Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Génération automatique de rapports d'analyse financière. Développement et évaluation d'un système à base de règles

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Ingénierie Multilingue

par

Milena CHAÎNE

Directeur de mémoire :

Cyril Grouin

Encadrants entreprise :

Guillaume Lacronique
Frank Megel

Année universitaire 2018/2019

TABLE DES MATIÈRES

Liste des figures

Liste des tableaux

Résumé

Ce mémoire décrit le développement et l'évaluation d'un système de génération automatique de rapports d'analyse financière. Nous avons défini, modélisé et codé des règles de génération pour un système commercial permettant de générer régulièrement plusieurs centaines de rapports simplifiés d'analyse. Le système doit modéliser des données informatiques non-textuelles complexes afin de les utiliser dans des règles de génération définies en collaboration avec des analystes financiers. Ces règles doivent être représentatives de l'expertise d'un·e analyste et permettre de générer un texte cohérent d'un point de vue rhétorique et linguistique. Enfin, les rapports d'analyse générés doivent employer la langue spécialisée du domaine qu'est la finance de marché. Pour évaluer la qualité des textes générés, nous avons mis en place un protocole d'évaluation humaine par des spécialistes du domaine.

Mots clés : génération automatique de textes, domaine de spécialité, langue de spécialité, approche symbolique, évaluation humaine

Remerciements

Je souhaite d'abord remercier mon directeur de mémoire, Cyril Grouin, pour ses conseils et l'aide précieuse qu'il m'a apportée lors de la rédaction de ce mémoire, ainsi que tous les membres de l'équipe pédagogique du master dont les enseignements m'ont permis d'enrichir ma réflexion durant ces deux années.

Je souhaite ensuite remercier mes encadrants entreprise, Guillaume Lacronique et Frank Megel, qui m'ont appuyée et aidée à orienter mon travail de recherche, mais également tou·te·s les collègues avec qui j'ai eu l'opportunité de travailler durant ces six mois à Exane. Je remercie tout particulièrement les analystes et vendeu·r·se·s qui ont pris le temps de participer à l'évaluation du système décrit dans ce mémoire.

Je remercie Margaux Duhayon, Boyu Niu, Elvira Quesada et Ferial Yahiaoui, pour leur aide et leur compagnie durant ces deux années de master.

Je remercie Floria Cophignon pour son soutien sans failles et ses conseils terminologiques, et mes parents pour leur soutien également sans failles et leurs conseils non-terminologiques. Enfin, je souhaiterais dédier ce mémoire à ma grand-mère.

INTRODUCTION

Présentation générale

Ce mémoire est le résultat d'un projet de recherche en traitement automatique des langues effectué en entreprise à Exane, une société de conseils en investissement financier, et dans le cadre du master Traitement Automatique des Langues (TAL) de l'Inalco. L'objectif du projet est de générer automatiquement et régulièrement des rapports d'analyse financière pour plusieurs centaines d'entreprises, à partir de données propriétaires collectées par Exane.

Dans ce cadre, nous utilisons un système de GAT commercial dont nous devons établir, modéliser et coder les règles. Pour ce faire, nous devons atteindre un niveau d'expertise satisfaisant dans le domaine de spécialité qu'est la finance de marché, et plus particulièrement sa langue de spécialité, dans laquelle les rapports seront rédigés. Nous devons maîtriser suffisamment le domaine pour pouvoir, en collaboration avec des analystes financi·er·ère·s, en extraire une représentation de connaissances satisfaisante.

Par ailleurs, nous avons également mis au point un protocole d'évaluation humaine de ces textes permettant de s'assurer de leur qualité. Nous devons définir des critères d'évaluation précis qui nous permettront de déterminer si les textes atteignent l'objectif qui leur est fixé.

Problématique de recherche

L'objectif de ce mémoire est d'offrir, entre autres, une réflexion autour des questions suivantes :

— quelle est l'influence d'un domaine de spécialité, comme la finance de marché, sur un système de GAT ?
— quelles sont les différentes étapes de génération automatique d'un texte linguistiquement cohérent ?
— un texte a-t-il un objectif, et si oui, est-il possible d'évaluer si ledit texte atteint cet objectif ?

En décrivant notre travail de recherche sur un système spécifique de GAT, nous espérons fournir des éléments de réponse à ces questions générales.

Plan

Le chapitre 1 décrit le contexte industriel dans lequel ce projet de recherche a eu lieu, ainsi que les besoins et contraintes exprimés par l'entreprise. Le chapitre 2 est un état de l'art de la génération automatique de textes qui introduit les grandes problématiques de recherche liées à la GAT. Dans le chapitre 3, nous dégageons les spécificités du projet qui orienteront les expérimentations des chapitres suivants.

Dans le chapitre 4, nous décrivons le corpus auquel nous avons accès ainsi que le système de GAT utilisé pour générer des rapports. Dans le chapitre 5, nous présentons nos expérimentations liées à différents modules du système. Le chapitre 6 consiste en l'évaluation globale du projet par des spécialistes du domaine.

Note

Ce mémoire a été approuvé après relecture par Exane BNP Paribas. Certains éléments présentés dans ce mémoire ont pu être altérés ou anonymisés afin de respecter le secret professionnel.

GLOSSAIRE

Ce glossaire définit les abréviations de termes employées dans ce mémoire. Les définitions linguistiques présentées dans ce mémoire sont adaptées de [Moeschler and Auchlin, 2018]. Les définitions financières sont adaptées de [Avenel et al., 2017].

Linguistique et informatique

ACL Association for Computational Linguistics

BLEU BiLingual Evaluation Understudy

GAT Génération Automatique de Textes

GER Génération d'Expressions Référentielles

INLG International Conference on Natural Language Generation

NLG Natural Language Generation

ROUGE Recall-Oriented Understudy for Gisting Evaluation

SIGGEN Special Interest Group in Generation

TAL Traitement Automatique des Langues

Finance

AMF Autorité des Marchés Financiers

EBITDA Earnings Before Interest, Taxes, Depreciation, and Amortization : bénéfice avant intérêts, impôts, dépréciation et amortissement

EPS Earnings Per Share : bénéfice par action

EV Enterprise Valuation : valeur de l'entreprise

FCF Free Cash Flow : flux de trésorerie disponible

FE First Estimated : première année estimée

GICS Global Industry Classification Standard

GMV Gross Merchandise Volumne : volume marchand brut

KPI Key Performance Indicator : indicateur clé de performance

LP Last Published : dernière année publiée

LT Long Term : long terme

ROCE Return On Capital Employed : rentabilité des capitaux investis

ROE Return On Equity : rentabilité des capitaux propres

ST Short Term : court terme

Autres

HPC Home and Personal Care : produits d'entretien ménager et de soins du corps

OEM Original Equipment Manufacturer : fabricant d'équipement d'origine

$OLAP$ Online Analytical Processing : traitement analytique en ligne

Première partie

Contexte

CHAPITRE 1

CONTEXTE INDUSTRIEL

Sommaire

1.1 Introduction

La problématique de recherche abordée dans ce mémoire repose sur les données que nous avons pu exploiter et les expérimentations que nous avons pu mener dans le cadre d'un stage de recherche. Nous présenterons d'abord l'entreprise dans laquelle notre stage a eu lieu, le groupe Exane [1], et plus spécifiquement, sa filiale Exane BNP Paribas. Nous fournirons une définition concise du métier d'analyse financière qui est au centre du système de génération que nous décrivons dans ce mémoire.

Nous décrirons ensuite les besoins et contraintes spécifiques au projet de génération automatique de rapports d'analyse financière envisagé par Exane, ainsi que les outils mis à notre disposition pour le réaliser, et nous tenterons d'en dégager une première problématique industrielle.

Nous enrichirons ensuite cette dernière d'un point de vue linguistique et informatique dans le chapitre 3, en nous appuyant sur notre description de l'état de l'art de la génération automatique de textes présenté dans le chapitre 2.

FIGURE 1.1 – Logo d'Exane BNP Paribas

1. www.exane.com.

1.2 Présentation de l'entreprise

1.2.1 Structure de l'entreprise

Fondé en 1990, le groupe Exane est une entreprise d'investissements financiers de la finance de marché. En 2004, Exane conclut un accord de partenariat avec BNP Paribas au niveau de son activité d'intermédiation actions, en apportant son expertise en recherche actions tout en préservant son indépendance. Présent sur neuf sites internationaux (dont entre autres Paris, où s'est déroulé ce stage, et Londres), Exane se spécialise dans trois types de métiers :

— l'intermédiation actions, particulièrement sur le marché européen : recherche, vente, exécution sur des actions européennes (*Exane BNP Paribas*)
— la gestion de contrats dérivés (*Exane Derivatives*)
— la gestion d'actifs financiers, via trois sociétés de gestion spécialisées (*Exane Asset Management*, *Ellipsis AM*, et *Ixios Asset Management*)

Ces métiers se rassemblent autour de trois fondamentaux de la culture d'Exane : le monde de l'*equity*, l'expertise en investissement et la qualité de la recherche, notamment de l'analyse financière.

1.2.2 Métiers concernés

Parmi les différents éléments d'expertise d'Exane, présentons plus en détails la notion d'analyse financière, qui est au cœur de notre projet. L'analyse financière consiste à présenter une synthèse de l'état d'une entreprise à un instant donné, c'est-à-dire ses forces et ses faiblesses par rapport à l'environnement économique et financier dans lequel elle se trouve (ou se trouvera dans le futur). Pour ce faire, un·e analyste financi·er·ère s'appuie en partie sur des données sectorielles et des données issues de la comptabilité de l'entreprise, qui sont confrontées au plan de développement de cette dernière via le calcul de flux de trésoreries, pondérés par un taux d'actualisation. L'objectif est de générer une valorisation ou un objectif de cours, en s'appuyant sur :

— la performance passée de l'entreprise
— la performance future estimée de l'entreprise
— la performance d'entreprises similaires (en termes de taille ou de secteur)

L'analyse financière est donc un rouage essentiel de la finance de marchés. Ici, on s'intéresse principalement au travail des analystes *sell side* (du côté du vendeur), dont la recherche se doit de conseiller les investisseurs en bourse lors de leurs décisions d'achat ou de vente. Les analystes présentent généralement leurs conseils d'investissement sous formes de rapports, ou notes[2], d'analyse qui se concentrent sur une entreprise ou son secteur.

Exane met le travail de recherche de ses analystes à disposition de ses clients via une interface web, Cube, qui leur permet d'avoir accès, entre autres, à des grilles financières contenant une partie des données numériques recueillies par Exane sur des entreprises tout comme les prévisions de future performance modélisées par la section recherche, ainsi qu'à des rapports d'analyse rédigés en anglais par des analystes qui expliquent, précisent ou complètent ces prévisions en fonction du contexte

2. Nous employons ici les termes « rapport » et « note » de façon interchangeable.

économique mais aussi politique ou parfois sociétal. Ces rapports délivrent également régulièrement une notation (*rating*) à chaque entreprise (table 1.1), qui décrit sa performance attendue. Les rapports mis en ligne sur Cube sont datés à la minute

Notation	Abréviation	Performance attendue de l'entreprise par rapport à celle du secteur
Outperform	(+)	Supérieure
Neutral	(=)	Similaire
Underperform	(-)	Inférieure

TABLE 1.1 – Définition des types de notation financière utilisés par Exane

de leur finalisation, qui peut différer légèrement de la date de leur mise en ligne (figure 1.2). Il est important de pouvoir contrôler la chronologie de la diffusion de ces rapports pour des raisons légales, en raison de l'influence que ces derniers peuvent avoir sur l'attitude des investisseurs qui les lisent.

> The investment recommendation was finalised at 06 :59 on 8 Oct. 2019 (London Time). It may differ from the date and time of broad dissemination on the website. Click here for Analyst Certification, Important Disclosures and Non-US Research Analyst disclosures.

FIGURE 1.2 – Exemple type de déclaration de datage d'une note d'analyse financière

On retrouve enfin toujours une section dédiée à différentes divulgations de la part des auteurs et d'Exane, ayant pour objectif d'éviter de potentiels conflits d'intérêt ou de rémunération (table 1.2) et d'établir clairement la situation de l'entreprise au moment de la rédaction du rapport. La structure interne d'un rapport d'analyse peut varier en fonction d'un nombre important de facteurs, dont par exemple :

— elle peut inclure une analyse des technologies auxquelles l'entreprise a recours, notamment si ces dernières sont nouvelles ou en évolution (par exemple, la construction de voitures électriques, les technologies 5G, etc.)
— si l'entreprise a récemment changé de dirigeant·e·s, le rapport peut présenter leur parcours
— on peut également retrouver des sections suggérant une liste de questions à poser à la direction (« Questions for management »), qui permettent d'orienter la réflexion du lecteur

Néanmoins, une note regroupe généralement en guise de résumé plusieurs points essentiels, présentés dans la table 1.3. Dans le cadre de ce mémoire, nous retiendrons qu'un rapport d'analyse est le résultat d'un travail d'analyse de données numériques mais aussi de tendances immatérielles, et qu'il est présenté comme une forme de discours argumentatif (dont l'objectif est de convaincre, si ce n'est de persuader le lecteur de la justesse de l'analyse) qui présente également une forte dimension temporelle. Enfin, le style et le format de ces documents fait l'objet de normes (légales ou simplement relatives au domaine de spécialité qu'est la finance de marché) plus ou moins strictes.

Divulgation (« Disclosure »)	Description
Analyst Certification	Certifie l'indépendance entre les recommandations exprimées et la rémunération des analystes
Research Analyst Disclosure	Décrit le type de relation, si elle existe, entre les analystes et l'entreprise concernée par le rapport
Regulatory Disclosures	Décrit le type de relation, si elle existe, entre Exane et l'entreprise concernée par le rapport
Prices & Ratings Chart	Historique comparatif des prévisions de performance de l'entreprise par Exane et de sa performance réelle

TABLE 1.2 – Liste simplifiée d'une partie des divulgations pouvant être associées à un rapport d'analyse

Section	Description
Investment case	Résumé du conseil d'investissement présenté
Valuation methodology	Méthode d'évaluation par laquelle l'analyste est arrivé·e à cette conclusion
Risks (upside)	Évènements incertains qui pourraient améliorer la situation décrite dans le futur
Risks (downside)	Évènements incertains qui pourraient aggraver la situation décrite dans le futur

TABLE 1.3 – Structure simplifiée d'un extrait de rapport d'analyse financière

1.2.3 Cadre et sujet du stage

Notre stage de fin d'études s'est déroulé au sein de l'équipe de recherche en Business Intelligence du department IT Advisory d'Exane BNP Paribas, à Paris. Il nous fallait donc nous familiariser avec les définitions de base du domaine financier présentées dans le reste de cette section 1.2, ainsi que leur implémentation dans l'écosystème informatique d'Exane. Pour répondre à la problématique de génération qui nous était posée (cf. section 1.4), nous avions accès à la base de données d'Exane servant à la génération des grilles financières mentionnées plus haut.

Cette base présente, pour chaque entreprise couverte par Exane, une série d'indicateurs, de données financières et de ratios fournis par l'entreprise et/ou calculés par Exane pour chaque année de couverture d'Exane, ainsi que des prévisions futures pour toutes ces données sur une période de plusieurs années (généralement entre deux et quatre ans)[3]. Ces données sont régulièrement mises à jour si un modèle d'analyse est modifié, ou encore lorsque l'entreprise publie son bilan annuel et ses résultat trimestriels. Elles sont présentées sous format HTML, PDF ou XLS. Sont également disponibles (et pertinents pour notre travail de génération) des tableaux comparatifs permettant d'évaluer la performance d'une entreprise par rapport à son secteur sur certains indicateurs clés.

3. Nous reviendrons sur la représentation temporelle de ces données dans le section 3.3.2.

1.3 Besoins et contraintes

1.3.1 Besoins

[redacted] [redacted] (=)

Growth and returns

FCF

Free Cash Flow should go up by 118% between 2018 and 2021, mostly due to EBITDA growth.

EPS

EPS has consistently gone up over the last 6 years, by 22% on average. Nevertheless, by comparison with recent trends, we expect EPS growth to slow down to 18.8% in 2019.

ROCE

ROCE remains high for 2019, although lower than its recent levels. [redacted] creates value with a ROCE of 27.1%, significantly above its cost of capital (8.7%).

ROE

The group's profitability is high, with an a ROE of 23.9% in 2019, in line with the historical average.

Financial structure

We anticipate that the dividend will increase to 6.3CHF in 2019, driven by earnings growth.

Valuation

The company uses limited financial leverage.

Le nom de l'entreprise et sa valorisation ont été anonymisés.

FIGURE 1.3 – Texte brut anonymisé d'un rapport généré

L'automatisation des tâches qui peuvent l'être avec un objectif d'optimisation de rendement est un processus déjà fermement engagé dans le monde de la finance. De nombreux outils (par exemple, les modèles de rédaction de rapport mentionnés en 1.2.2) sont à la disposition des analystes d'Exane pour les aider à intégrer rapidement les données numériques dont ils ont besoin. Un des projets de recherche d'Exane a pour objectif la rédaction des rapports courts d'analyse financière de façon automatique pour un très grand nombre d'entreprises (dans l'idéal, l'intégralité des entreprises couvertes par la section recherche : au moins six cents valeurs et, dans le futur, jusqu'à mille). Un exemple de rapport généré pendant notre stage est présenté dans la figure 1.3.

Ces rapports n'ont pas vocation, ni dans leur longueur ni dans leur complexité, à remplacer un rapport complet d'analyse financière. Un rapport tel qu'il est présenté sur Cube a une longueur minimale de plusieurs pages, et peut aisément en compter plusieurs dizaines (d'après nos calculs sur le corpus de rapports mis à notre disposition, la longueur moyenne d'un rapport récemment mis en ligne sur Cube était de plus de 500 mots [4]). Il présente sous forme textuelle le raisonnement derrière la valorisation proposée par l'analyste en s'appuyant sur des représentations graphiques de

4. On reviendra sur ce corpus et les expérimentations qu'il nous a permis d'effectuer dans les chapitres 4 et 5.

l'évolution de certains indicateurs de performance ou de tout autre élément permettant de justifier l'analyse. Par conséquent, le style de chaque rapport est également influencé par le style personnel du ou des analystes qui le rédigent (et qui se spécialisent généralement dans le suivi d'un secteur spécifique).

Cependant, pour s'adapter aux besoins de ses différents clients, Exane a également pour projet la génération de notes de recherche « simplifiées » accompagnant des résumés des grilles financières complètes de Cube. En effet, il est dans certains cas plus bénéfique d'avoir accès directement aux informations essentielles permettant de comprendre la situation d'une entreprise. L'objectif n'est donc pas d'effectuer directement le travail d'analyse, ce qui serait difficilement réalisable au vu des la variété des tâches (création de modèles, d'algorithmes, mais aussi recherche de terrain) qu'il recouvre, mais plutôt d'exprimer de façon textuelle les conclusions tirées à partir de ce travail. Notre modèle de génération devra donc répondre à un impératif de robustesse important dans la mesure où contrairement aux équipes d'analystes, il devra couvrir l'intégralité des secteurs suivis par Exane.

