mot* est ici à entendre dans un sens large (cf. les formes proposées dans le Tab. 2). En outre, l'exiguïté du nombre de lemmes qui compose le Corpus CAC40 (N = 156) devient

14'383). Ceci signifie que certaines formes se répètent très souvent dans le corpus. Pour identifier les mots qui se répètent le plus souvent, nous avons extrait tous les termes qui totalisent plus de 100 occ. (cf. Tableau 2, dans lequel les mots sont présentés par ordre de fréquence décroissante).

Tableau 2. Les « mots » avec N > 100 occ.

1.	de	956	**14.**	ont	168
2.	la	838	**15.**	l'	167
3.	à	817	**16.**	valeurs	167
4.	en	637	**17.**	hausse	158
5.	euros	527	**18.**	journée	156
6.	d'	371	**19.**	été	148
7.	séance	344	**20.**	baisse	142
8.	le	323	**21.**	termine	138
9.	et	284	**22.**	cac40	130
10.	avec	217	**23.**	boursière	118
11.	clôture	204	**24.**	transactions	112
12.	un	198	**25.**	milliards	109
13.	a	175			

Les mots les plus récurrents dans le Corpus CAC40 appartiennent à la classe des mots « vides » (dits aussi « fonctionnels »). Il s'agit de prépositions (*de/d', à, en, avec*), d'articles (*la, le, un, l'*), d'une conjonction de coordination (*et*), ainsi que de formes verbales à caractère grammatical (*a, ont, été*). Les mots qui occupent les quatre premiers rangs de la liste de fréquence (*de, la, à, en*) recouvrent 22,5% des occurrences totales du corpus (3'248/14'383). Ces données permettent d'identifier une première spécificité des textes générés. Si les mots fonctionnels sont fréquents dans tous les textes, et occupent donc les premiers rangs de toutes les listes de fréquence, ce qui est spécial dans le cas des textes générés (en l'occurrence de ceux qui nous occupent), c'est le fait que seul quatre mots couvrent presque un quart des occurrences du corpus. Une comparaison avec les quatre mots les plus fréquents du corpus frTenTen17 (à savoir *de, la, et, le* ; on notera que les deux premiers mots sont identiques à ceux du Corpus CAC40) permet

évidente quand on compare le résultat obtenu pour ce corpus avec celui du corpus frTenTen17 mentionné plus haut (N lemmes = 18'056'191).

d'observer qu'ils couvrent 13% du total (740'605'574/5'753'773'137), un chiffre deux fois moins grand que celui du Corpus CAC40[10].

Au sein du lexique qui revient fréquemment dans le Corpus CAC40, on trouve aussi un petit groupe de mots « pleins » (ou référentiels), typiques du domaine de la finance. Il s'agit surtout de termes appartenant à la classe des noms (cf. *euros, séance, valeurs, hausse, journée, baisse, transactions, milliards*), mais il y a aussi des adjectifs (*boursière*), des verbes (*termine*), ainsi que des mots qui entrent dans deux catégories lexicales différentes (c'est le cas de *clôture*, employé dans le corpus en fonction nominale et verbale).

Pour notre analyse, il est également pertinent d'identifier les formes (mots et *tokens*) qui ne se présentent qu'une seule fois dans tout le corpus (les *hapax legomena*). A ce propos, les données fournies par Sketch Engine permettent d'observer que la grande majorité des 1316 formes qui composent le Corpus CAC40 n'est employée qu'une seule fois. Plus précisément, on observe que, grosso modo à partir du rang de fréquence 200 (et ce jusqu'au dernier rang, à savoir 1316), on trouve presque exclusivement des données chiffrées : nombres entiers (1049 [euros]) et nombres décimaux (6267.39 [points], 4.42 [milliards d'euros], également sous forme de pourcentages : 0.19% / -4.28%). Les noms de valeurs (AXA, Carrefour, Alstom etc.) ne constituent par contre pas des hapax.

Une autre donnée quantitative intéressante à relever, qui reflète directement le fait d'être en présence de textes générés, basés sur des patrons fixes, est la présence de formes (mots ou autres) qui totalisent un nombre d'occurrences correspondant à un chiffre rond. Le Corpus CAC40 comprend par exemple 25 mots différents qui totalisent chacun 100 occ. (cf. Tableau 3) : on trouve donc chaque mot une fois par texte. Il s'agit tant de mots fonctionnels que de mots pleins (ces derniers sont toutefois plus nombreux).

Tableau 3. Mots avec N = 100 occ.

seconde – peloton – active – échanges – meilleure – points – attribuée – sur – traîne – podium– est – performance – millions – position – volume – titres – on – 2021 – suivi – queue – également – par – dataecriture – trouve – sont

Dans le corpus, on trouve en outre deux autres éléments qui totalisent un nombre d'occurrences correspondant à un chiffre rond. Il s'agit de signes de ponctuation : le point, présent 1000 fois (comme on le verra, il s'agit du signe qui clôt toutes les phrases), et la barre oblique, qui revient 200 fois (elle est présente deux fois dans chaque ligne d'information qui suit le titre).

[10] Dans le corpus frTenTen17, les lemmes *à* et *en* occupent respectivement les rangs 5 et 9.

3. Description des textes générés par le logiciel-rédacteur CAC40 : éléments fixes, variables et variantes

3.1 Nature des éléments qui composent les textes générés par le logiciel-rédacteur CAC40

Les textes générés automatiquement par le logiciel-rédacteur CAC40 peuvent être caractérisés – du moins dans un premier temps, parce qu'ils sont en réalité plus complexes – comme des « textes à trous » (on parle également de « formulaire ») :

> Un formulaire est un *texte partiellement rédigé* (un patron), et les parties manquantes (les *trous*) sont remplacées par des *variables* littérales liées à des champs d'une base de données. Cette méthode est couramment utilisée pour la production massive de textes très répétitifs et ne variant que sur des éléments locaux […]. (Danlos 2000; nôtre l'emploi de l'italique)

Les textes générés comportent des éléments fixes et des variables. Les *éléments fixes* peuvent être conçus comme des « briques de textes », qui se retrouvent inchangés dans tous les textes. Au niveau formel, il peut s'agir de segments plus ou moins longs et complexes d'un point de vue morphosyntaxique, composé d'un ou plusieurs mots, de collocations, de constituants syntaxiques ou autre. Les *variables* sont à leur tour les éléments qui remplissent les trous. Elles sont liées à des champs d'une base de données. Dans les textes générés par le logiciel-rédacteur CAC40, ces champs correspondent en général à des noms de valeurs cotées en bourse (*Arcelor Mittal / Airbus / Orange* etc.), au coût de la valeur (*44.85 euros / 27.02 euros / 124.54 euros*) et à la performance des valeurs à la clôture de la bourse de Paris (6279.35 points ; 2.29%). Dans la plupart des cas, comme nous l'avons vu dans le § 2.2, il s'agit de données chiffrées.

Les textes qui nous occupent se composent en réalité d'autres segments (pas identifiés dans la citation de Danlos 2000), qui ne constituent pas à proprement parler des variables. Il s'agit également de segments qui varient au niveau local, mais la variation n'intéresse pas ou peu le plan dénotatif : elle concerne plutôt la forme linguistique des expressions utilisées dans le texte. En termes saussuriens, ces segments varient au niveau du signifiant, mais pas ou minimalement à celui du signifié. Nous nous référerons à ces segments avec le terme *variantes* et distinguerons deux types de cas : les variantes *sémantiques* et les variantes *formelles*. Les segments qui varient au niveau *sémantique* sont par ex. les syntagmes nominaux du titre *l'indice vedette parisien / le CAC40 / l'indice phare de la Bourse de*

Paris : ils entretiennent un rapport de *synonymie* (on parlera donc dans ce cas de variantes synonymiques[11]). Pour ce qui est des segments qui varient au niveau formel, on distinguera les variantes morphologiques, morphosyntaxiques, syntaxiques etc. (comme par ex. *sa moyenne quotidienne* / *leur moyenne quotidienne*).