1.3.2 Contraintes

Dans ce contexte, nous notons plusieurs contraintes industrielles qui ont motivé la mise en place du projet de génération. D'une part, le métier de recherche et de conseil en investissement financier est soumis à diverses législations (notamment depuis la directive concernant les marchés d'instruments financiers MiFID II de l'Union Européenne en 2018) sous l'autorité de régulateurs locaux [5]. À ce titre, il est essentiel que les textes que nous génèrerons soient toujours le plus précis possible, et le moins à même d'induire un lecteur en erreur.

Par ailleurs, il sera nécessaire d'intégrer l'architecture de génération à un « pipeline » informatique déjà existant d'Exane, en utilisant les outils qui nous sont proposés. Par exemple, il faudra garder en tête que les données avec lesquelles nous travaillons sont susceptibles d'être modifiées (si un·e analyste modifie une de ses prévisions) et peuvent devenir « obsolètes » très rapidement (après un évènement perturbateur pour les marchés boursiers, par exemple) : le système développé devra donc regénérer les textes régulièrement [6].

Enfin, ce processus de génération peut s'apparenter à un processus de récapitulation de textes, ou du moins de simplification du raisonnement d'analyse. D'une part, on devra simplifier une partie du raisonnement d'analyse à une suite d'étapes logiques pouvant être interprétées par un programme, et d'autre part, on devra également synthétiser un style linguistique et rhétorique similaire à celui employé dans des rapports complets.

1.3.3 Outils

Un certain nombre d'entreprises privées proposent des solutions de génération automatique de textes à but commercial, avec un degré variable de multilinguisme et de fonctionnalité. On peut citer, entre autres :

— Arria (Arria NLG Studio [7])

5. En France, l'Autorité des Marchés Financiers (AMF).

6. Pendant notre stage, ce processus de regénération des données et des textes avait lieu toutes les nuits.

7. www.arria.com/studio/studio-overview.

— Automated Insights (Wordsmith [8])
— Yseop (Yseop Compose [9])

Dans le cadre de notre projet, Exane a choisi parmi ces différents fournisseurs de travailler avec Yseop et d'utiliser sa solution de génération, Yseop Compose. Dans le cadre de notre travail, nous avons donc pu profiter des fonctionnalités offertes par Compose pour assurer une partie du processus de génération.

Bien qu'un certain nombre de systèmes de génération à but non-commercial soient disponibles en ligne et/ou en open source [10], peu sont encore régulièrement maintenus et mis à jour. Parmi eux, on peut citer :

— le réaliseur de surface SimpleNLG [Gatt and Reiter, 2009] [11], codé en Java et disponible dans plusieurs langues
— OpenCCG [White, 2005] [12], un parseur et réaliseur de surface également codé en Java
— RosaeNLG [13], une bibliothèque JavaScript de génération de textes à base de modèles

Enfin, certains articles universitaires présentant des systèmes complets de génération publient l'intégralité du code de ces systèmes en accès libre, comme dans le cas de PASS [van der Lee et al., 2017] [14].

1.4 Problématique industrielle de génération

Que retenir du contexte industriel qui vient d'être présenté pour nous aider dans notre travail de génération automatique ? Nous pouvons poser deux questions liées :

> **pourquoi** et **comment** générer automatiquement des textes pouvant être qualifiés de rapports d'analyse financière ?

Pour répondre à la première question, nous pouvons relever les apports de la GAT au fonctionnement d'Exane dans le cadre de notre projet.

D'abord, la rédaction de ces rapports simplifiés d'analyse est un processus hautement automatisable qui serait par ailleurs beaucoup trop coûteux à réaliser entièrement manuellement. D'autre part, générer ces rapports de façon automatique nous permet également d'envisager de produire plusieurs versions présentant des variations (en terme de longueur, de style, etc.) pour différents publics ou clients. Il est possible de répercuter rapidement une modification des données dans le texte généré. Enfin, la mise en place d'un système de GAT nous permettra aussi de nous interroger sur le rapport entre les représentations textuelles et graphiques de la même information dans le domaine de l'analyse financière, et l'influence (positive ou négative) qu'a cette double représentation sur le niveau de compréhension des lecteurs.

Nous pouvons également identifier plusieurs questions complexes auxquelles il nous faudra apporter une réponse du point de vue linguistique et informatique.

8. www.automatedinsights.com/wordsmith.
9. www.compose.yseop.com.
10. Une liste non-exhaustive tenue par le Special Interest Group on Generation (SIGGEN) est disponible sur le site de l'Association for Computational Linguistics (ACL) : www.aclweb.org/aclwiki/Downloadable_NLG_systems.
11. www.github.com/simplenlg/simplenlg.
12. www.github.com/OpenCCG/openccg.
13. https://rosaenlg.org.
14. www.github.com/TallChris91/PASS.

Tout d'abord, nous devrons trouver un moyen de modéliser efficacement les données qui nous sont fournies pour qu'elles soient exploitables par un système de génération sans perte d'information, et intégrer cette modélisation au système d'Exane. Ensuite, comme dans le cas de la plupart des systèmes de GAT qui s'appuient sur un domaine de spécialité, nous serons confrontée à une question proche de l'ingénierie des connaissances : quelle approche ou architecture utiliser pour décomposer un processus complexe et hautement spécialisé comme l'analyse financière, à partir des informations que nous fournissent des experts du domaine (ici, donc, des analystes financiers) ? Enfin, nous devrons déterminer et développer une méthode pour évaluer les résultats de notre système et nous assurer que les textes que nous produisons sont au minimum conformes aux standards d'Exane, et idéalement qu'ils apportent une forte valeur ajoutée.

1.5 Conclusion

Nous avons présenté le contexte industriel dans lequel notre stage s'est déroulé et ce mémoire a été rédigé. À partir des besoins et des contraintes d'Exane et en s'appuyant sur les données et les outils qui nous ont été fournis, nous avons pu formuler une problématique industrielle de génération qui résume les attentes de l'entreprise : comment mettre en place un système de génération automatique de rapports d'analyse financière qui soit robuste, adaptable et intégrable au sein des outils déjà existants de l'entreprise ? Dans le chapitre 2, nous effectuerons un état de l'art de la génération automatique de textes, qui nous permettra d'enrichir cette problématique et de la transformer en problématique de recherche dans le chapitre 3.

CHAPITRE 2

ÉTAT DE L'ART

Sommaire

2.1 Introduction

Bien que son émergence en tant que domaine de recherche à part entière soit plus récente, la génération automatique de textes se trouve au cœur de nombre de domaines de traitement automatique des langues, et ce depuis l'émergence de ce dernier en tant que discipline [Zock and Sabah, 1992]. En dépit de cette relativement longue histoire, il reste difficile et controversé de définir strictement le domaine de la GAT : [Evans et al., 2002] notent que le seul consensus semble être qu'un système de génération de texte doit générer un texte (qui peut cependant être écrit ou parlé [Bateman and Zock, 2014]). De fait, comment définir la génération automatique de textes, ou les systèmes qui l'utilisent ? [Reiter and Dale, 1997, p.1] proposent une définition qui souligne plusieurs caractéristiques essentielles de la GAT de façon succincte :

> le sous-domaine de l'intelligence artificielle et de la linguistique informatique qui traite de la construction de systèmes informatiques capables de produire des textes compréhensibles, en anglais ou dans d'autres langues, à partir d'une représentation de données non-linguistique sous-jacente [1]

1. « the subfield of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information » (notre traduction).

Nous retiendrons de ce paragraphe plusieurs éléments qui permettront d'orienter ce tour d'horizon de l'état de l'art, avant d'approfondir dans le chapitre 3 les problématiques liées à notre projet :

— la variabilité des données, des domaines et des champs d'application concernés par la GAT
— l'importance accordée à la notion de système ainsi qu'à la structure desdits systèmes
— la description du texte généré par un système de GAT comme compréhensible, et les difficultés d'évaluation qu'un terme aussi flou peut générer

2.2 Périmètre de recherche

2.2.1 Domaines d'études

[Zock and Sabah, 1992] débutent leur liste chronologique de systèmes de génération avec [Yngve, 1961], un système de génération aléatoire de phrases en anglais via une grammaire générative qui se veut une partie d'un système de traduction automatique . Dans le champ des agents conversationnels et des systèmes de dialogue en général, la génération de réponses compréhensibles, pertinentes et intéressantes reste une problématique essentielle, de la création du système ELIZA [Weizenbaum, 1966] (souvent classifié comme un système de GAT) à nos jours : le jeu d'imitation proposé par [Turing, 1950], considéré comme une forme de test d'un agent de dialogue, évalue en grande partie la qualité du texte généré par ce dernier. Enfin, la GAT est également un domaine assez comparable à celui de la récapitulation de textes (*text summarization*), qui consiste également à générer de nouvelles données textuelles [Portet et al., 2009] [Manor and Li, 2019].

Si tout système de GAT se doit de produire des textes, l'objectif énonciatif de ces textes peut varier fortement, et donc influencer le développement de ces systèmes en les rapprochant de différents domaines d'études. Une majorité des systèmes de génération a pour objectif d'informer leur utilisateur, généralement en présentant des données difficiles à appréhender dans leur forme initiale (qu'elles soient trop abondantes [Yu et al., 2007] ou trop complexes [Hunter et al., 2012] pour pouvoir être présentées directement à un utilisateur) sous forme de texte. Cependant, certains systèmes de GAT, à l'image de NaturalOWL [Androutsopoulos et al., 2013], un système de textualisation d'ontologies, ou [Nguyen, 2013], qui explore l'utilisation de la GAT pour simplifier la création d'ontologies par des experts peu familiers avec des standards comme OWL, sont plus explicitement orientés vers le domaine d'étude de la représentation de connaissances. On notera que tous ces systèmes ont un objectif précis de génération, soulignant donc implicitement la difficulté que présente la création d'un système de génération multi-usages.

Dans d'autres cas, le texte généré n'a pas de visée objective, mais peut avoir une vocation narrative, humoristique, ou persuasive. Le système *How Was School Today ?* [Tintarev et al., 2016] est une expérience de génération de texte à visée narrative par des enfants pour raconter leur journée, tandis que JAPE [Binsted and Ritchie, 1994] est un système de génération automatique d'énigmes à base de contrepèteries dont la qualité humoristique fut évaluée par des enfants. Un système comme PASS [van der Lee et al., 2017], quant à lui, adapte le ton de son commentaire de match de football en fonction de l'équipe soutenue par son lecteur.

D'autres systèmes vont plus loin, et ont pour but d'inciter un changement de comportement chez le lecteur des textes produits. Par exemple, le système STOP [Reiter et al., 2003] avait pour objectif de persuader des fumeurs de réduire ou d'arrêter leur consommation ; SaferDrive [Braun et al., 2018], lui, encourage une conduite automobile plus sûre à partir de données sur les habitudes d'accélération et le freinage du conducteur. Enfin, on notera certains débats sur l'aspect éthique de la génération de texte [Smiley et al., 2017], notamment la notion de mensonge (volontaire ou par omission) de la part d'un système inanimé [van Deemter and Reiter, 2018].

2.2.2 Domaines de spécialité

On retrouve des systèmes de GAT, ou tout au moins des solutions inspirées par les paradigmes de GAT dans nombre de domaines de spécialité [2]. On peut notamment citer :

— le domaine biomédical avec des systèmes ayant pour but d'assister le personnel soignant [Hunter et al., 2012] ainsi qu'informer les proches des patients [Mahamood and Reiter, 2011]
— le « robojournalisme » qui permet la rédaction rapide d'articles convertissant des séries de données, comme les premiers résultats d'élections [3]
— dans le domaine financier, des systèmes de commentaires sur l'évolution des marchés financiers [Kukich, 1983] [Murakami et al., 2017] [Aoki et al., 2018]
— des systèmes de génération automatique de bulletins ou d'alertes météorologiques [Bourbeau et al., 1990] [Reiter et al., 2005]
— des systèmes de génération de commentaires [van der Lee et al., 2017] ou de résumés [Bouayad-Agha et al., 2011] de matchs sportifs
— l'aide à la génération de récits vidéoludiques [Garbe, 2019] [Mason et al., 2019]

En général, la génération automatique de textes répond donc à un besoin d'automatisation de la production de texte qui est soit dû à l'importance ou la complexité des données, ou à l'obligation de générer les textes en question de façon extrêmement rapide — trop pour des auteurs humains. Certains systèmes offrent en plus la possibilité de personnaliser le contenu généré ou de l'adapter à l'utilisateur du système. [Gatt and Krahmer, 2018] notent dans la conclusion de leur état de l'art que la GAT a tout intérêt à s'ouvrir à de nouveaux domaines de spécialités, tant d'un point de vue commercial qu'universitaire.

2.2.3 Données d'entrée

La GAT se distingue d'autres disciplines de traitement automatique des langues par l'extrême variabilité des données qu'un système peut accepter en saisie. [Gatt and Krahmer, 2018] reprennent la distinction entre les notions de *data-to-text* (données non-textuelles ou même non-linguistiques) et de *text-to-text* et à l'image de [Reiter and Dale, 2000], le précédent état de l'art, s'intéressent principalement aux systèmes *data-to-text*, qui sont par ailleurs en majorité. Cependant,

2. On reviendra, entre autres, sur la définition terminologique de la notion de domaine de spécialité [L'Homme, 2018] dans le chapitre 3.

3. En France, le moteur de rédaction de Syllabs (www.syllabs.com) a produit 36 000 articles résumant les résultats par commune des élections régionales et départementales de 2015 pour le journal Le Monde (www.lemonde.fr).

même en appliquant cette restriction, on peut retrouver en entrée de systèmes de génération des données numériques (structurées ou non) [Yu et al., 2007], des bases de connaissances ou même des ontologies comme on l'a vu en 2.2.1, des triplets RDF [Cimiano et al., 2013] [Gardent et al., 2017] ou d'autres formes de représentations sémantiques [Mauldin, 1984]. Enfin, certaines tâches, telles la génération de légendes et de descriptions d'images où les données sont donc visuelles [Tintarev et al., 2016] sont en plein développement.

Cette diversité a une forte influence sur la structure d'un système de génération, qui, quelle qu'elle soit, doit intégrer des étapes de traitement des données en entrée. Le modèle de représentation des données qui est choisi doit être à la fois aisément exploitable par le système et représentatif du domaine traité.

2.3 Méthodologies

2.3.1 Architectures modulaires et architectures globales

À la fin des années 1990, après une analyse des systèmes existants, Reiter et Dale suggèrent, dans [Reiter and Dale, 1997] et [Reiter and Dale, 2000], un consensus autour d'une architecture modulaire en « pipeline », constituée de trois étapes (figure 2.1). Cette architecture s'appuie notamment sur une division de la tâche de génération en six modules répartis dans ces trois étapes, sur lesquels nous reviendrons dans le chapitre 4 :

1. la planification de documents (*document planning*)
   — la détermination de contenu (*content determination*)
   — la structuration du texte (*text structuring*)
2. la planification de surface (*microplanning*)
   — l'agrégation de phrases (*sentence aggregation*)
   — la lexicalisation (*lexicalisation*)
   — la génération d'expressions référentielles (*referring expression generation*)
3. la réalisation de surface (*surface realisation*)
   — la réalisation linguistique (*linguistic realisation*)

Cette architecture, qui se veut un reflet des systèmes en activité à l'époque de sa création plutôt que d'une théorie linguistique, reste une référence à ce jour [Gatt and Krahmer, 2018], en partie parce qu'elle est facilement adaptable à un système spécifique (par exemple en regroupant certains modules avec d'autres) et parce qu'elle offre une forte souplesse de développement. Elle permet également de modéliser une distinction entre les tâches en début de pipeline, qui sont généralement liées à la question du *que dire* et donc souvent dépendantes du contexte ou système, et les tâches plus linguistiques (*comment le dire*) en fin de pipeline, qui dépendent surtout de la langue de génération.

Cependant, le coût de cette souplesse est, en autres, ce que [Meteer, 1991] décrit comme le *generation gap*, quand les décisions stratégiques du système (*que dire*) et ses décisions tactiques (*comment le dire*) entrent en conflit au fur et à mesure de la progression dans le pipeline. Pour répondre à ce type de problématiques, certains modèles d'architecture proposent des modules interconnectés [Danlos and Roussarie, 2000], mais plus complexes à mettre en place.

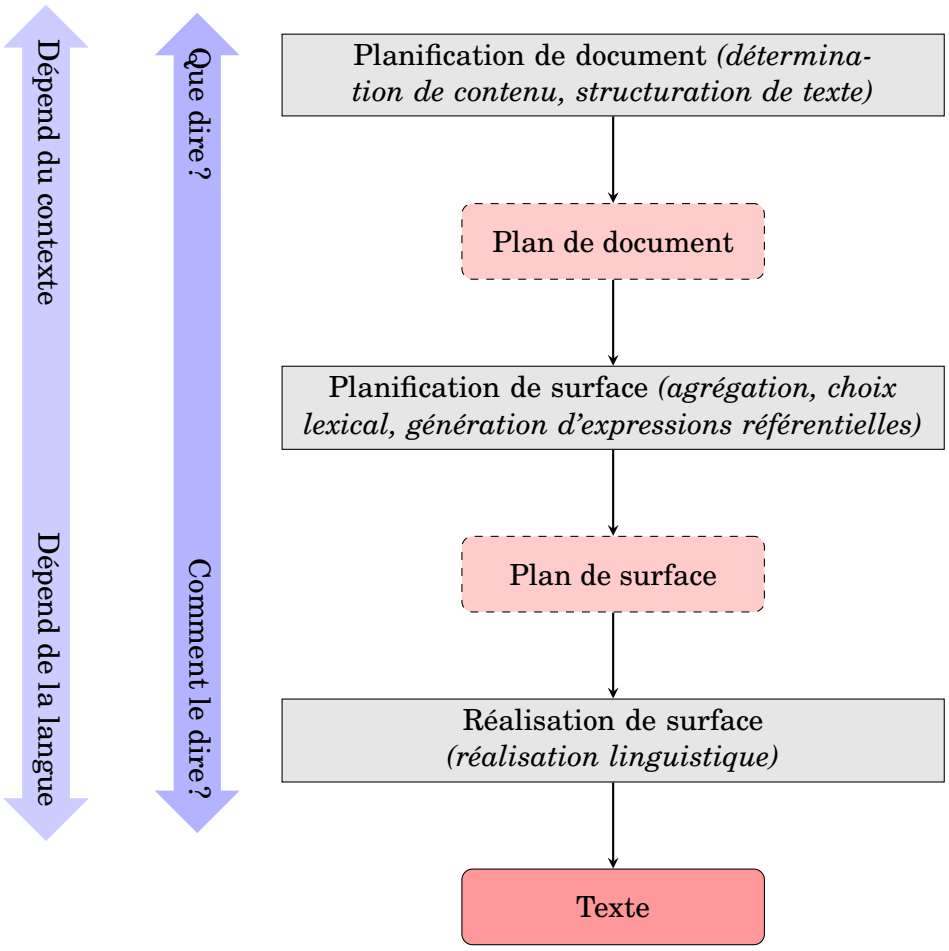


FIGURE 2.1 – Schéma d'une architecture tripartite de génération automatique de texte (adapté de [Reiter and Dale, 2000])

D'autres approches s'inspirent des méthodes de planification en intelligence artificielle, et considèrent la génération de textes comme une série d'actions planifiées ayant un but de communication, une approche qui repose en grande partie sur la notion d'acte de langage définie par Austin puis Searle (par exemple, KAMP [Appelt, 1985], qui s'appuie sur la théorie d'actes propositionnels de Searle).

Enfin, l'émergence et l'application de méthodes d'apprentissage statistique et automatique à partir de corpus a fortement favorisé l'apparition d'architectures plus globales ou intégrées, dont l'inconvénient est le fait qu'il est difficile de situer exactement les sources d'erreur dans le système.

2.3.2 Approches symboliques et approches stochastiques

Il est également possible de classifier les systèmes de GAT en fonction des différentes méthodes employées pour mettre en place les différentes architectures et modules mentionnés en 2.3.1. La méthode de réalisation linguistique la plus simple reste l'utilisation de modèles (*templates*) comme une forme de « texte à trous » où les données variables sont intégrées dans le modèle, de façon très contrôlée (les erreurs syntaxiques sont presque impossibles) mais par conséquent très rigide (les variations morphologiques peuvent être difficiles à gérer). L'appartenance de tels systèmes au

domaine de la GAT a été contestée en raison de leur apparente absence de complexité linguistique ; cependant, [van Deemter et al., 2005] minimisent l'écart entre les systèmes à base de modèles et les autres systèmes de GAT, notamment en soulignant un grand nombre d'approches hybrides, tandis que [Puzikov and Gurevych, 2018] démontrent qu'un système à base de modèles peut atteindre des résultats similaires à un système *encoder-decoder* pour certaines tâches.

Historiquement, une forte majorité des systèmes de GAT (y compris ceux mentionnés en 2.2) implémente des règles manuelles pour effectuer les choix nécessaires à la génération [4], et ce à tout ou partie des étapes mentionnées dans la figure 2.1. L'implémentation de ces règles est largement simplifiée si le système de génération est concentré sur un domaine de spécialité, ce qui limite le vocabulaire employé ainsi que la variabilité des situations rencontrées. Par ailleurs, la génération à base de règles offre un degré de contrôle sur le texte généré (jusqu'au niveau des mots) qui peut s'avérer extrêmement utile, surtout si l'on envisage de personnaliser le texte généré en fonction du public visé [Reiter et al., 2005]. Cependant, la rédaction desdites règles demande un travail humain non négligeable, et s'adapte difficilement aux domaines d'application qui requièrent une grande variabilité linguistique.

Depuis quelques années, la GAT, comme tous les domaines de traitement automatique des langues, assiste à la déferlante de méthodes stochastiques et statistiques de génération, dont [Langkilde and Knight, 1998] furent précurseurs. Plus récemment, les systèmes de génération neuronale font l'objet d'une attention particulière, notamment pour la génération de textes courts [Xie, 2017] ou de dialogues d'agents conversationnels [Gao et al., 2019]. Ce type a néanmoins le défaut de nécessiter un corpus robuste d'apprentissage, qu'il peut être long et coûteux de générer [Dušek et al., 2020] quand il n'est pas déjà disponible (ce qui est très souvent le cas en GAT [5] [Novikova et al., 2017]).

2.3.3 Intersections entre ces différents modèles

La raison pour laquelle nombre de systèmes modulaires ont recours à des méthodes symboliques plutôt que stochastiques est principalement historique, puisque les architectures modulaires sont de loin les plus anciennes.

Cependant, rien n'empêche un système modulaire d'employer des méthodes d'apprentissage statistique et d'apprentissage profond, par exemple pour la sélection de contenus [van der Lee et al., 2018], ou la génération d'expressions référentielles [Castro Ferreira et al., 2018], ou encore d'avoir recours à une architecture hybride où certains modules sont traités de façon symbolique et d'autres via l'apprentissage automatique [Moryossef et al., 2019]. De même, il n'est pas nécessairement évident que les approches stochastiques bénéficient toujours d'être implémentées dans une architecture intégrée plutôt que modulaire [Castro Ferreira et al., 2019].

Autrement dit, les architectures définies en 2.3.1 et les approches décrites en 2.3.2 ne correspondent pas directement les unes aux autres : elles peuvent se croiser de façon orthogonale (figure 2.2) en fonction des tâches à effectuer, des besoins et contraintes du système en question et des avantages et inconvénients de chaque méthode.

4. Cette tendance est également observable chez les fournisseurs de système de GAT commerciaux.
5. En particulier pour les systèmes ayant une visée commerciale ou industrielle, où l'objectif est souvent de générer un tel corpus.

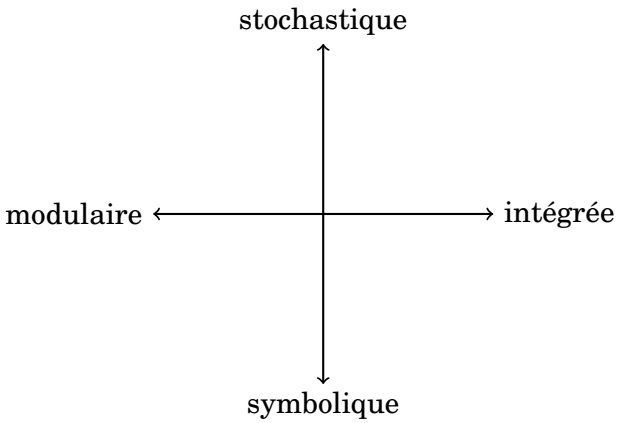


FIGURE 2.2 – Illustration des possibilités d'intersections entre différentes approches et architectures de génération

2.4 Évaluation(s)

2.4.1 Difficultés spécifiques à la GAT

Comme le notent entre autres [Gatt and Krahmer, 2018], la question de la méthodologie de l'évaluation des systèmes de GAT reste complexe et controversée : la diversité des systèmes de génération se reflète dans la diversité des méthodes d'évaluation proposées et implémentées dans la littérature. Ainsi, la variabilité de format et de complexité des données mentionnée en 2.2.3 que peut exploiter un système de GAT rend la comparaison de différents systèmes difficile sans la présence d'une forme de standard [6].

D'autre part, un système de GAT produit des textes, qui, presque par définition sont extrêmement variables et multiples. Un système est généralement capable de générer des phrases différentes (tant d'un point de vue syntaxique que lexical) pour exprimer la même idée, sans qu'il ne soit possible de désigner l'une d'entre elles comme un standard linguistique : [Mellish and Dale, 1998] soulignent l'importance de la notion de choix dans les questions de génération, et suggèrent que la distinction entre le *que dire* et le *comment dire* est principalement une tentative de classification entre les choix stratégiques et les choix tactiques qu'effectue le système.

Évaluer le résultat d'un système de génération consiste donc à évaluer une série de décisions prises par le système, ce qui nous renvoie aux problématiques évoquées en section 2.3 : faut-il évaluer les différents modules d'un système de façon indépendante ou se concentrer sur l'évaluation du système dans son ensemble ? Dans la pratique, [Hastie and Belz, 2014] (cf. table 2.1) et [Gkatzia and Mahamood, 2015] démontrent dans leurs vues d'ensemble l'éclatement des méthodes d'évaluation employées dans la littérature.

Cependant, avant même d'entamer un processus d'évaluation de la performance d'un système, il convient d'abord de déterminer ce qui est évalué exactement dans cette performance. La diversité des formes d'évaluation possibles reflète la diversité des usages et des fonctions visés décrite en 2.2.1 : il faut au minimum effectuer une

6. Ceci explique en partie pourquoi il existe un plus grand consensus autour de métriques d'évaluation comme BLEU ou ROUGE dans des domaines tels que la traduction automatique ou la récapitulation de textes.

Comment évaluer un système ?		
Type d'évaluation	Mesure	Description
intrinsèque	qualité des textes (*output quality*)	critères plus ou moins dépendants du contexte comme la lisibilité ou l'exactitude (évaluation humaine ou métrique)
intrinsèque	satisfaction de l'utilisateur (*user like*)	questions posées directement aux utilisateurs du système
extrinsèque	efficacité du système (*user task success*)	exécution réussie de la tâche (ex. un dialogue) pour laquelle le système a été conçu
extrinsèque	objectif global du système (*system purpose success*)	dépend de l'objectif du système (ex. questions aux utilisateurs sur leur ressenti sur le système dans son ensemble)

TABLE 2.1 – Extrait de la classification des méthodes d'évaluation de systèmes de génération de [Hastie and Belz, 2014]

distinction entre les systèmes dont l'objectif est d'informer l'utilisateur et ceux qui se doivent de le convaincre, voire de le persuader. Ainsi, l'évaluation est à la fois conceptuelle (d'un point de vue informatique) et utilitaire (du point de vue de l'utilisateur final).

2.4.2 Qu'évalue-t-on ?

Intuitivement, il est assez simple de déduire depuis la division en *que dire* et *comment dire* deux problématiques d'évaluation :

— le texte produit est-il vrai (c'est-à-dire une représentation correcte des données d'entrée) ?
— le texte produit est-il lisible (correct d'un point de vue grammatical et linguistique) ?

Données d'entrée	name[Cotto], eatType[coffee shop], near[The Bakers]
Phrase A	Cotto is a coffee shop located near The Bakers.
Phrase B	Cotto is a coffee shop **with a low price range**. It is located near The Bakers.
Phrase C	Cotto is **a pub** near The Bakers.

Les phrases B et C contiennent des informations (en gras) qui n'existent pas dans la représentation de données d'entrée.

TABLE 2.2 – Hallucinations de systèmes de GAT neuronaux ([Dušek et al., 2020], repris par [Reiter, 2018a])

Un texte généré peut en effet être faux mais parfaitement lisible (par exemple suite à un problème d'hallucination pour les systèmes utilisant l'apprentissage au-

tomatique, cf. table 2.2) ou au contraire être grammaticalement incorrect mais correspondre aux informations fournies au système (un problème plus courant pour les systèmes à base de règles).

[Mellish and Dale, 1998, p.5] présentent trois catégories plus profondes d'évaluation en GAT :

— l'évaluation de la ou des théorie(s) scientifique(s) sous-jacente(s) (notamment les théories linguistiques) du système
— l'évaluation du système lui-même (par exemple par rapport à d'autres systèmes)
— l'évaluation du potentiel d'utilisation dans un environnement précis (pourquoi utiliser un système de GAT plutôt qu'une présentation graphique, par exemple)

Enfin, [Gatt and Krahmer, 2018, p.123] reprennent la distinction méthodologique faite par [Spärck Jones and Galliers, 1995] entre les méthodes d'évaluation intrinsèques et extrinsèques, qui est particulièrement pertinente ici :

— une évaluation intrinsèque mesure la performance du système ou d'un de ses éléments sur un critère qui n'est pas directement lié à l'objectif final du système : l'évaluation de la lisibilité des textes produits par SaferDrive (2.2.1) est intrinsèque par rapport à l'objectif final de réduire le nombre d'infractions mineures des conducteurs utilisant le système
— une évaluation extrinsèque, au contraire, évalue la performance du système par rapport à cet objectif global : l'évaluation du système STOP, également mentionné en 2.2.1, était extrinsèque car elle consistait à vérifier si les utilisateurs du système parvenaient à baisser leur consommation de cigarettes de façon significative

L'intersection de ces différentes méthodologies explique l'existence d'un grand nombre de critères d'évaluation, sur lesquels on reviendra plus en détail dans le chapitre 6.

2.4.3 Évaluation métrique, évaluation humaine

Comme le soulignent [Gatt and Krahmer, 2018, pp.125-127], l'évaluation de systèmes de GAT via des mesures utilisant des corpus telles que BLEU (BiLingual Evaluation Understudy [Papineni et al., 2002] : métrique à l'origine conçue pour évaluer des systèmes de traduction automatique, qui calcule la « précision » d'un traduction automatique en comparant ses n-grammes à ceux d'une traduction de référence) ou ROUGE (Recall-Oriented Understudy for Gisting Evaluation [Lin, 2004] : utilisée en récapitulation de textes, ROUGE mesure également des chevauchements de n-grammes s'intéresse plus spécifiquement au « rappel » du résumé produit par rapport au texte de référence).

Ces mesures sont généralement employées pour évaluer la qualité linguistique du texte et sa réalisation plutôt que sa véracité [Reiter and Belz, 2009], ou parfois pour évaluer une tâche intermédiaire comme la génération d'expressions référentielles [Belz and Gatt, 2008].

Cependant, l'usage de ces mesures, dont la grande majorité n'a pas été conçue pour évaluer des tâches de génération, reste fortement débattu : d'une part, ces mesures ne sont pas nécessairement calibrées pour traiter les spécificités de la GAT (comme l'extrême variabilité des textes produits, cf. table 2.3). D'autre part,

d'après [Reiter, 2018b], les corrélations entre un score BLEU et une évaluation humaine dans le domaine de la GAT sont très variables ; dans le cadre de la génération d'expressions référentielles, [Belz and Gatt, 2008] soulignent que le critère d'une mesure comme BLEU (*humanlikeness*, soit la correspondance avec un texte rédigé par un humain, calculée à base de n-grammes dans le cas de BLEU), est une mesure intrinsèque qui n'est pas validée extrinsèquement.

Contexte	Why don't we go see a movie ?
Réponse A	Nah, I hate that stuff, let's do something active.
Réponse B	Oh sure ! Heard the film about Turing is out !

TABLE 2.3 – Variabilité sémantique et syntaxique entre deux réponses pertinentes pour un système de dialogue (adapté de [Liu et al., 2016])

On peut donc décider plutôt d'avoir recours à une évaluation humaine, ce qui ne nous absout cependant pas de devoir définir clairement les critères et protocoles d'évaluation employés, notamment dans le cas d'une évaluation extrinsèque où l'on cherche à évaluer le système en action : il faut être capable d'interpréter ces résultats non-linguistiques pour en tirer des conclusions sur la performance du système [Reiter et al., 2003]. D'autre part, si l'on choisit plutôt de demander à des experts du domaine traité d'évaluer les textes, il faut définir des critères précis permettant une évaluation efficace [Hunter et al., 2012].

2.5 Conclusion

Au terme de ce tour d'horizon de l'état de l'art, nous pouvons retenir plusieurs éléments essentiels pour la définition de notre problématique de génération. D'abord, la création d'un système de GAT sera fortement influencé d'une part par la visée du discours qui est généré, et d'autre part par le domaine de spécialité auquel appartiendra ce discours. Il est donc nécessaire d'avoir une maîtrise du domaine couvert ainsi que du langage spécifique à ce domaine, mais aussi de la rhétorique qui devra être maîtrisée pour produire du texte qui correspond aux besoins de l'utilisateur du système. C'est cette connaissance qui permettra ensuite de choisir une architecture et une approche adaptées en fonction des contraintes du système. Enfin, il faut être capable de définir un protocole d'évaluation qui permettra d'obtenir des résultats pertinents sur la performance du système en fonction de tous ces éléments.

CHAPITRE 3

DESCRIPTION DU PROJET

Sommaire

3.1 Introduction

Nous avons établi que notre projet consisterait à générer des rapports courts d'analyse financière explicitant la situation financière des entreprises suivies par Exane, à partir des données principalement numériques qu'Exane recueille sur ces entreprises. Nous avons également relevé dans notre état de l'art de la génération automatique de textes que les pierres d'achoppement d'un système de GAT sont très souvent le traitement des données non-textuelles à exploiter, la définition globale de l'architecture du système et l'évaluation des résultats.

Dans ce chapitre, nous reprenons le contexte décrit dans le chapitre 1 et l'état de l'art défini dans le chapitre 2 afin de mettre en relief certains aspects spécifiques à notre projet de génération et complexifier notre problématique.

Nous effectuons d'abord un rapide tour d'horizon de la recherche en TAL dans le domaine financier, et des quelques projets de génération existant dans le domaine financier. Nous pourrons donc ensuite préciser, pour orienter nos expérimentations et l'évaluation de notre système (chapitres 4, 5 et 6), des questions spécifiques qui ressortent de l'intersection entre notre domaine d'études (la GAT) et notre domaine de spécialité (la finance).

Ceci nous permettra de récapituler toutes les informations acquises lors de la description des contextes de recherche et d'industrie dans lesquels nous avons travaillé, avant de décrire notre système en détail et de commencer nos expérimentations.

3.2 La finance, un domaine de spécialité

3.2.1 Applications existantes du TAL

[Fisher et al., 2016] et [Xing et al., 2018] présentent l'intersection entre le traitement automatique des langues et le domaine financier, principalement autour de la question des prévisions de performance financière. Cependant, tous deux se concentrent principalement sur l'exploitation de données textuelles pour faciliter différentes tâches comme la prédiction de mouvements des marchés (*financial forecasting*).

Pour ce faire, [Xing et al., 2018] décrivent l'application de tâches comme l'analyse de sentiments, l'extraction d'événements et la fouille de textes sémantique. Les auteurs décrivent différents types de textes exploitables pour des travaux de TAL, comme les rapports financiers, les déclarations de résultats des entreprises, les journaux spécialisés, les réseaux sociaux... Quant à [Fisher et al., 2016], les auteurs établissent plutôt une distinction dans l'usage du TAL dans la finance entre les œuvres de classification (plus spécifiquement, la génération de taxonomies et thésaurus) et de prédiction (détection de fraude, prédiction d'activité du marché ou de résultats de certaines entreprises). Est aussi mentionnée, dans le cadre de la détection de fraude, la notion de lisibilité des textes financiers complexes comme les bilans financiers des entreprises.

Cependant, la génération automatique de rapports d'analyse financière s'inscrit en quelque sorte à l'opposé d'un grand nombre de ces problématiques de recherche : ici, l'idée n'est pas de remplacer l'expertise humaine d'un.e analyste financi·er·ère mais plutôt d'exprimer sous forme textuelle une partie du travail de recherche effectué par cet·te expert·e. Une grande partie des problèmes rencontrés dans d'autres tâches n'a donc pas lieu d'être, tandis que d'autres questions restent uniques à la problématique de génération.

3.2.2 Systèmes de génération

Parmi les systèmes de génération centrés sur le domaine financier, comme ceux mentionnés en 2.2.2, [Kukich, 1983] fait figure de précurseure. Kukich souligne l'importance des connaissances liées au domaine de spécialité et la possibilité de s'appuyer sur des corpus de textes déjà rédigés par des spécialistes du domaine. Elle reprend également la notion de « core [market] sentences » [1] pour définir les phrases ou éléments de texte qui peuvent être produits à partir de la base de données fournie, sans avoir accès et sans faire référence à des événements externes à cette base. Cette distinction entre l'« univers » du système et les « informations stratégiques » qui en sont exclues se retrouvera tout autant dans notre travail, dans la mesure où des analystes font amplement référence à ces informations dans le cadre de leur travail de recherche [2].

Nous pouvons aussi citer [Murakami et al., 2017], dont le travail fut repris par [Aoki et al., 2018]. Les auteurs décrivent un modèle neuronal de type *encoder-decoder* qui utilise des données sur le cours du Nikkei 225 (principal indice boursier de la bourse de Tokyo) pour générer des commentaires sur l'évolution de ce cours. Ils

1. Ici, « market » fait référence au domaine d'application des marchés financiers.
2. Par exemple, l'arrestation du PDG de Renault Carlos Ghosn au Japon en 2018, ou la vague de défaillances d'avions Boeing 737 MAX depuis 2018 sont des événements externes à la grille financière des entreprises concernées qui ont cependant une forte influence sur celle-ci.

mettent en avant les difficultés rencontrées pour encoder des données numériques sous formes de séries temporelles. Ce problème est directement lié à la question de l'expression de ce type d'information sous forme linguistique, d'une part parce qu'elle complexifie le traitement des données préalable à la génération. D'autre part, le vocabulaire employé pour décrire une évolution temporelle peut être compris de façons diverses et subjectives par différents lecteurs.