3.2 Identification des segments qui forment le titre des articles générés par le logiciel-rédacteur CAC40

Une analyse en segments (fixes, variables et variantes) peut être appliquée tout d'abord aux titres des articles générés par le logiciel-rédacteur CAC40. Pour illustrer la méthodologie employée, basée sur une comparaison des mêmes segments qui composent les textes du corpus, nous décrirons les titres qui chapeautent les dix premiers articles de notre corpus (cf. Tableau 4). Le titre correspond dans tous ces textes (et ceci vaut bien entendu également pour les 100 textes du corpus) à une phrase simple, basée sur l'ordre Sujet-Verbe-Objet-Adverbial des constituants.

Tableau 4. Analyse des segments qui forment les titres de 10 textes générés

Date	Sujet	Verbe	Object	Adverbial
11.5.21	Le marché boursier parisien	conclut	la séance	en baisse
12.5.21	La Bourse de Paris	termine	la séance	en hausse modérée
13.5.21	Le CAC40	clôture	la séance	en hausse modérée
14.5.21	Le marché boursier parisien	termine	la séance	en hausse
17.5.21	Le marché boursier parisien	clôture	la séance	en baisse modérée
18.5.21	Le marché boursier parisien	conclut	la séance	en baisse modérée
19.5.21	Le marché boursier parisien	termine	la séance	en baisse
20.5.21	La Bourse de Paris	termine	la séance	en hausse
21.5.21	La Bourse de Paris	termine	la séance	en hausse modérée
24.5.21	Le CAC40	termine	la séance	en hausse modérée

La nature des segments qui composent les quatre constituants syntaxiques des titres peut être aisément déterminée par une « lecture » verticale (ou paradigmatique) des formes employées. Ainsi, le sujet dénote toujours le même référent mais

[11] Au niveau terminologique, d'autres termes ont été proposés : François Portet distingue par exemple les variations « esthétiques » (qui correspondent à nos variantes sémantiques) des variations liées aux données (nos variables). Cf. Génération Automatique de Résumés d'Élections Régionales. https://lig-membres.imag.fr/portet/cours/idl/tp_elections.html (accès 7.1.2022).

avec des formes différentes : dans les dix titres analysés, trois syntagmes s'alternent (*Le marché boursier parisien / La Bourse de Paris / Le CAC40*), qui entretiennent un rapport de synonymie. Nous sommes donc en présence de variantes sémantiques. La même chose s'observe pour le verbe : trois variantes sémantiques, qui entrent dans un rapport de synonymie, s'alternent (*conclut / termine / clôture*). L'objet direct ne varie par contre jamais : il s'agit par conséquent d'un segment fixe.

L'adverbial final (réalisé sous forme de syntagme prépositionnel) présente un cas particulier de variable, qui dénote la manière dont s'est déroulée l'action exprimée par le prédicat (formé par le verbe et l'objet direct). Dans les 10 titres analysés, quatre formulations s'alternent (*en baisse / en baisse modérée / en hausse / en hausse modérée*) ; dans les autres titres du corpus, on trouve deux adverbiaux supplémentaires (*en très forte baisse / inchangé*). Dans ce cas, la variable correspond à une 'direction de variation' et la différence entre, par exemple, *baisse – baisse modérée – très forte baisse* est déterminée par des intervalles numériques précis, de sorte que lorsqu'on atteint un certain seuil la baisse est considérée comme 'modérée' ou 'très forte'. On est donc en présence d'une variable (non explicitement chiffrée) dépendante de la performance de la journée boursière qui vient de s'écouler. Ce segment est clairement différent des syntagmes qui forment le sujet et le verbe du titre : au niveau de l'adverbial la variation n'est en effet pas libre.

Globalement, on constate que le titre des articles générés comporte trois segments variables et un segment fixe. Les deux premiers segments du titre (le sujet et le verbe) constituent des variantes sémantiques libres, tandis que le segment final (l'adverbial) constitue une variable au sens propre, qui s'appuie sur des données chiffrées (bien que, dans ce cas, elles ne soient pas citées). L'emploi de variantes synonymiques permet de générer des titres moins répétitifs d'un texte à l'autre, notamment en ce qui concerne le lexique. Sur les dix titres analysés, seuls deux d'entre eux sont identiques en tout point (il s'agit des titres générés le 12.5.2021 et 21.5.2021), ce qui est en fait relativement peu probable. En effet, si on étend l'analyse en segments aux titres de tout le Corpus CAC40, force est de constater que le logiciel est également programmé pour gérer la fréquence des segments qui s'alternent au sein des paradigmes. Dans l'intégralité du corpus, les éléments d'un même paradigme (sujet et verbe) sont employés avec une fréquence quasi égale :

- sujet : *Le CAC40* (36 occ.) / *La Bourse de Paris* (35 occ.) / *Le marché boursier parisien* (29 occ.)
- verbe : *clôture* (35 occ.) / *conclut* (29 occ.) / *termine* (26 occ.)

La présence de trois segments variables au sein d'une structure somme toute relativement peu complexe (formée de quatre segments) est quelque peu surprenante. On peut donc se demander pourquoi le logiciel-rédacteur a été programmé ainsi. Pourrait-on attribuer à ce choix une fonction persuasive ? Il est vrai que, à lui seul, le titre illustre à merveille la flexibilité et précision de la technologie que la société à but lucratif (demain.ai) cherche à vendre.

3.3 Identification des segments qui forment le corps des articles générés par le logiciel-rédacteur CAC40

Suivant la même méthodologie que celle appliquée au titre (dans le § 3.2), une comparaison des 10 premiers articles de notre corpus permet de reconstruire de manière relativement précise le formulaire à la base des textes générés par le logiciel-rédacteur CAC40, en particulier de repérer les segments fixes, les variables et les variantes. Dans le texte du point (1), généré le 2.6.2021, nous indiquons les variables en gras et les variantes sur fond gris (sans distinction entre les variantes sémantiques et les variantes formelles) ; les segments (adverbial ou verbe) qui dénotent le degré de performance de la Bourse de Paris, et constituent comme nous avons vu des variables particulières, sont soulignés ; les segments fixes correspondent au reste du texte, qui n'est pas mis en relief au niveau graphique.

(1) Le CAC40 termine la séance en hausse modérée
CAC40 / **2 juin 2021** / DataEcriture
A l'issue des dernières transactions, l'indice phare de la Bourse de Paris progresse de **0.49**% à **6521.52** points.

La séance boursière a été moyennement active avec un volume d'échanges de 2.75 milliards d'euros. **54.14** millions de titres ont changé de main. **27** valeurs terminent dans le vert et **13** s'affichent en recul.

La meilleure performance de la journée est attribuée à **Total** avec **2.24**% à **39.69** euros. En seconde position sur le podium, **Airbus** clôture la séance à **110.7** euros, en hausse de **1.78**%, suivi par **Renault** à **34.58** euros (**1.65**%).

En queue de peloton, on trouve **Vivendi**. Le titre termine la journée en baisse de **-1.35**% à **29.21** euros. **Worldline** et **Téléperformance** sont également à la traîne.

L'action Alstom a enregistrée d'importants volumes de transactions, supérieurs d'un tiers à sa moyenne quotidienne.