[Murakami et al., 2017] utilisent notamment l'exemple du terme « rebound » [3], qui peut être employé par un système uniquement dans la mesure où ce dernier dispose d'informations sur une période entière. Il peut ensuite comparer ces informations avec celles de la période d'écriture (« delivery time »), qui peut elle-même être plus ou moins longue. Cette problématique s'associe à celle notée par [Aoki et al., 2018] de la diversité des sources d'information qui doivent être prises en compte (par exemple, le cours du Nikkei sera influencé par ceux d'autres indices boursiers dans le monde).

3.2.3 Difficultés d'accès aux données

Un problème récurrent décrit dans le cadre d'applications du TAL à des questions financières est celui de l'accès à des données, quel que soit leur format. Il peut être difficile de constituer des corpus ou des bases de données assez conséquents pour être exploités, dans la mesure où un grande partie de ces données peut être au moins propriétaires, si ce n'est confidentielles. Dans tous les cas, le recueil des données peut être difficile : d'après [Xing et al., 2018], la concentration du nombre de travaux de TAL sur des données liées aux marchés financiers (par rapport à par exemple, la prédiction de l'inflation ou l'évaluation des risques client) est en partie liée au fait que ces données (les cours de bourses) sont bien plus accessibles pour un travail de recherche. [Smiley et al., 2016], quant à eux, ont recours à un corpus d'articles de presse pour leur travail sur le choix lexical.

Dans le cadre de notre stage, nous avions accès aux données confidentielles recueillies et générées par Exane, ce qui nous a ouvert des perspectives nouvelles : nous disposons en effet d'une base de données extrêmement complète pour construire un système de génération. Ces données sont aussi relativement différentes de celles qui sont souvent utilisées dans des projets de génération dans la mesure où une partie d'entre elles correspond à des estimations futures. Par ailleurs, par rapport à [Murakami et al., 2017] et [Aoki et al., 2018], nous allons générer des rapports d'analyse spécifiques à des entreprises, plutôt que des rapports sur l'évolution du cours de la bourse.

Une autre conséquence du statut particulier des données sur lesquelles nous allons travailler est l'absence d'un corpus de textes qui soient similaires aux textes que nous allons générer. En effet, parmi les types de données textuelles aisément accessibles décrits par [Xing et al., 2018], aucun ne correspond véritablement de par ses caractéristiques textuelles (longueur, ton, etc.) à notre projet de rédaction de rapports courts d'analyse financière. En interne à Exane, nous avions accès à un corpus de rapports d'analyse financière complets extraits de Cube (cf. section 4.2). Ce dernier nous a été utile dans un certain nombre d'expériences intermédiaires et a pu nous servir de référence stylistique pour analyser la rédaction d'un rapport. Néanmoins, ces rapports divergent trop des textes générés, d'une part par leur longueur et par leur

3. Dans le domaine financier, « rebound » s'emploie pour définir la remontée d'une valeur ou d'un indicateur après une période plus ou moins prolongée de baisse. Pour l'utiliser, un système doit donc pouvoir appréhender la notion d'une baisse sur une période prolongée, mise en contraste avec une remontée brusque plus récente.

niveau de précision, et d'autre part par la variabilité de leur structure (cf. 1.2.2) pour pouvoir être utilisés directement comme un « gold standard », par exemple pour une évaluation métrique comme celles mentionnées en 2.4.3.

3.3 Spécificités du projet

3.3.1 Langue spécialisée ou sous-langage ?

> After three following years of decrease, we anticipate that EPS will rebound by 5.3% in 2019.

FIGURE 3.1 – Phrase extraite d'un rapport d'analyse généré

La phrase présentée dans la figure 3.1, extraite d'un des rapports d'analyse générés par notre système, est un bon exemple du type de langage spécialisé qui est employé dans ces rapports. Si elle est parfaitement compréhensible, en contexte, pour un·e analyste ou un·e lecteur·ice régulier·ère de ce type de documents, elle peut s'avérer bien plus obscure pour un·e non-expert·e. Comment qualifier le type de langage utilisé ici, et comment le modéliser efficacement dans notre système de génération ?

[Lopez, 2013, pp.41-58] donne une définition générale d'une langue spécialisée comme étant

> un ensemble constitué de nombreux aspects de la langue générale et de certains aspects spécifiques à un domaine de connaissances (dont le plus marquant est sans doute une terminologie propre à ce dernier)

Elle effectue ensuite une distinction entre les *langues de spécialité* ou *langues spécialisées* (ou « languages for specific purposes ») et les *sous-langages*, qu'elle considère comme deux types d'analyse différents de la même notion (un type de production langagière distinct de la langue générale et servant à exprimer des connaissances sur un domaine de spécialité). Nous pouvons retenir des éléments intéressants de ces deux approches pour nous aider dans notre travail de génération.

Comme le précise la définition précédente de Lopez, une langue spécialisée se définit par rapport à la langue générale ; pour répondre aux besoins de communication des personnes qui l'emploient pour traiter d'un sujet spécialisé, elle a tendance à privilégier des formes d'expression :

— univoques, c'est-à-dire monosémiques et non-ambigües
— concises, si ce n'est économes
— neutres, ou du moins tendant vers une forme d'objectivité

Ces caractéristiques, que l'on retrouve à divers degrés dans des phrases comme 3.1 et qui ont souvent dirigé nos choix de génération, s'expriment notamment par l'usage d'une terminologie spécifique, qui induit donc la définition de termes. [L'Homme, 2018] définit les termes au sens terminologique comme

> des unités lexicales dont le sens est envisagé par rapport à un domaine de spécialité, c'est-à-dire un domaine de la connaissance humaine, souvent associé à une activité socio-professionnelle

Dans notre phrase d'exemple, on peut repérer des termes comme « decrease » (baisser), une unité lexicale qui prend un sens spécifique dans le domaine financier

par rapport à sa signification dans la langue générale, ou encore « EPS », une abréviation de *earnings per share*, un terme suffisamment spécifique au domaine pour être défini dans des glossaires spécialisés comme celui du Nasdaq[4] ou le glossaire interne d'Exane. Le terme « EPS », comme beaucoup d'autres termes désignant des indicateurs de performance, est presque tout le temps abrégé dans les textes spécialisés, répondant aussi au besoin de concision des langues de spécialité. Seul·e un·e lecteur·ice ayant une connaissance précise de ces termes et de leur utilisation est à même de décider si la phrase 3.1 est correcte, tant linguistiquement que conceptuellement.

Une langue spécialisée est en effet définie par ses utilisateurs. [Lopez, 2013] reprend une distinction entre trois types de locuteurs d'une langue spécialisée, par ordre descendant de maîtrise de celle-ci : les experts, les semi-experts et les non-experts. Nous devions, en tant que non-experte travaillant avec des expert·e·s, atteindre un niveau de semi-expertise qui nous permettrait de percevoir l'intersection entre leurs connaissances linguistiques (comme l'utilisation de certaines structures ou certains termes) et leurs connaissances conceptuelles du domaine. C'est en effet en croisant ces connaissances que ces experts étaient capables de décider de la justesse et de la précision d'une tournure ou d'un terme. Cependant, dans la pratique, il peut être difficile pour des utilisateurs réguliers d'une langue spécialisée de percevoir quels en sont les termes spécifiques, tant ils sont habitués à l'emploi de cette langue. Dans ce cas, une perspective externe peut s'avérer utile en ce qu'elle révèle ce qui peut sembler évident à des spécialistes, ou, dans certains cas, des désaccords entre spécialistes sur la signification de termes censés être « évidents »[5].

	Variations morphosyntaxiques		
	Verbe	Nom	Adjectif
Synonymes	increase	increase	increasing
	grow	growth	growing
	go up	ø	ø

TABLE 3.1 – Variations morpho-syntaxiques et synonymie autour du terme « increase »

Pour illustrer ce point plus en détails, nous reprenons deux débats ayant eu lieu au cours du stage autour de l'expression lexicale dans les textes générés de deux termes relatifs dans notre exemple, « increase » (table 3.1) et « expect » (table 3.2).

Dans la table 3.1, nous reprenons la distinction entre synonymie et variation terminologique de [L'Homme, 2018][6] pour présenter les synonymes ayant été approuvés pour exprimer la notion de « monter » dans notre domaine de spécialité (« augmenter, croître en quantité, en intensité, en valeur ; atteindre telle ou telle valeur »)[7] dans différentes catégories grammaticales. Notre objectif de rédaction de textes suffisamment variables pour être agréables à lire doit être conjugué avec le besoin d'univocité et de précision d'une langue de spécialité : ainsi, les informations présentées dans la

4. www.nasdaq.com/glossary/e/earnings-per-share.

5. [Reiter et al., 2005] étudient l'influence de l'idiolecte de différents experts sur leurs choix lexicaux dans le cadre de la GAT. Nous revenons également sur cette question dans le chapitre 5.

6. Dans cet exemple, nous réduisons la notion de variation terminologique aux simples variations morpho-syntaxiques.

7. www.cnrtl.fr/definition/monter, définition I.B.6).

Verbe (infinitif)	+ proposition relative	+ groupe nominal	+ complément en TO
expect	**we expect that EPS will decrease*	we expect a decrease of the EPS	we expect the EPS to decrease
anticipate	we anticipate that EPS will decrease	we anticipate a decrease of the EPS	**we anticipate the EPS to decrease*
forecast	we forecast that EPS will decrease	we forecast a decrease of the EPS	**we forecast the EPS to decrease*

Les formes grammaticalement incorrectes sont notées en italique et précédées de *.

TABLE 3.2 – Variations syntaxiques des synonymes du terme « expect »

table 3.1 doivent être modélisées dans notre moteur de rédaction (cf. section 4.4.2) pour pouvoir être intégrées dans les mêmes structures syntaxiques.

Enfin, ce problème d'ambiguïté se retrouve également, via la notion de « neutralité » mentionnée plus haut, dans la table 3.2. La notion de prévision est essentielle pour le domaine dans lequel nous travaillons puisqu'elle désigne le cœur du travail d'analyse : à ce titre, les termes choisis pour la désigner doivent présenter un nombre limité d'ambiguités. Par exemple, la forme « predict » (prédire) a été rejetée dans la mesure où il a été jugé par nos experts qu'elle introduisait une connotation subjective indésirable dans la prévision. Les formes choisies (*expect*, *anticipate* et *forecast*) ne présentaient pas à leurs yeux ce problème ; cependant, elles ne s'emploient pas toutes dans les mêmes constructions grammaticales, un problème que nous devrons également prendre en compte lors de la modélisation.

La notion de sous-langage redevient ici utile pour définir certains aspects de notre langage de spécialité. Fondée sur l'approche de sous-systèmes mathématiques proposée par Zellig Harris, la théorie des sous-langages a déjà été employée dans la création de systèmes de génération [Kittredge et al., 1994]. Les sous-langages sont souvent délimités par des restriction syntaxiques, sémantiques, ou lexicales par rapport à la langue générale. Nous redéfinissons parmi les critères restrictifs définis par [Friedman et al., 2002], qui ont travaillé sur la définition de sous-langages dans le domaine biomédical, et par [Lopez, 2013, p.50], ceux qui nous semblent les plus applicables aux questions de génération :

— le vocabulaire et la terminologie limités : [Danlos and Roussarie, 2000] notent dans leur état de l'art des questions de génération que les limites lexicales intrinsèques à un domaine de spécialité facilitent le travail de génération en réduisant à la fois le vocabulaire employé mais également les risques d'homonymie
— les motifs paraphrastiques (*paraphrastic patterns*) : pendant la conception des règles de rédaction du système, nous avons pu repérer des modèles de co-occurrences de certains termes pour transmettre un type d'information « fixe » (cf. table 3.3). L'intérêt de la notion de motif paraphrastique, comme le soulignent [Friedman et al., 2002], est qu'elle permet d'établir une équivalence sémantique entre différentes constructions grammaticales, qui, dans le cadre

d'un domaine de spécialité, sont employées de façon interchangeable. On peut donc envisager d'utiliser des motifs paraphrastiques pour conjuguer l'impératif d'univocité d'une langue spécialité et la variabilité linguistique offerte par une système de GAT.
— le non-respect de règles de grammaire « générale » dans certains cas spécifiques : à titre d'exemple, l'usage de l'article défini avec des termes comme « EPS », comme dans la table 3.2 (« a decrease of **the** EPS », mais « EPS will decrease »)
— la structure plus ou moins stricte des documents produits, un critère qui se rattache directement à la notion de structuration des textes

Partie du motif	Spécifications
EXPECTATION	La table 3.2 présente différents synonymes du terme « expect »
KPI (Key Performance Indicator)	Un indicateur de performance dans un domaine de spécialité comme la finance, comme par exemple, l'EPS
FLUCTUATION	Un terme comme INCREASE, DECREASE ou STAGNATE, qui lui-même présente des variations terminologiques (cf. table 3.1)

Ici, le motif décrit permet de décrire les phrases générées dans la table 3.2, en omettant deux éléments (une référence temporelle ainsi qu'une valeur spécifique qualifiant la fluctuation) présents dans la phrase 3.1.

TABLE 3.3 – Description d'un motif paraphrastique dans le contexte du sous-langage de l'analyse financière

3.3.2 Représentation temporelle des données

Afin de pouvoir analyser les données contenues dans une grille financière, nous implémenterons la même représentation temporelle (à l'exception de quelques changements mineurs) de ces données que celle de la grille (figure 3.2 et table 3.4). On distingue de manière globale trois périodes :

— le passé, c'est-à-dire les données publiées et « confirmées » par l'entreprise (divisible en plusieurs périodes, cf. figure 3.2)
— le présent, c'est-à-dire le moment auquel le texte est rédigé
— le futur, c'est-à-dire la période des résultats estimés (le nombre d'années peut différer en fonction de la situation de l'entreprise ou de son secteur)

Par définition, le moment de rédaction se situe toujours entre la dernière déclaration des résultats d'une entreprise (par exemple, en septembre 2018) et la suivante (en septembre 2019). Dans ce cas, si le système générait un rapport en juillet ou en août 2019, il considérerait, dans sa représentation de données, 2019 comme la première année estimée (FE), en raison de l'absence de résultats publiés. En octobre ou novembre, après la publication des résultats, 2019 serait désormais la dernière année publiée (LP) pour le système.

Ce décalage entre l'année civile et les douze mois séparant deux déclarations de résultats peut causer des difficultés dans la mesure où différentes entreprises publient leurs résultats à des moments différents de l'année : comparer une entreprise

qui les publie en début d'année avec une qui les publie en fin d'année demande un prétraitement des données comparables. [Portet et al., 2009] mentionnent plusieurs méthodes d'abstraction temporelle pour pouvoir détecter et décrire des séquences d'événements efficacement, comme le séquençage d'événements répétitifs, la fusion ou la transformation de plusieurs événements en un seul.

D'un point de vue sémantique, il devient également plus difficile de déterminer les références temporelles des textes que le système va générer pour s'assurer de la compréhension du lecteur : on note qu'une phrase comme celle présentée dans la figure 3.1 contient deux références temporelles explicites (« after three following years of decrease » et « 2019 ») qui la rendent compréhensible.

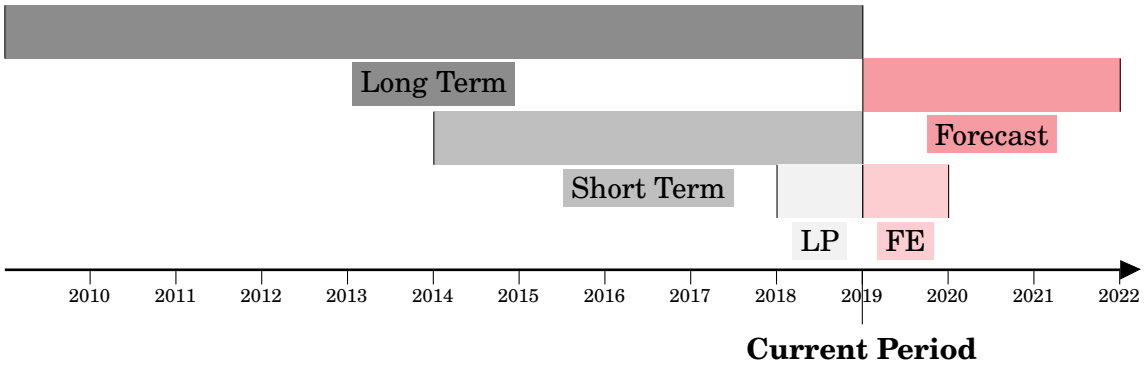


FIGURE 3.2 – Visualisation des termes employés pour décrire les périodes d'évolution d'une valeur (cf. table 3.4) en choisissant 2019 comme année courante

Terme	Définition	Exemple
Last Published (LP)	Année de publication des derniers résultats connus	2018
Current Period	Période de rédaction du rapport (le « delivery time » mentionné en 3.2.2)	2019
First Estimated (FE)	Première année d'estimation des résultats (soit l'année suivant l'année LP)	2019
Short Term (ST)	Résultats passés sur le *court terme* (entre trois et cinq ans)	2014–2019
Long Term (LT)	Résultats passés sur le *long terme* (entre cinq et dix ans)	2009–2019
Forecast	Nombre d'années pour lesquelles des résultats estimés sont fournis (entre deux et quatre ans)	2019–2022

TABLE 3.4 – Définition des termes employés pour décrire les périodes d'évolution d'une valeur (cf. figure 3.2) en choisissant 2019 comme année courante

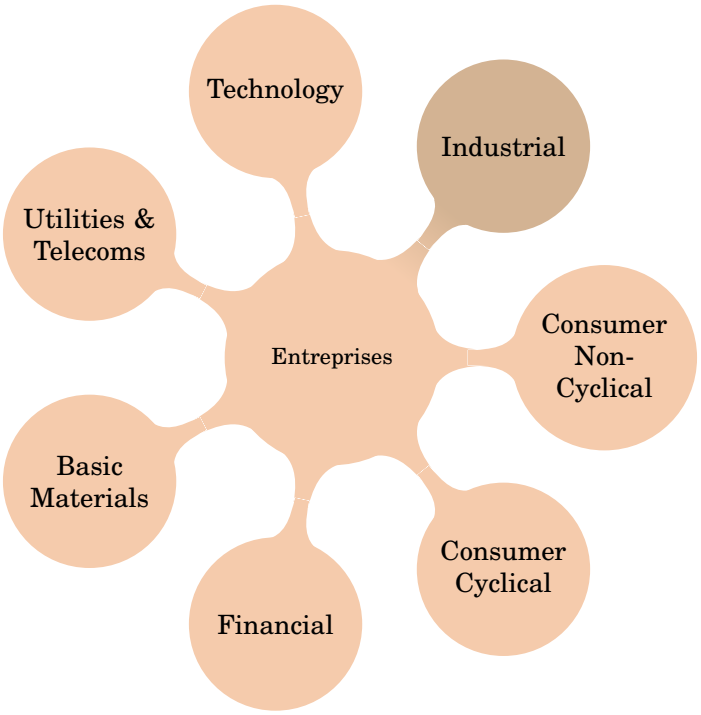


FIGURE 3.3 – Premier niveau de la classification Exane

3.3.3 Hiérarchisation de la représentation des données

Nous allons devoir générer des textes à partir de données numériques dont la représentation est complexe en raison des différents niveaux de hiérarchisation qui régissent leur organisation. Nous implémentons une partie de ces hiérarchies dans notre modélisation de données XML qui est ensuite reprise par le moteur de génération Yseop (cf. chapitre 4). Cependant, nous les présentons ici car elles influencent non seulement la représentation informatique des données mais également la représentation linguistique et sémantique des informations exprimées dans le texte (tout comme la représentation temporelle de ces données décrite en 3.3.2 a une influence sur les références temporelles des évènements décrits dans les textes générés).

Concrètement, on peut classifier une entreprise dans l'univers Exane selon de multiples critères :

— leur pays d'origine (en effectuant par ailleurs une distinction entre les États-Unis, les pays européens et potentiellement le Royaume-Uni dans le futur)
— leur secteur d'industrie, qui est lui-même divisé en secteurs et sous-secteurs
— leur taille : en fonction de leur capitalisation boursière, on distingue des entreprises *small caps*, *mid caps*, et *large caps* [8]
— l'équipe d'analystes qui est responsable du suivi du secteur de l'entreprise, et l'analyste (ou les analystes) en charge du suivi de l'entreprise

Plusieurs types de classifications et taxonomies (généralement propriétaires) du domaine financier sont employées par les acteurs du domaine en fonction de leurs besoins commerciaux, comme par exemple le Global Industry Classification Standard® (GICS), maintenu par Morgan Stanley Capital International et Standard

8. *Small capitalization*, *mid capitalization*, et *large capitalization*. Les small caps et les mid caps peuvent être regroupées dans nombre de nos cas d'analyse.

On choisit le ratio à partir duquel on veut comparer les entreprises (début de la liste de ratios à droite), puis on trie les entreprises par peers, market cap, ou pays (section « Group Stocks »), ou en sélectionnant une période.

FIGURE 3.4 – Extrait du système de réglage des paramètres de comparaison d'entreprises (Exane)

& Poor's [9]. La figure 3.3 présente le premier niveau de la classification propriétaire (en sept types d'industrie) développée par Exane ; la figure 3.5 présente les détails de classification d'un seul type d'industrie en secteurs et en sous-secteurs (Industrial), pour des raisons de lisibilité (cf. annexe A.3).

Il est donc possible de comparer un ratio ou un indicateur de performance d'une entreprise sur plusieurs années et par rapport à d'autres entreprises du même secteur. Ces entreprises peuvent elles-mêmes être regroupées par sous-secteurs (auquel cas on les qualifie de *peers* [10]), par taille, ou par pays (figure 3.4). Si nous souhaitons modéliser ce type de comparaisons via un moteur de génération de texte, nous devrons soit modéliser l'intégralité de ces données de comparaison pour chaque entreprise, ce qui ralentira énormément le processus informatique de génération, soit effectuer un travail d'ingénierie de la connaissance pour savoir quelles configurations permettent d'extraire les informations les plus essentielles.

9. www.msci.com/gics.

10. L'intérêt de la délimitation par sous-secteur (par exemple, dans la figure 3.5, « Defence ») est qu'elle permet de repérer les entreprises en concurrence les unes avec les autres.

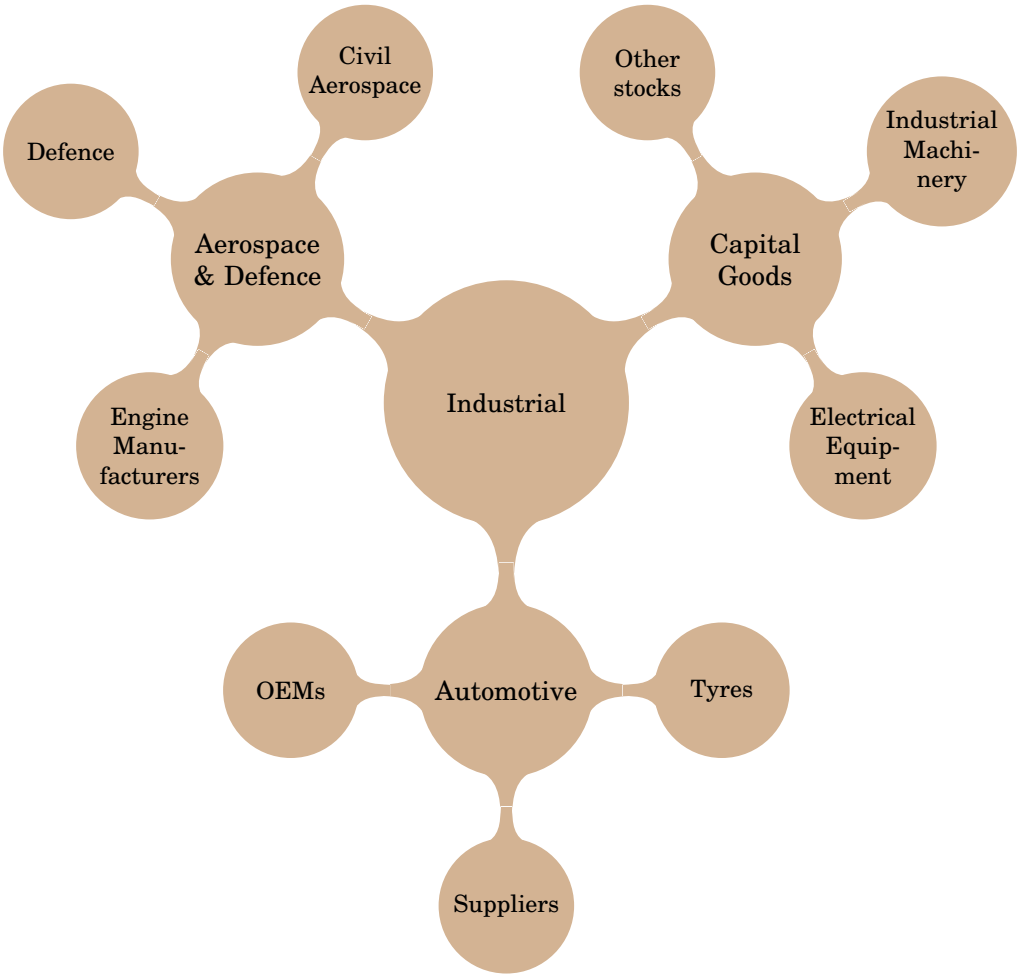


FIGURE 3.5 – Classification Exane d'un secteur (Industrial) en sous-secteurs

3.4 Récapitulatif et problématisation

Nous avons établi certaines caractéristiques spécifiques de notre projet de génération, parmi lesquelles :

— l'absence d'un corpus de textes directement comparables aux textes générés, en raison de la rareté des données financières en open source
— la complexité de la modélisation de ces données selon différentes représentations temporelles et hiérarchiques
— la définition d'une langue spécialisée associée au domaine de l'analyse financière qui contraint les possibilités de génération

Comment identifier et modéliser les informations essentielles à notre tâche pour les exprimer de la façon la plus idéale pour le public visé ? À partir de ces caractéristiques, nous pouvons identifier certains points d'approche pour le développement, l'évolution et l'évaluation de notre système :

— Comment modéliser la complexité de nos données efficacement (tant d'un point de vue informatique que linguistique) ? Il nous faut intégrer la représentation temporelle des données dans nos représentations sémantiques pour pouvoir

générer des textes cohérents, mais aussi la complexité des différentes classifications qui entrent en jeu dans cette représentation
— Comment rédiger des règles de génération qui permettent d'extraire les informations essentielles de cette représentation, via un travail d'ingénierie des connaissances avec des expert·e·s du domaine ?
— Comment utiliser la langue spécialisée associée à notre domaine de spécialité pour exprimer les informations déterminées par ces règles d'une façon précise, utile, et lisible pour nos utilisateurs ? Comment évaluer, ensuite, si ces objectifs ont été atteints ?

C'est à ces questions, entre autres, qu'il nous faudra tenter de répondre par nos expérimentations sur la modélisation du système, la rédaction des règles de génération et la réalisation de surface de ces règles. Par ailleurs, nous devrons aussi prendre en charge l'évaluation du système d'après ces critères, en prenant en compte le fait que nous n'avons pas accès à un corpus directement comparable à celui formé par les textes générés.

3.5 Conclusion

Dans ce chapitre, nous avons présenté le contexte dans lequel notre système de génération a été construit. Pour ce faire, nous avons défini plusieurs caractéristiques saillantes de notre projet qui ont dirigé la tenue de nos expérimentations. Nous avons rappelé l'influence du domaine de spécialité qu'est la finance sur le système, qui génère donc des textes dans un langage spécialisé, à partir de règles et de données spécifiques à ce domaine.

Dans la deuxième partie de ce mémoire, nous décrivons les opportunités d'expérimentations qui se sont présentées à nous lors du développement du système.

Deuxième partie

Expérimentations

CHAPITRE 4

SYSTÈME ET DONNÉES

Sommaire

4.1 Introduction

Dans ce chapitre, nous spécifions les détails des données numériques et textuelles qui étaient à notre disposition pour mener nos expérimentations, ainsi que le moteur de rédaction Yseop que nous avons pu employer pour construire le système de génération.

Nous commençons par décrire un corpus de référence contenant des rapports d'analyse rédigés par des analystes et extraits de Cube. Bien que ce corpus puisse constituer une base pour certains types d'expérimentations, nous expliquons également pourquoi ces textes ne sont pas directement comparables à ceux générés, et nous envisageons d'autres possibilités d'expérimentations avec ce corpus.

Ensuite, nous présentons le moteur de rédaction Yseop qui constitue une partie du cœur de notre pipeline de génération. Pour intégrer les données d'Exane dans Yseop Compose, nous avons dû travailler sur un format de données spécifiques prenant en compte les besoins du système.

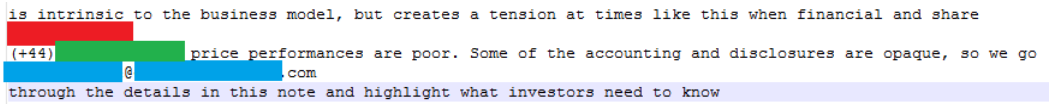
Enfin, nous reprenons ce format de données et nous détaillons la mise en place du pipeline de génération, depuis la modélisation des données jusqu'à la génération des textes finaux.

4.2 Corpus de textes de référence

4.2.1 Prétraitement du corpus

Nous disposons d'un corpus, sous forme de fichiers TXT, qui avait été recueilli avant notre arrivée en extrayant des documents publiés sur Cube (des notes d'analyse courtes ou longues, des annonces de mises à jour de notation, ou encore des compilations de rapports d'analyse) pendant l'année 2018 sous la forme d'un premier fichier, et entre janvier et mai 2019 sous la forme d'un deuxième fichier. Le texte composant ces documents avait été extrait depuis leur version HTML, et par conséquent, il était nécessaire d'effectuer plusieurs opérations de prétraitement avant de pouvoir utiliser ce corpus.

Après une observation manuelle, nous avons noté que le problème le plus important était probablement l'insertion d'éléments de texte présents dans une colonne dans le fichier HTML à l'intérieur de phrases du texte principal (figure 4.1). Outre le fait que ces éléments étaient le plus souvent des données personnelles d'analystes, leur placement à l'intérieur de phrases était susceptible de perturber des opérations de traitement comme l'étiquetage morpho-syntaxique.



Des informations personnelles d'analystes (en rouge, un nom, en vert, un numéro de téléphone, et en bleu, une adresse mail) qui se situaient dans une colonne dans le fichier HTML sont ici insérées au milieu de phrases.

FIGURE 4.1 – Problèmes de prétraitement du corpus (1/4)

Par ailleurs, tous les tableaux de données, qui constituent une part importante des rapports, étaient convertis en lignes de texte au format relativement variable (figure 4.2) et donc difficile à neutraliser. Cependant, la présence de ces lignes est moins problématique dans la mesure où elles sont isolées dans les textes.

```
- High (x) 1.18 0.83 - -    - High (x) 1.45 0.99 - -
- Low (x) 0.67 0.65 - -    - Low (x) 0.91 0.86 - -
Net yield (%) 1.8 2.2 2.3 2.4    Net yield (%) 4.6 6.0 6.4 6.8
ROTE, adjusted (%) 6.2 8.2 9.5 9.8    ROTE, adjusted (%) 9.3 9.3 10.4 11.8
```

Les tableaux de données numériques et autres figures sont convertis en séries de chiffres.

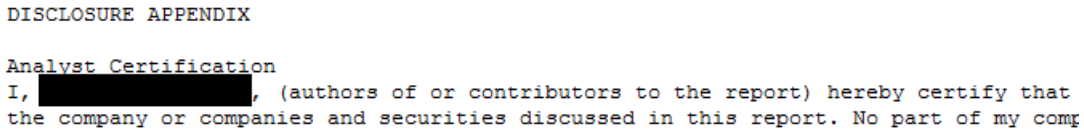
FIGURE 4.2 – Problèmes de prétraitement du corpus (2/4)

```
Contents
Valuation context _____ 3
Bonus pools and deferral accounting _____ 4
```

Les tables des matières et autres éléments de mise en page sont insérés dans les textes.

FIGURE 4.3 – Problèmes de prétraitement du corpus (3/4)

Enfin, il nous fallait supprimer certains éléments récurrents à tous les textes comme les tables des matières (figure 4.3) ou les annexes contenant les divulgations légales mentionnées dans la section 1.2 (figure 4.4). Ces textes sont tous extrêmement similaires et peuvent donc nous induire en erreur lors de notre travail statistique sur le corpus (notamment l'estimation de sa taille). Cependant, ils sont aussi relativement simples à repérer du fait de cette régularité.


DISCLOSURE APPENDIX

Analyst Certification
I, [redacted], (authors of or contributors to the report) hereby certify that
the company or companies and securities discussed in this report. No part of my comp

Les appendices contenant toutes les divulgations et textes légaux (soit l'équivalent de plusieurs pages de texte) se répètent de façon extrêmement similaire dans tous les rapports.

FIGURE 4.4 – Problèmes de prétraitement du corpus (4/4)

Pour effectuer notre travail de prétraitement, nous avons fait le choix d'un script Python qui se chargeait également de transformer les fichiers en un fichier CSV valide. Nous souhaitions nous assurer en particulier que les problèmes de segmentation des phrases présentés dans la figure 4.1 étaient traités correctement, mais nous avons décidé de tolérer la présence d'un minimum de contenus comme ceux de la figure 4.2, qui ne devraient pas perturber outre-mesure notre analyse.

Le script parcourt chaque fichier ligne par ligne, et, après avoir supprimé les lignes indésirables à l'aide d'expressions régulières, reconstitue un article complet, auquel il associe son identifiant et sa date. Les expressions régulières ayant été utilisées dans ce script pour éliminer les contenus superflus sont définies dans la figure 4.5. Le script complet est décrit dans l'annexe B.1.

```
1  re_names = re.compile(r"([A-Z][a-zéè]+(-([A-Za-zéè])+)?␣
       ([A-Za-z']{2,3}␣?)?([A-Z][A-Za-z-éèñíó]+)(,␣CFA)?(,␣)?)+(␣
       \(\d{1,2}␣[A-Z][a-z.]+\))?␣") # 'CFA' is a trade certification
2  re_phone_numbers = re.compile(r"(\(?\+\d{2,3}\)?)?␣((\d{9})|(\d{1,}␣
       )+)")
3  re_dates = re.compile(r"\d{1,2}␣[A-Z]+␣\d{4}")
4  re_email = re.compile(r"[a-zA-Z-]+(\.[a-zA-Z-]+)?@[a-z]+\.[a-z]{2,4}")
5  re_page = re.compile(r"[A-Za-z0-9␣]+␣page␣\d{1,2}␣") # 'Exane BNP
       Paribas Research INVESTMENT BANKS 8 MAY 2019 page 3'
6  re_contents = re.compile(r"(Contents␣)|([^_]+_{2,}␣\d{1,2})") #
       'Investment summary _____ 4'
7  re_misc = re.compile(r"([().,x0-9%]{2,7}␣){2,}") # '20% 450 16%'
8  re_rating = re.compile(r"^[A-Za-z␣-.]+␣\([-=+]\)") # 'CompanyName (+)'
9  re_spec_sales = re.compile(r"Specialist␣sales␣") # Specialist sales
10 re_figure = re.compile(r"((^\(1\)$)|(^Figure))") # beginning of figure
       caption
11 re_valuation = re.compile(r"Prices␣at␣\d+␣[A-Za-z]+(␣\d{4})?") #
       beginning of valuation
```

FIGURE 4.5 – Expressions régulières définies pour le prétraitement du corpus (Python)

4.2.2 Caractéristiques et statistiques

Caractéristiques	Avant prétraitement	Après prétraitement
Format	TXT	CSV
Encodage	CP-1252	UTF-8
Taille (Mo)	167 (2018) + 50 (début 2019)	16 (2018) + 5 (début 2019)

TABLE 4.1 – Comparaison du corpus pré et post-traitement

Une fois ce travail de prétraitement achevé, nous disposions de deux fichiers au format CSV. Le fichier correspondant à janvier-mai 2019 contenait 1719 articles, tandis que celui regroupant les articles de l'année 2018 en contenait 4793 [1]. Ce ratio se retrouve dans la taille des fichiers : en effet, le fichier de début 2019 qui couvre 5 mois est environ trois fois plus petit que celui de 2018 qui en couvre douze (cf. table 4.1). La réduction importante de la taille des fichiers est en partie due au prétraitement, mais également au format des textes originaux (l'extraction des fichiers HTML avait dispersé le texte sur un grand nombre de lignes avec une quantité importante d'espaces). Les fichiers CSV suivent le format présenté dans la table 4.2, qui est repris des fichiers non-prétraités où ces trois éléments étaient déjà séparés par des points-virgule.

Colonne	DIID	DIXML	DIDATE
Contenu	Identifiant à 6 chiffres d'un rapport	Texte du rapport	Date de publication du rapport (jour et heure)

TABLE 4.2 – Format CSV du corpus post-traitement

La table 4.3 résume l'ordre de grandeur de la taille et la distribution du corpus (principalement calculés à l'aide de commandes bash comme `wc` et du module `statistics` de Python) [2]. On dispose donc d'un corpus d'une taille relativement conséquente, composé de rapports d'analyse et d'autres textes relativement similaires en termes de style et de contenu. Cependant, nous allons procéder à une analyse plus fine pour déterminer à quel point ce corpus est similaire au corpus que nous tentons de générer, et dans quelle mesure nous pouvons nous en servir dans le reste de notre travail.

Nombre d'articles	6512
Longueur moyenne d'un article (mots)	~557
Longueur médiane d'un article (mots)	~301
Nombre moyen d'articles par mois	~383

TABLE 4.3 – Caractéristiques du corpus post-traitement (tous fichiers compris)

1. Certains articles qui avaient particulièrement tronqués lors du processus d'extraction ont été retirés du corpus.

2. Ces calculs restent cependant approximatifs, d'abord en raison des problèmes de prétraitement mentionnés en 4.2.1 mais aussi parce que, comme on l'a vu, des articles aux formats très différents peuvent être mis en ligne sur Cube.