EssilorLuxottica se distingue avec un plus haut de 52 semaines à **143.66** euros et une valorisation boursière grimpant à **62,947.97** M€.

Sur les 174 formes (mots, chiffres, abréviations) qui composent l'article analysé (y compris le titre), 110 correspondent à des segments fixes, 36 constituent des variables (27 chiffrées et 9 non chiffrées) et 28 des variantes. Ces données permettent donc d'observer deux choses importantes : plus de la moitié des textes générés par le logiciel-rédacteur CAC40 est fixe (63%) ; en même temps, une portion non négligeable de ces textes est flexible et change en fonction de différents paramètres (notamment sémantiques et grammaticaux). Au sein des parties non fixes du texte, on compte 21% de variables[12] et 16% de variantes.

Le prochain pas de notre analyse consiste à déterminer de manière plus précise les propriétés textuelles et informationnelles des segments fixes, des variables et des variantes. Pour ce faire, nous tiendrons compte de la distribution des segments au sein des unités du texte, notamment des micro-unités que nous appelons *unités informationnelles* (pour une définition, cf. le § 4.3). Avant de déterminer dans quelles unités textuelles se distribuent les différents segments (§ 5), il faut toutefois encore identifier la structure des textes générés par le logiciel-rédacteur CAC40 (§ 4). Dans ce qui suit, l'analyse concerne uniquement le corps du texte.

4. La structure des textes générés par le logiciel-rédacteur CAC40

Les textes générés par le logiciel-rédacteur CAC40 peuvent être segmentés à différents niveaux : à un premier niveau, le texte est découpé en macro-unités qui correspondent à des blocs (§ 4.1) ; les blocs s'articulent à leur tour en énoncés (§ 4.2) et les énoncés en micro-unités qui correspondent aux *unités informationnelles* (§ 4.3).

[12] À noter que le pourcentage de variables fluctue quelque peu d'un texte à l'autre : celui associé au texte reproduit au point (1) est relativement bas. Certains textes comptent en effet jusqu'à un tiers de variables. Il s'agit notamment de ceux qui contiennent une parenthèse dans ce qui est l'avant-dernier bloc du texte reproduit en (1). Cette parenthèse, remplie par des variables, est parfois très longue (dans le texte généré le 11 mai 2021, on dénombre par exemple 23 noms de valeurs ; cf. ex. (12)).

4.1 Segmentation des textes en blocs

Le corps des textes générés par le logiciel-rédacteur CAC40 se compose de blocs, à savoir d'unités textuelles graphiquement autonomes, séparées les unes des autres par un espace vide (sur le concept, cf. De Cesare *et al*. 2016 : 125-126). Le nombre de blocs varie dans les textes du corpus. La plupart des textes se compose de 5 blocs (71/100), mais on trouve plusieurs textes qui en contiennent 4 ou 6 (respectivement, 15 et 14 sur 100). Le Tableau 5 propose la segmentation en blocs du texte généré le 2.6.2021, ainsi que les macro-thèmes développés par chacun d'entre eux.

Tableau 5. Segmentation en blocs des textes du Corpus CAC40

Blocs	Texte généré le 2.6.2021	Macro-thèmes
1	A l'issue des dernières transactions, l'indice phare de la Bourse de Paris progresse de 0.49% à 6521.52 points.	Performance globale de la Bourse de Paris à sa fermeture, décrite par rapport au jour précédent.
2	La séance boursière a été moyennement active avec un volume d'échanges de 2.75 milliards d'euros. 54.14 millions de titres ont changé de main. 27 valeurs terminent dans le vert et 13 s'affichent en recul.	Evaluation du degré d'activité de la séance boursière. Des précisions sont fournies sur différents aspects (nombre de titres échangés, valeurs en hausse et en baisse).
3	La meilleure performance de la journée est attribuée à Total avec 2.24% à 39.69 euros. En seconde position sur le podium, Airbus clôture la séance à 110.7 euros, en hausse de 1.78%, suivi par Renault à 34.58 euros (1.65%).	Identification des trois meilleures valeurs individuelles de la journée.
4	En queue de peloton, on trouve Vivendi. Le titre termine la journée en baisse de -1.35% à 29.21 euros. Worldline et Téléperformance sont également à la traîne.	Identification des trois plus mauvaises valeurs individuelles de la journée.
5	L'action Alstom a enregistrée d'importants volumes de transactions, supérieurs d'un tiers à sa moyenne quotidienne.	Faits marquants de la séance, en termes de volume de transactions.

6	EssilorLuxottica se distingue avec un plus haut de 52 semaines à 143.66 euros et une valorisation boursière grimpant à 62,947.97 M€.	Valeur qui se distingue par un record à la hausse (dans un cas à la baisse[13]).

Comme nous l'avons dit plus haut, les articles générés par le logiciel-rédacteur CAC40 ne se composent pas tous du même nombre de blocs. La présence / absence de certains blocs de texte est déterminée par la présence / absence d'une certaine donnée structurée liée à la journée en question. On constate cependant que tous les textes présentent au moins les quatre premiers blocs. Les deux derniers blocs du texte, qui fournissent des indications sur les valeurs qui ont enregistré une performance spéciale (presque toujours à la hausse), ne sont en revanche pas toujours présents. Quand une journée n'est marquée par aucune performance spéciale, les deux derniers blocs sont absents (on a donc des articles à 4 blocs). Dans la plupart des articles, la valeur qui fait défaut est celle qui marque un record particulier, décrite dans le Bloc 6. Ceci explique pourquoi la majorité des textes du corpus se compose de 5 blocs. Dans le corpus, on relève également des textes à 5 blocs dans lesquels il manque le Bloc 5 et donc dans lesquels le dernier bloc coïncide avec le Bloc 6 des textes qui se composent de tous les blocs.

4.2 Segmentation des blocs de textes en énoncés

Les blocs de texte se composent à leur tour d'énoncés (abrégés : E), à savoir d'unités qui, à l'écrit, ont principalement une fonction de composition textuelle (elles s'agencent autour de relations logico-argumentatives et thématiques ; cf. Ferrari 2014 : 32-37). La segmentation en énoncés des blocs qui forment le texte généré le 2.6.2021 est proposée dans le Tableau 6 (par convention, la frontière des énoncés est indiquée par une double barre oblique).

Tableau 6. Segmentation des blocs de textes en énoncés

Blocs	**Texte généré le 2.6.2021**
1	A l'issue des dernières transactions, l'indice phare de la Bourse de Paris progresse de 0.49% à 6521.52 points. //**E1**
2	La séance boursière a été moyennement active avec un volume d'échanges de 2.75 milliards d'euros. //**E2** 54.14 millions de titres ont changé de main. //**E3** 27 valeurs terminent dans le vert et 13 s'affichent en recul. //**E4**

[13] Le corpus présente un seul cas de ce type (« Cloturant [sic] à 39,42 euros (-2,83%), l'action touche son plus bas annuel »).