4.2.3 Utilité pour les expérimentations

Après avoir observé le corpus, il nous semble que s'il n'est pas directement comparable aux textes produits par notre tâche de génération (notamment dans le cadre d'une évaluation métrique), il a cependant de la valeur en tant qu'outil d'expérimentations.

Parmi les différences qui expliquent pourquoi ce corpus ne peut pas servir de corpus de référence pour une évaluation automatisée de nos textes, ou encore comme outil de base pour une forme d'apprentissage automatique, nous pouvons noter :

— la différence de taille entre les rapports écrits par des analystes et ceux que nous générons (un problème déjà mentionné dans la section 1.3) [3]
— la variabilité des types de textes produits
— la variabilité de la structure des textes produits

Cette variabilité est le reflet d'une différence plus profonde entre ces textes et les textes générés : ils ont des visées discursives plus ou moins différentes. Le but de nos rapports est de fournir une description aussi objective que possible de la situation financière d'une entreprise et des prévisions de performance que les analystes d'Exane lui attribuent. Ce n'est pas nécessairement le cas des rapports rédigés par des analystes : ces derniers ont souvent une visée plus argumentative (par exemple, convaincre du bien-fondé d'une prévision). C'est cette distinction qui informe également la structure des documents [4] et qui distingue les textes que nous souhaitons générer des rapports dont ils s'inspirent.

Néanmoins, cela ne veut pas dire que ce corpus n'a aucun intérêt pour notre travail de génération. Il est notamment un excellent exemple de la langue spécialisée qui est employée dans les rapports générés et que nous avons commencé à décrire dans la section 3.3.1. [Reiter et al., 2005] et [Smiley et al., 2016] sont deux exemples d'utilisation d'un corpus de textes similaires aux textes générés pour mener des expériences sur le choix lexical lors de la génération. Ici, dans la mesure où un de nos objectifs est de générer des textes dont le style est similaire à celui d'un·e analyste, l'usage du corpus dans nos expérimentations (cf. section 5.4) se justifie.

4.3 Moteur de rédaction

4.3.1 Représentation des données

Afin de modéliser les données employées par le moteur de rédaction Yseop, plusieurs types de représentation de données ont été envisagés. Comme nos données doivent souvent être analysées et comparées sur plusieurs axes, le modèle utilisé est proche d'un cube OLAP (« online analytical processing », ou « traitement analytique en ligne »). L'intérêt du cube est qu'il permet de définir des **dimensions** (dans notre cas, les entreprises, les indicateurs de performance, et les années) ayant chacune différents membres, comme illustré dans la figure 4.6 [5].

À l'intersection des dimensions se trouve un **fait** (« fact »), comme par exemple le *Free Cash Flow* (FCF) de l'entreprise Comp1 en 2019, dont on stocke la **valeur**

3. On pourrait cependant considérer cette différence de taille et de complexité comme un problème de récapitulation de textes : cette approche peut nous aider à orienter notre réflexion de façon plus globale, comme on le verra dans le chapitre 5.

4. Par exemple, certains rapports débutent par une section intitulée « Why you should read this report », une forme explicite d'argumentation.

5. Malgré son nom, un cube OLAP ne comporte pas nécessairement trois dimensions.

(ou « measure », une donnée numérique). Pour accéder à certaines valeurs dans le cube, on effectue différents types de **jointures** entre différentes dimensions pour accéder à différents faits (par exemple, des informations sur une entreprise pendant une année).

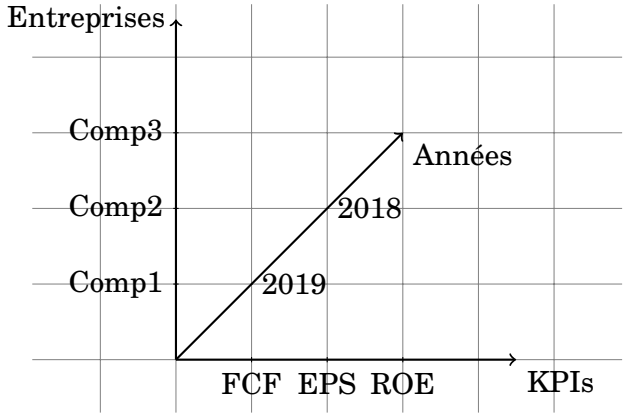


FIGURE 4.6 – Visualisation des trois dimensions de la représentation de données envisagée

L'avantage de ce type de modélisation est qu'il permet de connecter et de comparer un grand nombre de données à l'intérieur du cube. Cependant, l'implémentation de cette modélisation peut s'avérer extrêmement complexe en termes de ressources informatiques. On peut donc essayer d'établir au préalable quelles données devront être comparées ou utilisées, afin de simplifier notre modèle.

4.3.2 Modélisation des règles

if *les données contiennent le taux de variation de l'EPS (EPS change)* **then**
  créer une variable estimatedEPSChange;
  estimatedEPSChange = la valeur de l'EPS change pour l'année FE;
end

Une règle métier permet d'identifier quel fait est important pour l'analyse.

if *estimatedEPSChange > 0* **then**
  générer une phrase sur la croissance de l'EPS ;
else
  générer une phrase sur la baisse de l'EPS;
end

Une règle de génération utilise ces informations pour générer une phrase.

FIGURE 4.7 – Distinction entre les règles métiers et les règles de génération

Le moteur de rédaction d'Yseop, Yseop Compose, va intégrer la représentation de données que l'on vient de décrire et les associer à deux types de règles afin de générer du texte. On peut d'une part appliquer des **règles métier** aux données modélisées afin de générer de nouvelles données ou variables, et d'autre part écrire des **règles de**

génération qui combinent les règles métiers et les données modélisées pour générer du texte (cf. figure 4.7).

Les règles métier permettent en quelque sorte d'extraire les faits essentiels de notre base de données : comme leur nom l'indique, leur écriture nécessite l'intervention d'expert·e·s métier. Quant aux règles de génération, elles sont au centre de notre travail de génération en ce qu'elles assurent le travail de planification et de réalisation de surface.

4.3.3 Réalisation linguistique

Yseop Compose utilise un langage propriétaire, Yseop Markup Language (YML), un langage similaire en certains aspects à Java, pour gérer la planification et la réalisation des textes générés. Les règles de génération rédigées en YML prennent en charge l'essentiel de la réalisation linguistique (accords de genre et de nombre, accords avec des verbes avec leur sujet, mise en forme...). YML possède également plusieurs fonctionnalités liées à la planification de surface que nous avons pu exploiter dans le cadre de ce projet :

— la génération de synonymes, que nous avons pu utiliser pour augmenter la variabilité des textes produits tout en respectant le vocabulaire de la langue spécialisée
— la génération d'expressions référentielles pour les éléments modélisés dans la base de données
— le style d'écriture d'éléments comme les montants ou les quantités à l'intérieur des textes

```
1  Lexicon lexicon = Lexicon.getDefaultLexicon();
2  NLGFactory nlgFactory = new NLGFactory(lexicon);
3  Realiser realiser = new Realiser(lexicon);
4
5  NPPhraseSpec eps = nlgFactory.createNounPhrase("EPS");
6  SPhraseSpec s1 = nlgFactory.createClause(eps, "fall", "sharply");
7  s1.setFeature(Feature.TENSE, Tense.PAST);
8  NPPhraseSpec year = nlgFactory.createNounPhrase("2019");
9  PPPhraseSpec in_year = nlgFactory.createPrepositionPhrase();
10 in_year.addComplement(year);
11 in_year.setPreposition("in");
12 s1.addComplement(in_year);
13 //prints "EPS fell sharply in 2019"
14 System.out.println(realiser.realiseSentence(s1));
```

FIGURE 4.8 – Génération d'une phrase via la bibliothèque SimpleNLG (Java)

La figure 4.8 est un extrait de code Java issu de la bibliothèque open source SimpleNLG présentée rapidement dans la section 1.3.3, permettant de générer une phrase similaire à celle présentée dans la figure 3.1 comme « EPS fell sharply in 2019 ». Elle permet d'illustrer certaines caractéristiques intéressantes d'un langage de génération, notamment :

— l'obligation de s'appuyer sur un lexique (l.1) : le verbe « fall » fait partie du lexique par défaut de SimpleNLG, ce qui explique pourquoi la phrase générée est capable de le conjuguer malgré son irrégularité

— la possibilité de modéliser un groupe nominal comme un objet (l.5) afin de le définir comme le sujet d'une phrase verbale : c'est à partir de cette possibilité que l'on peut envisager de générer des expressions référentielles en YML

Ce dernier point est particulièrement intéressant dans la mesure où il souligne l'importance de la modélisation des données pour la réalisation linguistique (sous forme de lexique, par exemple). Yseop Compose permet de regrouper les problématiques de représentation des données numériques extraites de la base de données et des données linguistiques qui vont servir à exprimer ces données.

4.4 Pipeline de génération

4.4.1 Format des fichiers

Pour générer un rapport sur une entreprise, le pipeline final requiert un fichier XML contenant les informations nécessaires modélisées sur un format proche de celui décrit dans la section 4.3.1. Nous reprenons deux dimensions, les années et les indicateurs de performance, tandis que la dimension des entreprises est séparée par les différents fichiers XML. Environ 500 fichiers sont générés chaque jour afin de regénérer les textes. La table 4.4 décrit le format d'un fichier type plus en détails [6].

Tag XML		Détails
cube	dimensions	définit les membres des dimensions « TIME » (2019, 2018, etc.) et « MEASURETYPE » (FCF, EPS, etc.)
	facts	définit chacun des faits concernant l'entreprise en les rattachant à un membre de chaque dimension
stockName, rating, etc.		données sporadiques représentées sous forme de chaîne de caractères (nom de l'entreprise, sa notation, etc.)
ev_multiples, etc.		contient des informations stratégiques spécifiques à l'entreprise spécifiées par un·e analyste et recueillies sous forme de chaîne de caractères (données libres)
specificMeasures		définit une représentation des membres de MEASURETYPE sous une forme exploitable lors de la rédaction des règles
Currency		définit la représentation (abréviation, symbole...) de la monnaie dans laquelle l'entreprise est valorisée

TABLE 4.4 – Format simplifié d'un fichier XML généré pour une entreprise

Lors de notre travail d'interprétation de données avec des analystes, nous avons dû faire face au problème de l'explicitation de nos règles métier sous un format facile à manipuler pour les analystes mais qui puisse également être converti en code informatique. Nous avons principalement utilisé des fichiers Excel, principalement parce que les tableurs comme ceux produits par Excel sont de loin le format de données le plus utilisé et le plus apprécié dans notre contexte industriel : nous avons donc privilégié cette facilité d'usage.

6. La longueur d'un fichier varie en fonction du nombre d'indicateurs sur lequel on dispose de mesures, ainsi que le nombre d'années pour lesquelles on dispose de ces informations. Cependant, un fichier type contient généralement plus de 1000 lignes.

4.4.2 Données linguistiques

Nous avons présenté dans la section 4.3.3 le moteur de réalisation linguistique employé par Yseop Compose : nous décrivons ici la complexité de certaines structures linguistiques relatives à la GAT qui doivent être modélisées dans le moteur de rédaction. La notion de lexique que nous avions évoquée plus haut s'avère essentielle pour pouvoir modéliser une intersection des variations morpho-syntaxiques et de la synonymie telle que celle décrite dans la table 3.1 : on doit entrelacer d'une part des informations sémantiques de synonymie (par ex. « grow » est un synonyme de « increase ») qui influencent la signification de la phrase générée et d'autre part des variations morpho-syntaxiques (par ex. « grow » est un verbe irrégulier, contrairement à « increase ») pour assurer son sens grammatical.

Nous devons donc générer une modélisation des attributs linguistiques de termes essentiels comme les indicateurs de performance (la table 4.5 présente un exemple pour le terme « EPS », qui existe donc à la fois comme une dimension dans la base de données et comme un objet linguistique) pour que le moteur de rédaction puisse non seulement générer des phrases grammaticalement correctes mais aussi prendre en charge des éléments plus complexes de la planification de surface comme la génération d'expressions référentielles : pour générer une expression anaphorique (comme un pronom à la troisième personne), le système doit être capable de reconnaître un terme quand il est employé plusieurs fois, mais aussi posséder suffisamment d'informations linguistiques pour générer une expression correcte.

Type d'objet linguistique	Attributs
substantif	nombre (singulier) ; genre (neutre) ; lexème (EPS)
groupe nominal	déterminant (indéfini) ; tête du groupe nominal (le substantif « EPS »)
expression référentielle	expression anaphorique pouvant remplacer l'expression référentielle (pronom personnel)

TABLE 4.5 – Modélisation linguistique d'un indicateur de performance, l'EPS

4.4.3 Synthèse du pipeline

La figure 4.9 décrit le pipeline de génération mis en place dans le cadre du projet à partir des caractéristiques du moteur de génération que nous venons de décrire. Nous reprenons également dans la figure les étapes du pipeline de génération décrit dans l'état de l'art (2.3.1) :

— la planification de document (détermination de contenu, structuration de texte) permet de déterminer quelles informations parmi les données analysées seront textualisées et dans quel ordre
— la planification de surface : l'agrégation (la réunion de plusieurs informations en une seule phrase), le choix lexical (la gestion de la synonymie et de la langue spécialisée), et la génération d'expressions référentielles
— la réalisation de surface qui assure la génération d'un texte cohérent et lisible d'un point de vue grammatical

Comme le soulignent [Gatt and Krahmer, 2018, p.71], cette architecture modélise l'avancée du pipeline comme une progression depuis des décisions centrées sur les données, très spécifiques au système, à des décisions linguistiques plus générales.

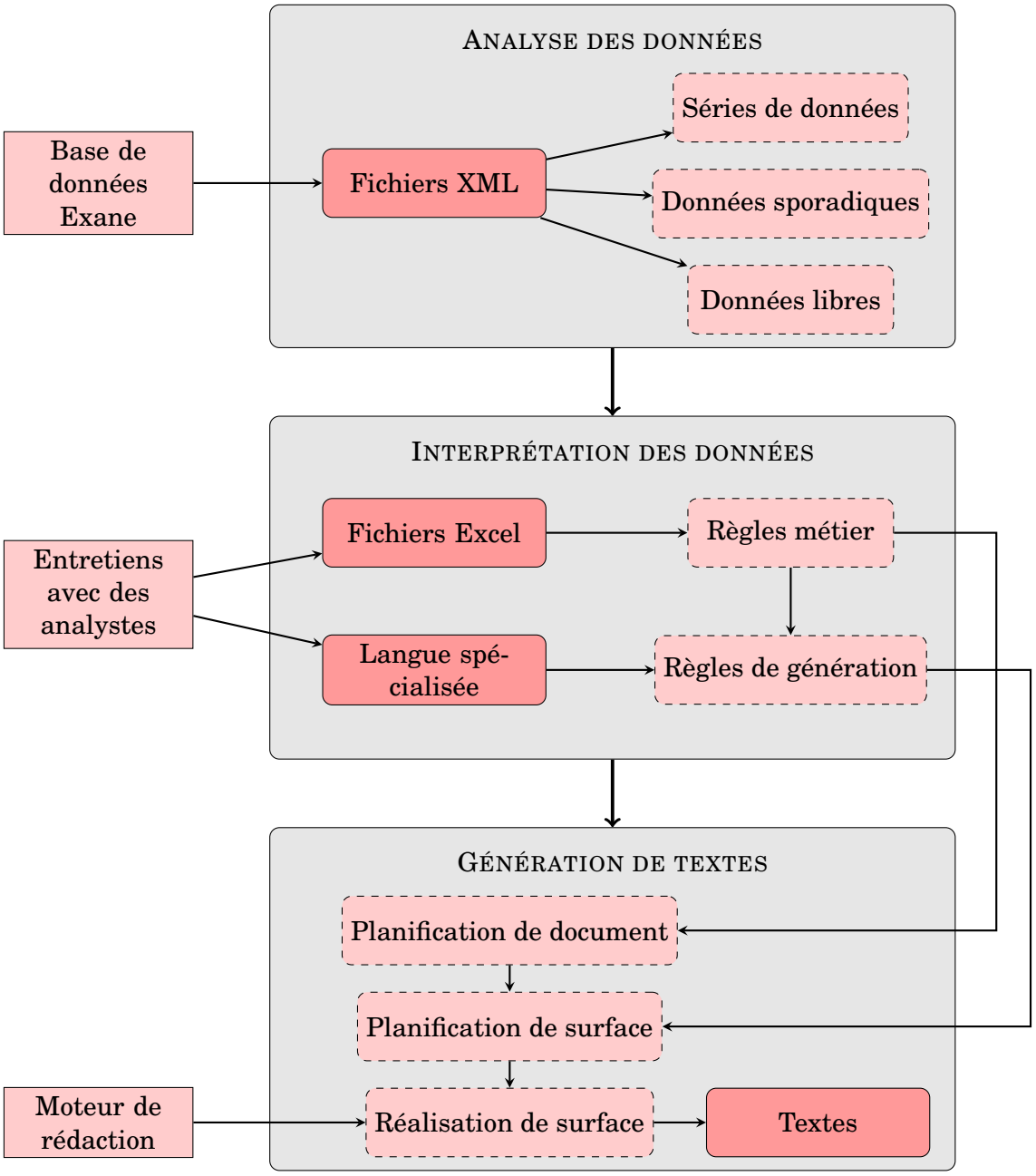


FIGURE 4.9 – Schéma du pipeline de génération d'Exane

4.5 Conclusion

Dans ce chapitre, nous avons présenté les données à notre disposition : d'une part, un corpus de textes de référence que nous avons prétraité, et d'autre part, la base de données dont nous extrayons les informations nécessaires à la génération des rapports. Nous avons également présenté le pipeline de génération qui permet la rédaction de ces textes, notamment son moteur d'écriture Yseop. Nous pouvons donc désormais décrire nos expérimentations plus en détails dans le chapitre 5.

CHAPITRE 5

MODULES DE GÉNÉRATION

Sommaire

5.1 Introduction

Dans ce chapitre, nous décrivons plus en détails certains modules clés du système de génération.

Nous nous intéressons d'abord plus en détails à la rédaction des règles de génération. Les problèmes associés à la rédaction des règles sont généralement liés à la sélection des données : comment déterminer quelles sont les informations essentielles à fournir dans un rapport, et par conséquent, quelles données doivent être extraites pour parvenir à générer automatiquement ces informations ?

Nous présentons ensuite la variabilité de la structure des documents générés. Après avoir décidé quelles types d'informations doivent être exprimées, le système doit être capable de les organiser d'une façon cohérente et rhétoriquement satisfaisante. On verra que la structure d'un rapport est également influencée par certaines caractéristiques de l'entreprise qui est traitée dans le rapport.

Enfin, nous présentons une expérience que nous avons menée afin d'approfondir notre réflexion sur l'influence de la langue spécialisée qu'est l'anglais financier sur les textes générés. Nous nous intéressons particulièrement dans cette expérience aux connotations de différents synonymes des termes « monter » et « baisser », mais nous présentons également d'autres problématiques de lexicalisation spécifiques à notre domaine, comme la lexicalisation d'expressions numériques.

5.2 Règles de génération

5.2.1 Ingénierie des connaissances

Nous avons vu dans la section 4.3.2 comment le moteur de génération interprétait les règles de génération. Ici, nous décrivons plus en détails le processus de sélection des données et de décisions sur la rédaction des règles. Dans notre expérience, ces deux phénomènes sont particulièrement liés : il faut savoir quelles sont les données à notre disposition pour générer des règles, mais il faut aussi extraire ces données en fonction des règles que l'on souhaite rédiger.

Dans la pratique, nous avons employé plusieurs méthodes lors de notre travail avec des analystes pour réussir à formaliser les règles essentielles de l'analyse. D'abord, les analystes tentaient d'établir des concepts généraux autour d'un terme spécifique, comme par exemple l'EPS (ces concepts généraux se retrouvent également souvent dans la structure des textes). En général, nous avons noté que les analystes comparaient les résultats d'un KPI sur différentes plages temporelles et essayaient de repérer celle qui permettait l'analyse la plus pertinente (cf. figure 5.1).

```
if on observe une tendance d'évolution du KPI sur le LT then
    if cette tendance est inverse aux prédictions then
        évoquer ce contraste ;
    else
        évoquer cette continuité ;
    end
end
if la prédiction FE est très différente des résultats LP then
    évoquer ce changement brutal ;
end
```

Une forte majorité des phrases générées dans les rapports est le résultat d'une analyse temporelle de la performance d'un KPI.

FIGURE 5.1 – Analyse simplifiée de tendances sur différentes plages temporelles pour un KPI

Une fois ces grandes lignes décrites, nous avons parfois demandé aux analystes de rédiger des rapports ou des extraits de rapport, puis de nous expliquer le raisonnement les ayant conduit à ce rapport. Ce processus, bien qu'efficace, s'est avéré très gourmand en temps, ce qui explique pourquoi nous y avons eu recours de façon sporadique, ou dans le cadre de questions très précises [1].

5.2.2 Cohérence temporelle

La question de la représentation temporelle des données est commune à un grand nombre de systèmes de GAT, et elle a par conséquent fait l'objet d'un grand nombre d'analyses. [Portet et al., 2009] définissent la notion de cohérence temporelle (« temporal coherence ») pour désigner l'importance d'une représentation temporelle précise des évènements décrits. Si nous disposons dans notre domaine de spécialité d'un

1. Par exemple, pour déterminer le seuil à partir duquel une augmentation n'est plus « légère » mais « modérée », nous avons demandé à une analyste de verbaliser plusieurs phrases afin d'essayer de déterminer où elle plaçait instinctivement ce seuil.

certain nombre de termes temporels (table 3.4) non-ambigus qui peuvent nous assister dans cette tâche, nous rencontrons néanmoins le même problème que Portet et al., c'est-à-dire le conflit entre la cohérence temporelle et la cohérence rhétorique du texte, qui veut qu'on mette en avant les éléments essentiels d'un rapport.

[Sambaraju et al., 2011] propose deux types d'évaluation des textes produits par [Hunter et al., 2012] : l'analyse de contenu et l'analyse de discours. Ces deux analyses se concentrent sur la force rhétorique des textes générés, et démontrent d'une part que la cohérence temporelle est un élément essentiel des textes rédigés par des humains, mais d'autre part que cette cohérence temporelle est un élément essentiel de la « narration » (*narrative*) du texte. Dans la pratique, les textes rédigés par notre système favorisent, pour reprendre la classification de [Sambaraju et al., 2011], des marqueurs ou intervalles temporels spécifiques (ex. « 2019 », « between 2011 and 2016 ») plutôt que des références indirectes (ex. « prior to this ») ou subjectives (ex. « recently »).

[Smiley et al., 2017] présentent une autre difficulté liée à la représentation temporelle des données : les données incomplètes (cf. figure 5.2). Ils soulignent qu'un système de GAT utilisant des plages de données, comme le nôtre, peut être incapable de repérer un manque ponctuel de données et effectuer par conséquent des calculs incorrects. Cette difficulté s'étant présentée dans notre travail, nous avons dû modifier nos règles métier pour la gérer.

Apr. 16	Apr. 17	Apr. 18	Apr. 19	Apr. 20e[1]
NC	NC	NC	NC	NC
46.3%	47.2%	46.8%	46.8%	48.8%
28.6%	28.2%	28.0%	28.1%	28.1%
33.9%	34.5%	NC	24.8%	25.0%
16.8%	16.3%	17.1%	18.6%	18.0%

Les données manquantes (NC) sont particulièrement problématiques quand il n'en manque qu'une seule dans une ligne.

FIGURE 5.2 – Exemple d'une grille de données incomplète

5.2.3 Complexité des données

Bien que la plupart des données exploitées par le système soient des données strictement numériques, nous avons dû gérer des cas spécifiques où le format et la complexité des données ont rendu leur représentation plus difficile.

Nous avons déjà mentionné (3.2.2) la notion d'informations stratégiques dans le domaine de la finance. Un exemple récurrent de d'information stratégique considérée par les analystes comme pertinente pour la rédaction du rapport était la présence de provisions structurelles d'une entreprise [2] comme par exemple des provisions pour pensions de retraite. Cette information varie par entreprise : nous avons donc dû trouver un moyen d'insérer un commentaire libre sur cette information stratégique dans notre texte généré. Ces commentaires sont généralement limités à une phrase non-verbale ou un groupe nominal (ex.« predominantly pensions ») et elles sont inté-

2. Une provision est une charge probable qu'on doit rattacher à l'exercice comptable d'une entreprise.

grées via des parenthèses pour ne pas compromettre la structure grammaticale des autres phrases.

Par ailleurs, nous avons été confrontée à un problème majeur : la comparaison de différentes entreprises entre elles. Nous avons déjà vu (3.3.3) qu'il était possible de comparer une entreprise avec d'autres selon un nombre conséquent de critères, en utilisant la notion de « comparables ». Le problème qui se pose est la complexité informatique de modélisation des données nécessaires pour ces comparaisons. Dans l'état actuel, les données liées à chaque entreprise sont modélisées de façon indépendante et stockés dans des fichiers XML séparés. Pour pouvoir effectuer des comparaisons entre les entreprises, il faudrait soit modéliser l'intégralité des entreprises comparables dans chaque fichier de données (ce qui complexifie grandement leur traitement), soit effectuer des pré-calculs de comparaison de ratios pour disposer d'éléments simplifiés de comparaison. Nous avons choisi cette dernière solution, malgré les limites qu'elle pourrait potentiellement nous imposer.

5.3 Structure des rapports

5.3.1 Texte, paragraphe et phrases

La structure des rapports générés est intimement liée à leur sens. Nous rappelons ici les différents niveaux d'organisation des rapports.

Section	Paragraphe
Growth and returns	FCF (*Free Cash Flow*)
	EPS (*Earnings Per Share*)
	ROCE (*Return On Capital Employed*)
	ROE (*Return On Equity*)
Financial structure	Cash flow/EBITDA (*Earnings Before Interest, Taxes, Depreciation, and Amortization*)
	Dividend
Valuation	EV (*Enterprise Valuation*)

Dans la pratique, la distinction entre les deux paragraphes de « Financial structure » n'est pas marquée visuellement. D'autre part, la section « Valuation » est encore incomplète, d'où la présence d'un seul paragraphe.

TABLE 5.1 – Paragraphes pouvant être générés par le système dans un rapport

La table 5.1 présente tous les paragraphes pouvant être générés dans un rapport. Chaque paragraphe est constitué de phrases ou de groupes de phrases générés par une règle de génération. Si aucune phrase n'a été générée dans un paragraphe, le titre du paragraphe n'est pas généré. De la même façon, une section n'apparaît pas si aucun de ses paragraphes ne contient de texte (ce qui reste relativement rare). Ce format de document, où les sections correspondent à de grandes notions d'analyse financière et les paragraphes sont nettement liés à un type d'indicateur, divergent assez nettement du format d'un rapport humain qui est nécessairement beaucoup plus libre dans son organisation, mais sa simplicité est avantageuse à la fois pour le moteur de génération et pour le lecteur.

Pour assurer la cohérence rhétorique du texte généré, il faut souvent s'assurer que les phrases générées par différentes règles de génération sont cohérentes tant du

point de vue de leur exactitude que de leur lisibilité, afin que les textes conservent leur cohérence d'analyse. Il faut donc être capable d'établir des liens rhétoriques (comme un lien de causalité) entre différents paragraphes et parfois différentes sections du texte. Nous revenons sur cette question dans la section 5.3.3.

5.3.2 Statistiques sur la structure des rapports

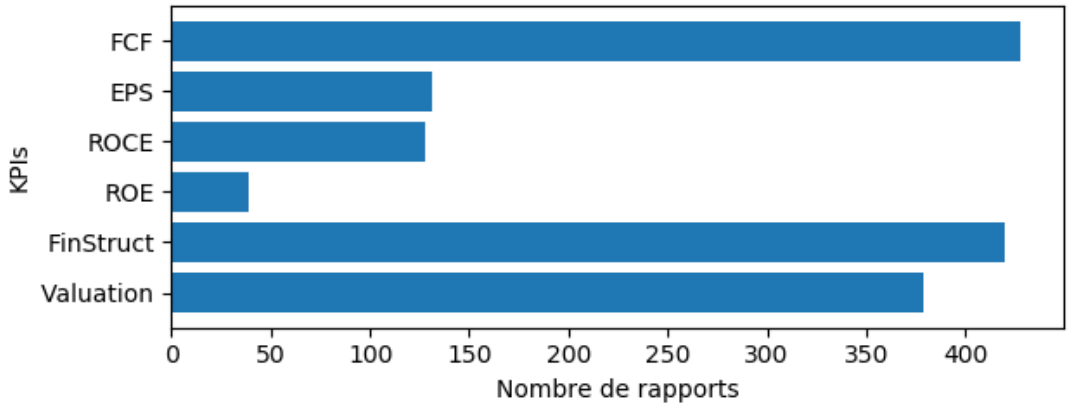


FIGURE 5.3 – Fréquence des paragraphes concernant différents KPIs dans les rapports générés

Vers la fin de notre stage, nous avons décidé d'analyser la répartition des différents paragraphes générés par le système. La figure 5.3 présente la répartition des paragraphes et sections [3] dans 423 textes générés le jour de l'analyse.

Nous pouvons noter que certains paragraphes sont quasiment omniprésents : dans le cas de « FCF », cela peut s'expliquer par le fait que ce paragraphe est le plus ancien et contient le plus grand nombre de règles. Dans le cas de « Financial structure », cette proportion peut s'expliquer par le fait qu'une des règles de cette section génère du texte quelque soit la valeur du ratio utilisé dans cette règle [4]. Ce type de statistiques est utile notamment pour les analystes, qui sont plus à même de déterminer s'il est normal, par exemple, que le paragraphe sur le ROE soit généré si peu souvent et éventuellement de proposer une modification des règles. En effectuant ce type d'analyse régulièrement, il serait possible de suivre l'évolution des textes en fonction de l'évolution des règles.

Nous voulions aussi avoir un ordre d'idée de la longueur des paragraphes générés, qui est un indicateur plus fin de la répartition du temps consacré à chaque indicateur dans les textes. La figure 5.4 modélise la longueur médiane de chaque paragraphe [5], leur longueur minimale et maximale, ainsi que leur répartition entre ces valeurs. Nous pouvons noter des regroupements autour de certaines longueur, pour exemple pour le FCF, l'EPS et le ROCE, ce qui peut nous indiquer que certaines règles se

3. Nous détaillons les différents paragraphes de la section « Growth and returns », mais restons au niveau de la section pour les deux autres, dans la mesure où la distinction n'est pas faite dans le texte.

4. La raison pour laquelle le texte n'est pas généré pour 100% des rapports est l'absence, pour certaines entreprises, de ce ratio.

5. La longueur médiane des paragraphes sur l'EPS, le ROCE et le ROE est égale à zéro dans la mesure où ces paragraphes n'apparaissent pas dans plus de la moitié des rapports.

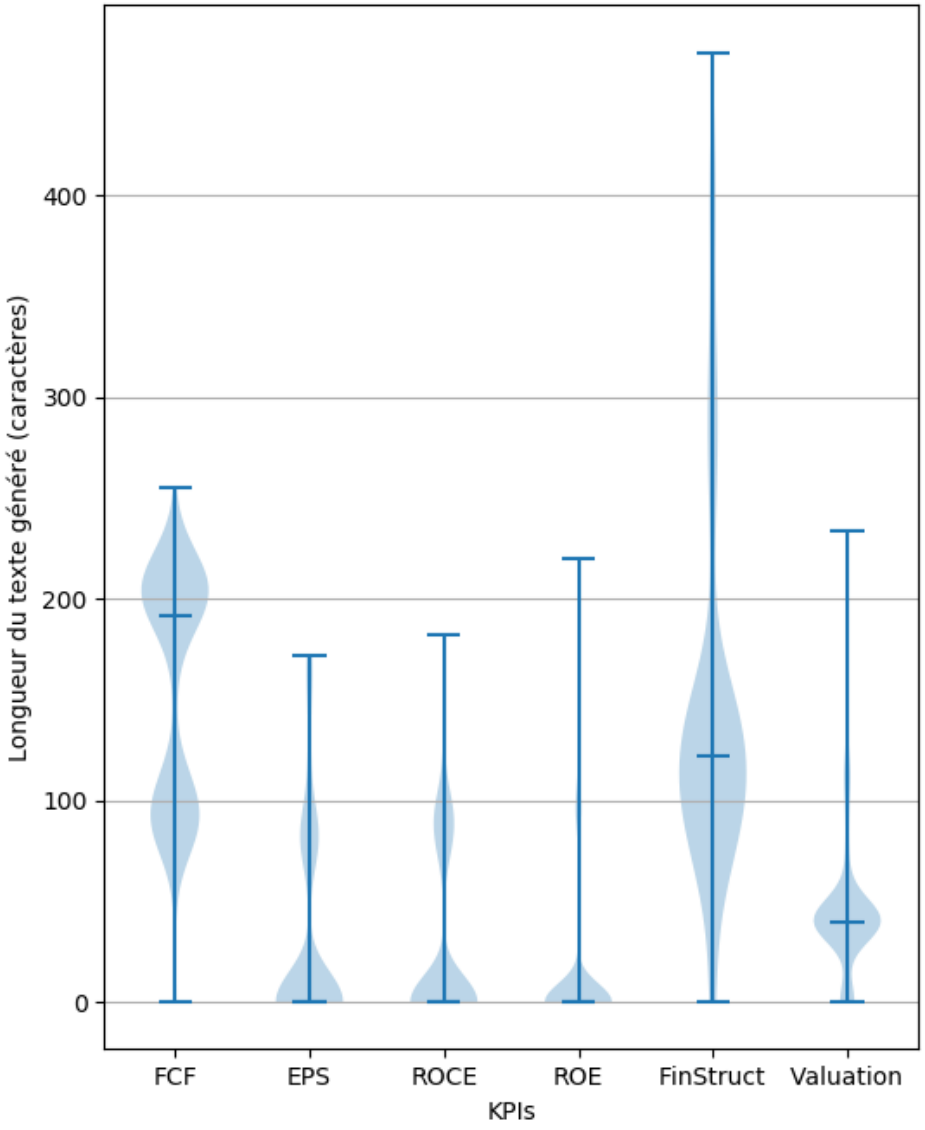


FIGURE 5.4 – Variabilité de la longueur des paragraphes générés

déclenchent toujours dans les mêmes rapports. Cette observation peut orienter la rédaction de nouvelles règles.

5.3.3 Cohérence discursive

Toutes les parties d'un texte cohérent doivent être liées sinon explicitement, au moins implicitement. Cette obligation a une forte influence sur la façon dont le moteur de GAT envisage la structure des textes générés.

Au niveau de l'enchaînement entre différentes règles, le système doit être capable d'exprimer, par exemple, un lien de causalité simple entre deux notions :

> EBITDA conversion to FCF has been weak since 2014, averaging 24% over the period, on account of CAPEX levels. **However**, Free Cash Flow should go up by 89% from 2018 to 2021, mainly driven by EBITDA growth.

Cette obligation n'est pas entièrement indépendante de l'obligation de cohérence temporelle décrite en 5.2.2, et elle suppose de modéliser un nombre important d'indicateurs liés à l'analyse de l'entreprise dans le système de GAT qui lui permettent de « reconnaître » ces liens rhétoriques et de contrôler la structure des textes générés pour pouvoir rendre cette analyse plus claire.

Pour garantir la clarté du texte, le système peut également choisir de ne pas générer une suite de règles : dans le cas décrit dans la figure 5.1, le système ne décrit l'évolution court terme de l'indicateur que si aucune tendance long terme n'a été repérée.

Enfin, on peut essayer de structurer les textes en fonction d'indicateurs externes, comme par exemple la classification de l'entreprise (cf. section 3.3.3). Un algorithme en développement pendant notre stage permettrait de définir quels paragraphes générer en fonction du secteur et du sous-secteur de l'entreprise. Si nous devions reproduire l'analyse statistique de la structure de notre corpus, nous intégrerions le secteur et le sous-secteur de l'entreprise à nos diagrammes, afin de faciliter ce travail de structuration.

5.4 Choix lexical

5.4.1 Influence de la langue spécialisée

Phrase	Notes
The company has lost share and guidance disappointed but it is our top pick.	Le complément d'objet de « disappoint » est omis. La référence de l'anaphore « it » est ambigüe.
GMV growth in 2019 will be around 60% we think.	Le style de la phrase (« ...we think ») est fortement oral.

TABLE 5.2 – Exemples de langue spécialisée grammaticalement inhabituelle

Un système qui doit générer des textes dans une langue spécialisée doit-il reproduire les singularités grammaticales courantes de cette langue spécialisée ? Comme le soulignent les exemples de la table 5.2, l'anglais de la finance de marché est une langue dont la grammaire peut être souple, et qui tolère des constructions grammaticales inhabituelles [6]. Nous avons choisi d'essayer de générer des phrases grammaticalement standard. Cependant, nous avons pu effectuer un travail de recherche plus complet sur la synonymie des termes essentiels des rapports d'analyse.

Dans le cadre de ce travail, nous avons pu reprendre les formes suggérées dans la table 3.2 et approfondir notre travail de recherche de variations morpho-syntaxiques. Nous avons à nouveau pu noter l'existence de ce que [Kittredge et al., 1994] désignent comme une « sémantique floue » (*fuzzy semantics*) de certains termes, où les nuances de sens repérées par différents analystes semblent fortement subjectives.

Ce flou peut s'avérer problématique car, comme le notent [Portet et al., 2009, p.802], la lexicalisation, surtout de termes essentiels comme « expect » ou « increase », est directement liée aux représentations sémantiques qu'utilise le système de GAT. Nous voulions donc essayer de déterminer s'il était possible d'utiliser notre corpus

6. On peut supposer, entre autres, que cet aspect est dû à l'internationalisation de l'usage de l'anglais dans la finance, qui a amené beaucoup de non-locuteurs de l'anglais à devoir l'utiliser.

Verbe (infinitif)	expect	anticipate	forecast
proposition relative	ø	we anticipate that EPS will decrease	we forecast that EPS will decrease
complément en TO	we expect EPS to decrease	ø	ø
complément passif en TO	*the EPS is expected to decrease*	ø	ø
groupe nominal (complément)	we expect a decrease of the EPS	we forecast a decrease of the EPS	we anticipate a decrease of the EPS
groupe nominal (tête)	*we expect a decreasing EPS*	*we forecast a decreasing EPS*	*we anticipate a decreasing EPS*
construction modale	EPS should decrease		

Les formes grammaticalement incorrectes ont été retirées du tableau par rapport à la table 3.2. Les formes rejetées lors d'entretiens avec des analystes sont notées en italique et précédées de *.