3	La meilleure performance de la journée est attribuée à Total avec 2.24% à 39.69 euros. //**E5** En seconde position sur le podium, Airbus clôture la séance à 110.7 euros, en hausse de 1.78%, suivi par Renault à 34.58 euros (1.65%). //**E6**
4	En queue de peloton, on trouve Vivendi. //**E7** Le titre termine la journée en baisse de -1.35% à 29.21 euros. //**E8** Worldline et Téléperformance sont également à la traîne. //**E9**
5	L'action Alstom a enregistrée d'importants volumes de transactions, supérieurs d'un tiers à sa moyenne quotidienne. //**E10**
6	EssilorLuxottica se distingue avec un plus haut de 52 semaines à 143.66 euros et une valorisation boursière grimpant à 62,947.97 M€. //**E11**

Tous les blocs de textes générés par le logiciel-rédacteur CAC40 peuvent être segmentés en énoncés comme celui que nous avons analysé dans le Tableau 6. Dans les textes qui comportent tous les blocs (soit six) on relève 11 énoncés. La plupart des textes du corpus se composant de 5 blocs, il résulte toutefois que le texte 'type' inclue au total 10 énoncés. La segmentation des blocs de textes en énoncés permet également de constater que ces blocs sont généralement courts : dans les textes à 6 blocs, la moitié d'entre eux se compose d'un seul énoncé (blocs 1, 5, 6) ; on dénombre aussi un bloc à deux (blocs 3) et deux blocs à trois énoncés (blocs 2 et 4).

Les textes générés par le logiciel-rédacteur CAC40 incluent également des incises délimitées par des parenthèses : une première incise, toujours présente, se situe à la fin de E6 ; une deuxième incise, présente seulement dans certains textes, se trouve au début de E10. Les incises constituent des énoncés indépendants à l'intérieur d'autres énoncés (à noter que nous avons décidé de ne pas les numéroter).

4.3 Segmentation des énoncés en Unités informationnelles

Les énoncés peuvent être à leur tour segmentés en unités informationnelles (dorénavant UIs), définies comme les plus petites unités fonctionnelles qui structurent l'énoncé et le texte ; ces unités sont autonomes et hiérarchisées (Ferrari 2014 : 37ss). Dans notre modèle théorique de référence, dit « de Bâle » (cf. Ferrari *et al*. 2008 ; Ferrari 2014), on distingue trois types de UIs : le Noyau, le Cadre et l'Appendice. Une définition de ces trois UIs est proposée dans le Tableau 7.

Tableau 7. Les Unités Informationnelles (UIs) structurant les énoncés (Ferrari 2014 : 38-43).

Type de UIs	Fonction	Propriétés
Noyau	Le Noyau définit la fonction textuelle et la force illocutoire (à l'écrit, il s'agit généralement d'assertion) de l'E dans son ensemble. Il constitue le premier plan de l'E.	Le Noyau est une UI nécessaire (un E contient au moins un Noyau) et suffisante (un E peut être formé seulement du Noyau). Un E formé du seul Noyau est simple d'un point de vue informationnel ; un E qui comporte le Noyau et au moins une autre UI (Cadre ou Appendice) est en revanche complexe.
Cadre	Le Cadre constitue le champ d'application du Noyau. Il fournit, à l'arrière-plan de l'E, des indications sur les coordonnées spatio-temporelles dans lesquelles se déroule le contenu nucléaire, mais présente aussi les liens logico-argumentatifs qui lient le Noyau au cotexte ainsi que des évaluations et des jugements du locuteur.	Le Cadre est une UI optionnelle, qui ouvre l'E. Sa portée est généralement large et s'étend aux E du cotexte, notamment à l'E qui suit celui qui réalise le Cadre. Un E peut inclure plus d'une UI de Cadre. Le Cadre garantit la continuité et cohérence du texte à différents niveaux (notamment logico-argumentatif et thématique).
Appendice	L'Appendice sert à transmettre des informations que le locuteur décide de véhiculer en arrière-plan du message. Il permet de préciser, d'enrichir, de clarifier ou de moduler la ligne centrale du message, correspondant aux informations véhiculées dans les Noyaux.	L'Appendice est une UI optionnelle, qui suit ou interrompt l'UI à laquelle il (à savoir l'Appendice) se rattache (Cadre, Noyau ou autre Appendice). L'Appendice a une portée locale, limitée à l'E dont il fait partie. En raison de sa portée locale, l'Appendice n'a pas d'impact direct sur la cohérence textuelle.

La segmentation en UIs des énoncés du texte généré par le logiciel-rédacteur CAC40 le 2.6.2021 est proposée dans le Tableau 8. Par convention, les UIs qui forment les énoncés sont délimitées par une barre oblique simple et leur nom est fourni en exposant après la barre.

Tableau 8. Segmentation des énoncés en unités informationnelles (Cadre, Noyau, Appendice)

Blocs	Texte généré le 2.6.2021
1	A l'issue des dernières transactions, /[Cadre] l'indice phare de la Bourse de Paris progresse de 0.49% à 6521.52 points. /[Noyau] //**E1**
2	La séance boursière a été moyennement active avec un volume d'échanges de 2.75 milliards d'euros. /[Noyau] //**E2** 54.14 millions de titres ont changé de main. /[Noyau] //**E3** 27 valeurs terminent dans le vert et 13 s'affichent en recul. /[Noyau] //**E4**
3	La meilleure performance de la journée est attribuée à Total avec 2.24% à 39.69 euros. /[Noyau] //**E5** En seconde position sur le podium, /[Cadre] Airbus clôture la séance à 110.7 euros, /[Noyau] en hausse de 1.78%, /[Appendice] suivi par Renault à 34.58 euros (1.65%). /[Appendice] //**E6**
4	En queue de peloton, /[Cadre] on trouve Vivendi. /[Noyau] //**E7** Le titre termine la journée en baisse de -1.35% à 29.21 euros. /[Noyau] //**E8** Worldline et Téléperformance sont également à la traîne. /[Noyau] //**E9**
5	L'action Alstom a enregistrée d'importants volumes de transactions, /[Noyau] supérieurs d'un tiers à sa moyenne quotidienne. /[Appendice] //**E10**
6	EssilorLuxottica se distingue avec un plus haut de 52 semaines à 143.66 euros et une valorisation boursière grimpant à 62,947.97 M€. /[Noyau] //**E11**

Tous les textes générés par le logiciel-rédacteur CAC40 peuvent être segmentés comme celui dans le Tableau 8. Les 11 énoncés des textes qui se composent de six blocs présentent 11 Noyaux, 3 UIs de Cadre et 3 UIs d'Appendice, toujours placées en position finale d'énoncé. Si la plupart des textes du corpus comporte des énoncés simples, formés d'une seule UI correspondant au Noyau (7/11), on relève également un groupe d'énoncés complexes, articulés en plus d'une unité informationnelle (4/11). Les énoncés complexes réalisent en outre des schémas informationnels variés : on relève deux énoncés articulés en 'Cadre / Noyau' (E1, E7), un énoncé articulé en 'Noyau / Appendice' (E10) et même un énoncé composé de quatre UIs, réalisant le schéma informationnel complexe 'Cadre / Noyau / Appendice / Appendice' (E6[14]).

La segmentation des énoncés en unités fonctionnelles plus petites permet de relever que les énoncés générés automatiquement par le logiciel-rédacteur CAC40 présentent une structuration informationnelle variée et relativement fine. Il s'agit

[14] On pourrait envisager une autre interprétation de E6, dans laquelle le dernier segment (« suivi par […] ») constitue la suite d'un Noyau interrompu par une information d'arrière-plan correspondant à une UI d'Appendice (« en hausse de 1.78% »). Étant donné que l'information présente dans le Cadre (*En seconde position sur le podium*) se réfère uniquement à la première valeur (à savoir *Airbus*), nous retenons pour notre analyse l'articulation proposée dans le Tab. 8.

toutefois d'une structuration planifiée à l'avance, qui ne varie jamais d'un texte à l'autre. On trouve en effet toujours le même nombre de UIs par énoncé, ces UIs sont toujours de la même nature et sont présentées dans le même ordre. Une conclusion importante que l'on peut tirer de ce constat est que l'articulation informationnelle des énoncés fait partie intégrante du patron des textes générés par le logiciel-rédacteur CAC40[15].

5. La distribution des segments au sein des micro-unités qui forment les textes générés par le logiciel-rédacteur CAC40

Considérons maintenant la distribution des segments (fixes, variables et variantes) dans les UIs (Noyau, Cadre et Appendice) qui composent les énoncés des textes générés par le logiciel-rédacteur CAC40. Compte tenu des différences fonctionnelles entre les trois UIs (cf. les définitions présentées dans le Tableau 7), nous nous attendons à ce que les segments ne se distribuent pas au hasard dans ces UIs. Nos hypothèses sont plus particulièrement les suivantes :

(i) les segments fixes peuvent se réaliser dans tous les types de UI, à savoir les Noyaux, les Cadres et les Appendices ;

(ii) les variables se réalisent en premier lieu dans les Noyaux, qui contiennent le contenu situé au premier plan de l'énoncé ; les UIs d'Appendices, qui spécifient et élaborent les informations présentées dans le reste de l'énoncé, pourraient elles aussi inclure des variables, mais celles-ci se situeraient à l'arrière-plan de l'énoncé et auraient donc un statut informationnel secondaire par rapport aux variables présentées dans le Noyau ;

(iii) les variantes peuvent en principe, elles aussi, occuper tous les types de UI.

5.1 Nature des segments qui composent le Noyau des énoncés

Toutes les UIs de Noyau se composent de plusieurs segments de nature différente : elles comportent toujours au moins un segment fixe et une variable. Il s'agit donc clairement d'une UI remplie par une structure à trou. Dans le cas de figure le plus simple, à un premier segment fixe (qui n'est pas coloré dans les exemples

[15] Cette conclusion n'est pas valable pour tous les textes générés (cf. De Cesare, Eliasson & Weidensdorfer en prép.).

ci-dessous) suit une variable (en gras). Dans le Noyau de E7, la variable coïncide avec une valeur boursière :

(2) / on trouve **Renault** /[Noyau]. // **E7**

Ce cas de figure est toutefois présent une seule fois dans les textes du corpus (dans E7 justement). Dans les autres cas, on observe des configurations plus complexes, dans lesquelles plus de deux segments différents sont juxtaposés au sein d'un même Noyau. C'est notamment le cas des Noyaux de E1 à E4 :

(3) / l'indice vedette parisien recule de -**1.86**% à **6267.39** points /[Noyau]. // **E1**

(4) / La séance boursière a été moyennement active avec un volume d'échanges de **4.42** milliards d'euros /[Noyau]. // **E2**

(5) **97.86** millions de titres ont été négociés /[Noyau]. // **E3**

(6) / **Toutes les** valeurs terminent en baisse /[Noyau]. // **E4**

Dans E1, on trouve deux variables (en gras) sous forme de données chiffrées (-1.86 et 6267.39), ainsi qu'un segment (le sujet *l'indice vedette parisien*) qui varie uniquement au niveau de sa formulation ; il s'agit plus particulièrement d'une variante sémantique de nature synonymique. Le verbe (*recule*) est également constitué d'un segment variable, qui s'adapte selon la performance de la journée (pour plus de détails, cf. Eliasson en prép.). Dans E2, on retrouve une variable sous forme de donnée chiffrée (4.42), le reste est fixe, à l'exception du degré d'activité de la bourse, qui constitue un type particulier de variable (non chiffrée). On relève dans ce cas une modalisation de l'adjectif prédicatif *active* par l'adverbe de degré *moyennement* (dans les textes du corpus, le paradigme des modalisateurs comporte trois formes, qui dénotent les degrés bas, modéré et haut : *peu / moyennement / très* ; sur la classe des adverbes de degré, cf. De Cesare 2002). E3 s'ouvre par une variable chiffrée (97.86), suivie d'un segment fixe (le syntagme nominal *millions de titres*) et d'une variante, à savoir d'un segment qui varie au niveau de la formulation mais minimalement sur le plan sémantique (*ont été négociés / ont changé de main*). Pour finir, le Noyau de E4 s'ouvre par une variable (le quantificateur *Toutes les*) et se termine par une autre variable (l'adverbial *en baisse*), dépendant de la performance de la journée qui vient de s'écouler.

Cette brève analyse des segments réalisés dans les Noyaux des énoncés générés par le logiciel-rédacteur CAC40 permet de relever trois points importants,

auxquels on ne s'attendait en partie pas et qui montrent que les textes générés par ce logiciel sont plus complexes que ce que l'on avait pensé. Il s'agit en premier lieu du fait que les Noyaux ne se composent pas d'une structure à trou simple, dans laquelle à une première partie fixe suit une variable. Deuxièmement, dans les Noyaux les segments qui varient (variables ou variantes) ne se placent pas de manière rigide après les segments fixes. Nous reviendrons sur ce point dans le § 5.4. Troisièmement, les variables (notamment les données chiffrées et les noms de valeurs boursières) occupent différentes positions du Noyau. Comme attendu, elles se placent à la fin du Noyau (comme en E7, mais aussi E1 et E2, où elles se réalisent au sein des derniers constituants syntaxiques). On les trouve toutefois également au début du Noyau (E3 et E4 ; cf. aussi E10).

5.2 Nature des segments qui composent les Cadres

Les UIs de Cadre ouvrent trois énoncés (E1, E6 et E7) réalisés dans trois blocs textuels différents (B1, B3 et B4). Une première observation sur ces trois Cadres est que chacun d'entre eux se compose d'un segment uniforme (il s'agit soit de segments fixes soit de variantes) : on ne trouve donc pas les structures à trou décrites pour les Noyaux (§ 5.1).

Le Cadre de E1 (qui ouvre le Bloc 1) est rempli par un segment qui coïncide avec une variante sémantique. Dans les textes du corpus, on relève un paradigme de trois syntagmes prépositionnels adverbiaux temporels alternatifs, qui ont une fréquence d'emploi relativement équilibrée : *En clôture* (38 occ.) */ A l'issue des dernières transactions* (31 occ.) */ A la clôture du marché* (31 occ.). En ce qui concerne la fonction de ce Cadre, on relève qu'il joue un rôle cohésif important. Le segment qui réalise le Cadre du Bloc 1 est en effet lié sémantiquement au verbe de la phrase du titre qui chapeaute l'article, qui dénote lui aussi le moment de fermeture de la Bourse parisienne (pour plus de détail sur ce lien sémantique et une analyse détaillée des verbes et prédicats qui composent les textes du Corpus CAC40, cf. Eliasson en prép.). Le lien qui s'instaure entre le titre du texte, notamment le verbe, et le corps du texte, via le contenu du Cadre du premier énoncé, est explicité au point (7) (à noter que les accolades incluent toutes les formulations du Corpus CAC40 qui s'alternent au sein d'un même segment) :

(7) {Le marché boursier parisien / Le CAC 40 / La Bourse de Paris} {***conclut / termine / clôture***} la séance {en hausse (modérée) / en baisse (modérée) / inchangé}. [titre]

{***En clôture / A l'issue des dernières transactions / A la clôture du marché***} /**Cadre** […] //**E1**.

Les deux autres UIs de Cadre sont remplies par un segment fixe, qui se retrouve tel quel dans les 100 textes du corpus. Dans les deux cas, la fonction de ce segment est encore une fois clairement cohésive et permet de lier différentes unités textuelles. Le Cadre interne au Bloc 3 (*En seconde position sur le podium*) lient entre eux deux énoncés, à savoir E6 (qui identifie la deuxième meilleure valeur de la journée) à E5 (qui identifie la meilleure valeur de la journée). On a donc ici raccord entre deux énoncés au sein d'un même bloc :

(8) // / La meilleure performance de la journée est attribuée à Total avec 2.24% à 39.69 euros. /Noyau //**E5 En seconde position sur le podium**, /**Cadre** Airbus clôture la séance à 110.7 euros, /Noyau en hausse de 1.78%, /Appendice suivi par Renault à 34.58 euros (1.65%). /Appendice //**E6**

Le Cadre suivant (*En queue de peloton*) occupe une double position au niveau structurel : il ouvre un énoncé (E7) et en même temps le bloc dont il fait partie (Bloc 4). Par sa position-pivot, ce Cadre permet donc de lier des contenus appartenant à deux niveaux textuels hiérarchiquement distincts. A un premier niveau (à savoir celui des blocs), il permet la concaténation de deux énoncés au sein du bloc ; en effet, le contenu du Cadre de E7 reste valable pour interpréter le Noyau de E8 :

(9) // / **En queue de peloton**, /**Cadre** on trouve Vivendi. /Noyau //**E7** Le titre termine la journée en baisse de -1.35% à 29.21 euros. /Noyau //**E8**

A un niveau hiérarchique supérieur, le Cadre de E7 permet de lier entre eux deux blocs de texte : le contenu du Bloc 4, qu'il ouvre, se rapporte au contenu du Bloc 3. On identifie dans ce cas un regroupement d'énoncés relativement ample, qui intéresse le contenu sémantico-pragmatique des Blocs 3 et 4. La fonction de cette macro-unité textuelle consiste à présenter un classement des valeurs en fonction de leur performance individuelle.

5.3 Nature des segments qui composent les Appendices

Par rapport aux trois UIs de Cadres, les trois UIs d'Appendices (réalisées au sein de E6 et E10) présentent une plus grande variation au niveau des segments qui les composent.

Les deux UIs d'Appendices de E6 sont remplies par des structures à trou, à savoir par des segments fixes dans lesquels viennent s'insérer des informations nouvelles correspondant à des variables (pour une autre interprétation de la structure informationnelle de E6, cf. note 14) :

(10) En seconde position sur le podium, /Cadre Airbus clôture la séance à 110.7 euros, /Noyau **en hausse de 1.78%**, /Appendice **suivi par Renault à 34.58 euros** (1.65%). /Appendice //**E6**

La variable de la première UI d'Appendice de E6 coïncide avec une donnée chiffrée (*en hausse de* [chiffre]%) ; sa fonction consiste à spécifier la performance de la valeur nommée dans le Noyau (dans l'exemple fourni : *Airbus*). La deuxième UI d'Appendice de E6 contient à son tour deux variables : le nom d'une valeur boursière (dans l'ex. *Renault*) et une donnée chiffrée (*suivi par* [valeur boursière] *à* [chiffre] *euros*). La fonction de cette deuxième UI d'Appendice consiste à présenter, à l'arrière-plan de l'énoncé, des informations pertinentes sur une troisième valeur qui a marqué la journée boursière par sa performance à la hausse.

La troisième UI d'Appendice des textes générés par le logiciel-rédacteur CAC40, présente à la fin de E10, se compose d'un segment quasiment fixe, qui varie uniquement au niveau formel. En fonction du nombre grammatical du sujet, réalisé dans le Noyau (*L'action / Plusieurs valeurs*), le pronom possessif singulier *sa* (cf. ex. 11) alterne avec le pluriel *leur* (comme en 12). La fonction de cet Appendice consiste à spécifier le volume de transaction de la ou des valeurs identifiées dans le Noyau :

(11) L'action Alstom a enregistrée [sic] d'importants volumes de transactions, /Noyau **supérieurs d'un tiers à sa moyenne quotidienne** /Appendice. //E10 (texte généré le 2.6.2021)

(12) Plusieurs valeurs (Alstom, AXA, Bouygues, Cap Gemini, Carrefour, Crédit Agricole, Engie, EssilorLuxottica, Hermès, Legrand, LVMH, Orange, Renault, Safran, Saint Gobain, Schenider Electric, Société Générale, ST Microelectronics, Téléperformance, Vivendi et Worldline) ont enregistrées [sic] d'importants volumes de transactions, /Noyau **supérieurs d'un tiers à leur moyenne quotidienne** /Appendice. //E10 (texte généré le 11.5.2021)

La présence de variables au sein de deux UIs d'Appendice confirme une de nos attentes (cf. les hypothèses exprimées au début du § 5). Cette distribution informationnelle des variables est très intéressante, puisqu'on a un cas de figure dans lequel des informations nouvelles sont présentées à l'arrière-plan de l'énoncé. Cet

aspect sera développé dans le paragraphe suivant, centré sur la nature des segments qui constituent le Focus informationnel de l'énoncé.

5.4 Nature des segments qui réalisent le Focus informationnel

Dans le cadre théorique adopté, le Focus se définit comme l'information la plus importante de l'énoncé, réalisée dans le Noyau. Le Focus coïncide généralement (et donc pas nécessairement) avec une information nouvelle, située en fin de Noyau. En termes syntaxiques, le Focus recouvre typiquement le constituant post-verbal ou une partie de ce constituant (pour plus de détails, cf. Ferrari *et al*. 2008 : 95-99). Dans le Tableau 9, le Focus informationnel de chaque énoncé est mis en relief avec le caractère gras.

Tableau 9. Identification du Focus informationnel des énoncés

Blocs	Texte généré le 2.6.2021
1	A l'issue des dernières transactions, /Cadre l'indice phare de la Bourse de Paris progresse **de 0.49% à 6521.52 points**. /Noyau //**E1**
2	La séance boursière a été moyennement active **avec un volume d'échanges de 2.75 milliards d'euros**. /Noyau //**E2** 54.14 millions de titres **ont changé de main**. /Noyau //**E3** 27 valeurs terminent **dans le vert** et 13 s'affichent **en recul**. /Noyau //**E4**
3	La meilleure performance de la journée est attribuée **à Total avec 2.24% à 39.69 euros**. /Noyau //**E5** En seconde position sur le podium, /Cadre Airbus clôture la séance **à 110.7 euros**, /Noyau en hausse de 1.78%, /Appendice suivi par Renault à 34.58 euros (1.65%). /Appendice //**E6**
4	En queue de peloton, /Cadre on trouve **Vivendi**. /Noyau //**E7** Le titre termine la journée **en baisse de -1.35% à 29.21 euros**. /Noyau //**E8 Worldline et Téléperformance** sont également à la traîne. /Noyau //**E9**
5	L'action Alstom a enregistrée **d'importants volumes de transactions**, /Noyau supérieurs d'un tiers à sa moyenne quotidienne. /Appendice //**E10**
6	EssilorLuxottica se distingue **avec un plus haut de 52 semaines à 143.66 euros** et une valorisation boursière **grimpant à 62,947.97 M€**. /Noyau //**E11**

Dans les textes générés, on observe trois schémas informationnels différents, dont l'un d'entre eux est clairement dominant. Il s'agit du cas dans lequel le Focus de l'énoncé est placé en fin de Noyau et correspond à ou se compose de variables : données chiffrées (dans le cas de E1 : « de [x]% à [x] points »), non chiffrées (*en recul*, E4) ou noms de valeur(s) (cf. E7 : *Vivendi*). Les deux autres configurations informationnelles qui se dégagent sont marginales. On relève tout d'abord le cas d'énoncés à Focus final réalisé par un segment fixe (cf. E10) ou une variante

sémantique (E3). La troisième configuration, qui dévie également du schéma informationnel prédominant, voit en revanche le Focus en début de Noyau (E9). L'identification du Focus à l'initiale du Noyau repose sur deux éléments : la présence d'une variable (dans E9 on trouve en l'occurrence toujours une coordination de deux valeurs : *Wordline et Téléperformance*), vers laquelle pointe une marque de focalisation ; dans le cas en question, il s'agit de l'adverbe paradigmatisant *également* (sur cet adverbe, cf. Andorno & De Cesare 2017).

6. Bilan final : entre *répétition* et *variation*

Dans la première partie de cet article nous avons présenté une méthodologie permettant de décrire et d'évaluer la 'qualité' des textes générés de manière automatique, ainsi que d'identifier certaines de leurs spécificités par rapport aux textes non générés. En particulier, nous avons identifié une série de concepts théorico-descriptifs qui permettent d'aboutir à une description adéquate des textes générés. Dans la deuxième partie du travail, nous avons ensuite décrit un échantillon de textes générés dans le domaine de la finance.

Le bilan que l'on peut tirer à l'issue de notre analyse est relativement clair : les textes générés par le logiciel-rédacteur CAC40 sont très similaires les uns aux autres. Ceci est évident au niveau lexical, où l'on constate que les mêmes (paradigmes de) mots se retrouvent dans tous les textes du Corpus CAC40. Ces mots constituent des segments fixes (à savoir des « briques de texte », qui couvrent plus de 60% de chaque texte) ou des variantes sémantiques (15%). Les éléments présents une seule fois dans le corpus (à savoir les *hapax legomena*) correspondent en revanche uniquement à des chiffres, qui ne font bien entendu pas à proprement parler du lexique.

La similarité – et donc répétitivité – des textes générés par le logiciel-rédacteur CAC40 peut être observée à d'autres niveaux, moins évidents à percevoir que le choix du lexique. Dans la présente étude, nous nous sommes penchés sur les propriétés textuelles et informationnelles des textes générés. Un premier constat, relatif à la segmentation des textes en unités, est que la plupart des articles générés présente la même macro-structuration en blocs (et ceci est un autre aspect relativement évident à cueillir quand on procède à une comparaison intertextuelle), qui se découpent à leur tour pratiquement toujours de la même manière : chaque bloc se compose du même nombre d'énoncés, basés sur les mêmes schémas informationnels, y compris la même distribution du Focus informationnel au sein du Noyau des énoncés. Au niveau de la macro- et micro-structuration des textes

générés, on a donc une structure fixe, commune à tous les textes du corpus, qui présente l'architecture dessinée dans la Figure 3.

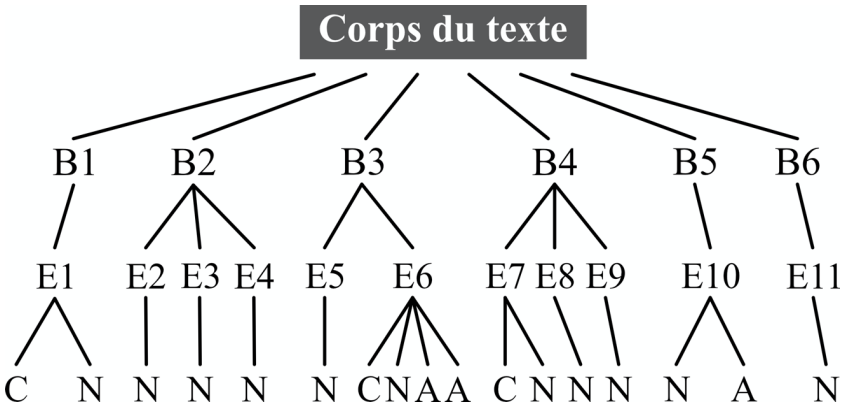

Figure 3. Architecture des textes générés par le logiciel-rédacteur CAC40.
Légende : B = Bloc ; E = Énoncé ; C = Cadre ; N = Noyau ; A = Appendice.

Les variations structurelles que l'on observe dans l'architecture d'un certain nombre de textes (par ex. le fait que certains articles comportent moins de blocs) s'expliquent par l'absence de données chiffrées, notamment liées à des performances spéciales au cours de la journée dont le texte présente un compte-rendu. Au niveau structurel, l'absence de ces données chiffrées se traduit par la non-réalisation de l'unité textuelle (énoncé, voire bloc) chargée de véhiculer l'information en question.

Un constat similaire vaut également pour les segments (fixes, variables et variantes) qui remplissent les plus petites unités qui composent les textes, à savoir les UIs. Toutes les UI des textes du Corpus CAC40 présentent les mêmes segments, qui se suivent dans le même ordre. Ici se dessinent des constantes qui s'expliquent tant par la nature des segments que par celle des UIs. Conformément à nos hypothèses de départ (formulées dans le § 5), nous avons observé que les variables sont distribuées principalement dans les Noyaux de l'énoncé, mais qu'on en trouve également dans les UIs d'Appendice. Les variables, qui constituent un trait caractérisant des textes générés, occupent donc aussi bien le premier plan que l'arrière-plan des énoncés. Les Cadres (qui sont directement impliqués dans la cohésion textuelle) sont en revanche remplis par des formes fixes ou des variantes sémantiques, qui changent uniquement au niveau de leur formulation.

La grande répétitivité des textes générés par le logiciel-rédacteur CAC40 n'est en soi pas gênante. Elle devient en effet évidente dans deux cas particuliers.

Premièrement, quand, dans un même texte, on retrouve les mêmes mots pleins (comme par exemple *clotûre*) ; ici, il faut préciser que plus les mots se répètent à brève distance, plus cela peut gêner la personne qui lit (pour détails, cf. Eliasson en prép.). Deuxièmement, quand on compare entre eux un grand nombre de textes. Pris individuellement, chaque texte généré présente une variation interne suffisamment importante pour résulter naturel. Nous avons par exemple observé la déclinaison de schémas informationnels variés à différents niveaux. Au niveau des énoncés, chaque texte comporte des énoncés simples, constitués du seul Noyau, et des énoncés complexes, constitués de deux, voire trois UI (cf. Fig. 3). Au sein des Noyaux, ensuite, le Focus informationnel ne se présente pas toujours à la fin et ne coïncide pas forcément avec une variable (chiffre ou valeur boursière).

Tous ces aspects constituent à notre avis des éléments décisifs pour évaluer positivement la qualité des textes générés[16]. Un texte dans lequel on trouverait un seul schéma informationnel, répété dans tous les énoncés, pourrait en effet vite être perçu comme anormal, et donc artificiel. La personne qui lit un tel texte pourrait peut-être même se douter qu'elle a sous les yeux l'output d'un logiciel d'écriture. La qualité des textes générés est bien entendu également assurée par leur cohérence interne, prévue au niveau du template. Chaque texte intègre même dans son tissu des éléments cohésifs explicites, comme l'adverbe paradigmatisant *également*, qui permet de lier deux énoncés contigus (cf. Andorno & De Cesare 2017). La cohésion textuelle est en outre assurée par les contenus réalisés dans les UIs de Cadre.

Notre bilan est bien entendu très provisoire : il se base en effet pour l'instant sur des textes générés par un seul logiciel d'écriture et sur un faisceau relativement restreint de traits linguistiques et textuels. Les prochaines étapes de notre recherche sur les textes générés consisteront donc à élargir notre grille d'analyse, à considérer le produit d'autres « logiciels-rédacteurs », notamment dans le domaine de la finance (cf. De Cesare, Eliasson, Weidensdorfer, en prép.), et à comparer les textes générés à des textes rédigés par des journalistes en chair et en os.

[16] Au niveau grammatical, on relève toutefois une erreur d'accord sur un participe passé du verbe *avoir* : *L'action [x] a enregistrée / Plusieurs valeurs ont enregistrées* (cf. les exemples 11 et 12). Cette faute, présente dans les premiers textes générés, a été corrigée à partir du 23.6.2021. En matière d'erreurs, mentionnons encore le fait que, comparés à d'autres textes générés, dans lesquels se glissent différentes catégories récurrentes de fautes typographiques (espace non requis avant une virgule ou autres signes de ponctuation ; absence de majuscule après un point etc.), ceux du logiciel CAC40 sont très soignés. Nous avons relevé une seule coquille de ce genre (il s'agit de l'absence d'un accent circonflexe sur le 'o' de *clôturant* ; cf. le texte cité à la note 13).

Remerciements

Je remercie Tom Weidensdorfer pour la sélection des 100 textes qui composent le Corpus CAC40, pour différentes observations sur les propriétés de ces textes ainsi que pour son aide dans la mise en page de cette contribution et la création de la Fig. 3. Je remercie également les deux relecteurs/relectrices anonymes de ce texte pour leurs commentaires constructifs.

Bibliographie

Andorno, C.M. & De Cesare, A.-M. 2017. Mapping additivity through translation: From French *aussi* to Italian *anche* and back in the Europarl-direct corpus. In A.-M. De Cesare & C.M. Andorno (eds), *Focus on Additivity. Adverbial Modifiers in Romance, Germanic and Slavic Languages (Pragmatics & Beyond New Series n°278*). Amsterdam-Philadelphia: John Benjamins, 157-200.

Belen Baez, M. 2018. *Génération de récits à partir de données ambiantes. Informatique et langage* [cs.CL]. Thèse de doctorat, Université Grenoble Alpes.

Danlos, L. 1991. Génération automatique de textes en langue naturelle. In J. Anis & J.-L. Lebrave (eds), *Texte et ordinateur. Les Mutations du Lire-Ecrire (Linx, hors-série n°4)*, 197-214; doi: https://doi.org/10.3406/linx.1991.1198.

Danlos, L. 2000. Génération automatique de textes. In J.-M. Pierrel (ed), *Ingénierie des langues*. Paris: Hermès Science, 311-330.

De Cesare, A.-M. 2002. *Intensification, modalisation et focalisation. Les différents effets des adverbes* proprio, davvero *et* veramente. Bern: Peter Lang.

De Cesare, A.-M. & Laura Baranzini. 2011. La variété syntaxique des dépêches d'agence publiées en ligne. Réflexions à partir d'un corpus de langue italienne. In A. Ferrari & L. Lala (eds.), *Variétés syntaxiques dans la variété des textes* online *en italien : aspects micro- et macrostructuraux* [*Verbum* XXXIII/1-2], 247-298.

De Cesare, A.-M., Garassino, D., Agar Marco, R., Albom, A. & Cimmino, D. 2016. *Sintassi marcata dell'italiano dell'uso medio in prospettiva contrastiva con il francese, lo spagnolo, il tedesco e l'inglese. Uno studio basato sulla scrittura dei quotidiani online (Linguistica contrastiva 5)*. Frankfurt am Main: Peter Lang.

De Cesare, A.-M., Eliasson, E. & Weidensdorfer, T. En prép. Gerarchie testuali della scrittura generata automaticamente in ambito finanziario. Italiano-francese a confronto. In A.-M. De Cesare, A. Ferrari, F. Pecorari (a c. di), *Forme della scrittura italiana contemporanea in prospettiva contrastiva. La componente testuale*. Firenze: Cesati.

Dierickx, L. 2019. *Production automatisée d'informations: une ligne du temps*. Publié par Ohmybox. https://journodev.tech//generation-automatique-de-textes-et-journalisme-une-ligne-du-temps/ (visité le 15.11.2021).

Dierickx, L. 2020. *La production automatisée d'informations en appui aux pratiques journalistiques: Analyse des représentations, des conditions d'association et de la structuration des usages en Belgique francophone*. Thèse de doctorat, Université Libre de Bruxelles.

Dierickx, L. 2021. Journalisme algorithmique : un état de l'art de la recherche. In L. Dierickx (ed.), *Journalisme algorithmique*. *Les carnets du Laboratoire des pratiques et des identités journalistiques*. In *Les Carnets du LaPIJ* 2, 8-10.

Eliasson, Elina. En prép. L'emploi des verbes dans les textes générés automatiquement sur l'indice boursier CAC40 : une perspective aspectuo-temporelle.

Ferrari, A. 2014. The Basel Model for paragraph segmentation: the construction units, their relationships and linguistic indication. In S. Pons Bordería (ed.), *Discourse Segmentation in Romance Languages*. Amsterdam: John Benjamins, 23-54.

Ferrari, A. et al. 2008. *L'interfaccia lingua-testo. Natura e funzioni dell'articolazione informativa dell'enunciato*. Alessandria: Edizioni dell'Orso.

GPT-3. 2020. A robot wrote this entire article. Are you scared yet, human? Publié par The Guardian. https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3 (visité le 07.01.2022).

Graefe, A., Haim, M., Haarmann, B., & Brosius, H. 2018. Readers' perception of computer-generated news: Credibility, expertise, and readability. In *Journalism*, 19(5), 595–610.

Jung, J., Song, H., Kim, Y., Im, H., & Oh, S. 2017. Intrusion of software robots into journalism: The public's and journalists' perceptions of news written by algorithms and human journalists. In *Computers in Human Behavior*, 71(C), 291-298.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. 2014. The Sketch Engine: ten years on. In *Lexicography*, 1, 7-36.

Linden, C.-G. & Dierickx, L. 2019. Robot Journalism: The damage done by a metaphor. In *Unmediated: Journal of Politics and Communication* 2, 152-155.

Mégean, O. 2021. *Les robots rédacteurs: une technologie stable, efficace et éprouvée.* Publié par Demain.ai. https://www.demain.ai/nos_publications/les-robots-redacteurs-une-technologie-stable-efficace-et-eprouvee (visité le 09.09.2021).

Meier-Vieracker, S. 2020. Die Verdatung des Fußballs: Spuren von Algorithmen in der Fußballberichterstattung. In *Muttersprache*, 130, 304-318.

Meier-Vieracker, S. 2021. Wer schreibt? Automatisierter Fußballjournalismus aus kommunikations- und sprachwissenschaftlicher Sicht. Vortrag, TU Dresden, 21.04.2021.

Ponton, C. 1997. Génération automatique de textes : 30 ans de réalisations. In *Génération Automatique de textes GAT'97*, 1-14.

Schevenels, H. 2019-2020. *La génération automatique de textes en presse écrite : historique problématisé, questions d'éthique et analyse de contenu*. Thèse de Master, Université de Liège.

Stalph, F., Thaesler-Kordonouri, S. & Thurman, N. 2021. Exploring audience perceptions of, and preferences for, data-driven 'quantitative' journalism. Working paper.

Thurman, N. 2019. Computational Journalism. In K. Wahl-Jorgensen & T. Hanitzsch (eds), *The Handbook of Journalism Studies*, Second Edition. New York: Routledge, 180-195.

Van der Lee, C., Krahmer, E. & Wubben, S. 2018. Automated learning of templates for data-to-text generation: comparing rule-based, statistical and neural methods. In *Proceedings of the 11th International Natural Language Generation Conference*, 35-45. Association for Computational Linguistics. https://aclanthology.org/W18-6504.pdf.