TABLE 5.3 – Variations syntaxiques enrichies des synonymes du terme « expect »

de rapports humains pour repérer des synonymes d'un terme courant, sans être sous l'influence de l'idiolecte de chacun des analystes participant au système.

5.4.2 Qualifier l'évolution d'un indicateur

Nous avons choisi de reproduire, en modifiant certains détails pour des raisons pratiques, une partie de l'expérience décrite par [Smiley et al., 2016] pour déterminer statistiquement des synonymes des termes « increase » et « decrease » dans le domaine financier. Les auteurs observent que l'emploi de certains termes est corrélé à l'intensité de la variation décrite.

VERBE	ADVERBE ou PRÉPOSITION (facultatif)	NOMBRE CARDINAL	POURCENT
VB,VBD,VBG, VBN,VBP,VBZ	up, down, by	CD	%
grew	by	17	%
go	down	0.3	%
increasing		50	%

Stanford CoreNLP et TokensRegex utilisent le Penn Treebank Tagset [Jurafsky and Martin, 2009] pour leurs annotations morpho-syntaxiques, et nous reprenons ici ces étiquettes.

TABLE 5.4 – Motif verbal à extraire dans le corpus et exemples de ce motif dans le corpus

En utilisant le corpus décrit dans la section 4.2 et Stanford CoreNLP [Manning et al., 2014], nous avons extrait tous les motifs correspondant

au format décrit dans la table 5.4. Comme le précisent [Smiley et al., 2016], réduire les verbes extraits à ceux employés pour exprimer une variation de pourcentage permet de s'assurer que toutes les variations sont relativement comparables [7]

Nous souhaitions, en reproduisant cette expérience, valider nos choix manuels de synonymes pour ces deux termes, mais aussi observer la répartition des verbes extraits pour pouvoir éventuellement générer des synonymes plus subtils en cas d'augmentation/baisse légère, modérée ou forte.

```
1 ENV.defaults["ruleType"] = "tokens"
2 pos = { type: "CLASS", value:
      "edu.stanford.nlp.ling.CoreAnnotations$PartOfSpeechAnnotation" }
3 ner = { type: "CLASS", value:
      "edu.stanford.nlp.ling.CoreAnnotations$NamedEntityTagAnnotation" }
4
5 { ruleType: "tokens",
6   pattern: ( ( [ { pos:/VB.*/}] /up|down/?) /by?/ ( [ { pos:CD } ])
        /%/),
7   action: ( Annotate($1, ner, "VERBE"), Annotate($2, ner, "POURCENT"
        )) }
```

FIGURE 5.5 – Fichier de règles TokensRegex (Stanford CoreNLP)

Nous avons utilisé TokensRegex [Chang and Manning, 2014], un module d'extraction de motifs de Stanford CoreNLP, pour modéliser le motif de la table 5.4, annoter les motifs dans le corpus et les extraire dans un fichier CSV. Après avoir prétraité le fichier et retiré certains verbes non-pertinents, nous avons suivi la présentation de [Smiley et al., 2016] et modélisé l'écart interquartile de la valeur des pourcentages associés à chaque verbe (figures 5.6 et 5.7).

Les synonymes que nous avions choisis manuellement pour exprimer la baisse et la hausse étaient :

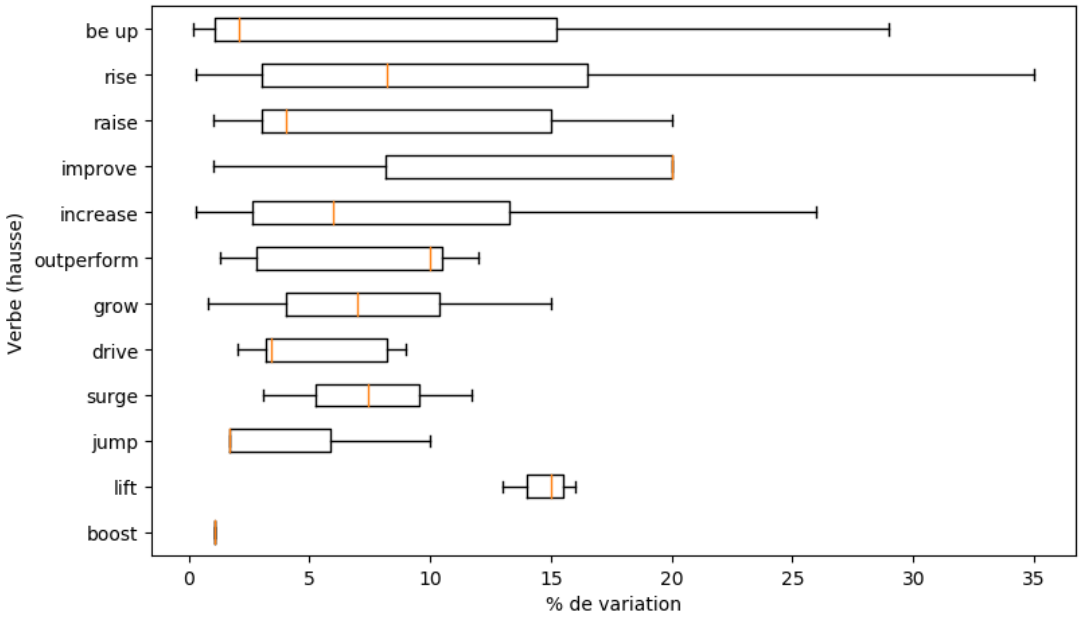
— pour la hausse : *increase*, *go up*, *grow*
— pour la baisse : *decline*, *decrease*, *go down*

Le seul terme qu'on ne retrouve pas dans les graphes est *go up*, terme qui par ailleurs fut contesté plusieurs fois par des analystes (cf. commentaires de l'évaluation dans la table A.3) [8]. Cependant, nous n'obtenons pas des résultats aussi satisfaisants que [Smiley et al., 2016], où les verbes avec l'écart le plus important sont des verbes exprimant clairement une forte intensité comme *rocket* ou *plunge*. Ici, les verbes employés le plus largement ont un sens assez neutre, comme *rise* ou *increase*.

Nous pouvons proposer plusieurs explications à cette déception : d'une part, notre corpus n'était pas un corpus de presse mais un corpus plus spécialisé, dont le lexique est probablement plus neutre (les verbes *rocket* et *plunge*, par exemple, n'apparaissaient pas dans notre liste). D'autre part, nous supposons que le sujet du motif verbal sélectionné influence l'importance de l'augmentation. En fonction de l'indicateur mentionné, une augmentation de 5% peut être légère, modérée ou forte. Une possibilité aurait été d'extraire le sujet du motif verbal, ce qui aurait probablement nécessité une analyse de dépendances syntaxiques du corpus et complexifié l'expérience.

7. Par exemple, une augmentation de 10 euro et une augmentation de 10 centimes ne sont pas, dans la plupart des cas, comparables.

8. Nous avons choisi en nous appuyant sur les résultats de cette expérience de remplacer *go up* par *rise*.



La visualisation ci-dessus n'inclut pas les valeurs atypiques, au contraire de celle ci-dessous.

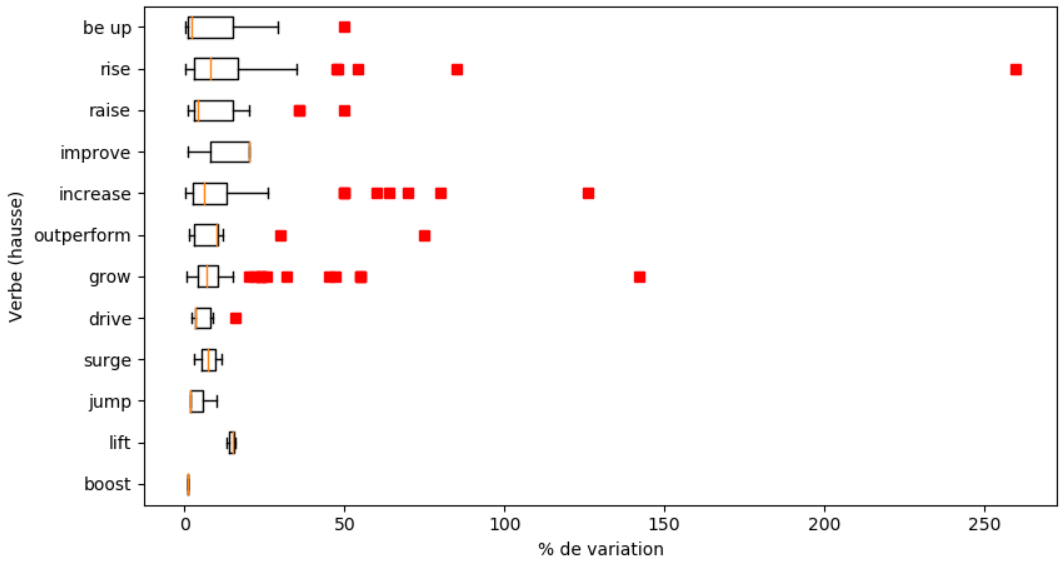
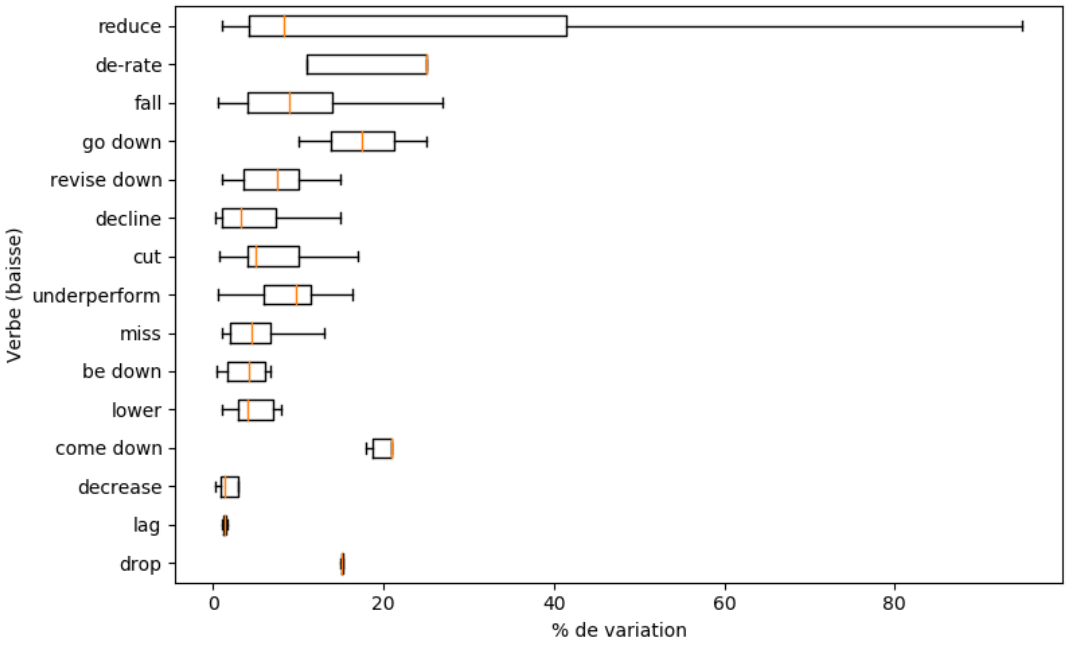


FIGURE 5.6 – Répartition interquartile des verbes de hausse d'après le pourcentage exprimé

Par ailleurs, les termes choisis présentent plus qu'une simple variation d'intensité. Un lexique financier [9] fournit un grand nombre de synonymes du terme *increase* dont la signification varie en fonction d'autres critères que l'intensité (par exemple, le terme *rebound*).

Enfin, l'influence de l'idiolecte et des choix personnels des analystes n'est encore

9. Par exemple, ce dictionnaire de synonymes :https://dictionary.cambridge.org/topics/finance/price-increase/.



La visualisation ci-dessus n'inclut pas les valeurs atypiques, au contraire de celle ci-dessous.

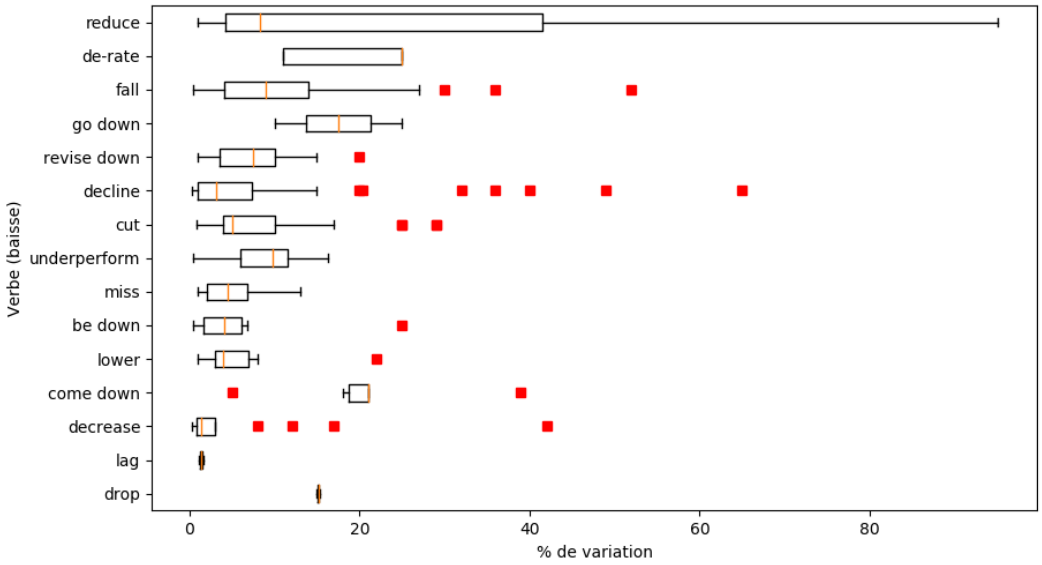


FIGURE 5.7 – Répartition interquartile des verbes de baisse d'après le pourcentage exprimé

une fois pas à négliger [Reiter et al., 2005]. Le corpus que nous avons utilisé est constitué de rapports rédigés par un nombre relativement limité d'analystes dont le style personnel est assez fortement établi. Il n'est pas nécessairement évident que les verbes extraits lors de cette expérience seraient utilisés de la même façon dans le contexte de la rédaction de rapports simplifiés comme celle effectuée par notre système de GAT.

[Smiley et al., 2016] présentente un protocole d'évaluation de leur système où des

humains évaluent des phrases générées en utilisant les verbes extraits dans différents contextes. Dans la mesure où nous avions déjà prévu une évaluation globale du système, nous n'avons pas proposé d'évaluation spécifique de cette expérience de choix lexical, à part une validation des termes choisis lors des réunions de travail sur le système.

5.4.3 Lexicalisation d'expressions numériques

Au cours de l'expérience décrite dans la section précédente, nous avons pu nous apercevoir pendant le prétraitement des valeurs numériques que l'expression de ces valeurs de façon textuelle pouvait être fortement significative. L'expression des valeurs numériques pouvait être nuancée de différentes façons :

— le nombre de chiffres après la virgule sous-entend un niveau de précision différent (« a relatively low 22% adjusted tax rate », « c2.3% of the current market cap », « a 13.77% 2017 »)
— l'usage d'une plage (« a 50-60% cut ») ou d'un marqueur d'imprécision (« c2.3% of the current market cap »)

[Power and Williams, 2011] et [Williams and Power, 2013] se sont intéressés aux problèmes liés à la génération d'expressions numériques, notamment les questions d'arrondi [Williams and Power, 2013] et d'approximation [Power and Williams, 2011]. Parmi les exemples présentés, nous pouvons noter que l'usage de l'expression « c.2.3% » peut sembler légèrement paradoxale, puisque la décimale induit un degré de précision qui est nié par l'utilisation du *circa*.

Dans le cadre de notre système, nous avons souvent décidé, après consultation avec des analystes, de spécifier le nombre de décimales après la virgule dans les règles en fonction de l'indicateur dont le système décrit la variation afin d'obtenir un niveau de précision adapté à l'objet décrit. En effet, l'ordre de grandeur de certains indicateurs permet de préciser le degré de précision avec lequel il faut exprimer leur valeur.

La question de l'approximation est plus complexe pour notre système de GAT, qui doit normalement essayer d'être le plus précis possible. Cependant, paradoxalement, générer un nombre trop précis alors que l'on décrit une estimation ou une approximation peut induire le lecteur en erreur par sous-entendu [Power and Williams, 2011].

5.5 Conclusion

Dans ce chapitre, nous avons présenté certains aspects des modules de génération de notre système qui soulevaient des questions particulièrement saillantes.

Nous avons établi la notion de cohérence temporelle et son importance dans l'écriture des règles de génération du système. Nous avons également établi que la comparaison de différentes entreprises dans le même rapport, bien qu'elle soit un étape presque naturelle de l'analyse pour les analystes, peut être particulièrement complexe pour un système de GAT. Nous avons décrit la structure des textes générés et le raisonnement soutenant cette structure, ainsi que des possibilités d'adaptation de cette structure en fonction de classifications externes. Enfin, nous avons effectué une expérience traitant de la synonymie de termes essentiels de l'analyse financière, et nous avons présenté certaines pierres d'achoppement de la lexicalisation de ce type de textes.

Dans le chapitre suivant, nous procédons à l'évaluation du système.

CHAPITRE 6

ÉVALUATION DU SYSTÈME

Sommaire

6.1 Introduction

Nous présentons dans ce chapitre les critères, le protocole et les résultats de l'évaluation du système que nous avons essayé de mettre en place à la fin de notre stage. Nous avons décidé de développer un protocole d'évaluation humaine par des expert·e·s à Exane afin d'obtenir une vue d'ensemble de la performance du système, d'une part parce qu'une évaluation est généralement considérée comme standard en GAT (cf. section 2.4.3) et d'autre part parce que nous ne disposions pas d'un corpus permettant d'utiliser des mesures d'évaluation (cf. section 4.2.3).

Nous expliquons d'abord plus en détails la démarche de mise en place de l'évaluation, du choix des données à évaluer à la rédaction du formulaire d'évaluation. Nous devions notamment prendre en compte durant ce processus le temps limité que peuvent accorder des expert·e·s à l'évaluation.

Nous décrivons ensuite les critères d'évaluation que nous avons choisis en fonction de ce qu'ils peuvent nous apprendre des performances du système : la lisibilité des textes produits, leur exactitude, et leur utilité dans le cadre du projet de génération.

Enfin, nous présentons les résultats obtenus, proposons une interprétation de ces résultats, et décrivons les autres types d'évaluation possibles pour un système comme le nôtre.

6.2 Protocole d'évaluation

6.2.1 Type d'évaluation

Dans cette section, nous abordons les distinctions entre différents types d'évaluation que nous avions décrites lors de l'état de l'art (2.4) afin de développer un protocole d'évaluation adapté à nos besoins et contraintes. Nous décrivons plus en détails la logique derrière le choix de nos critères d'évaluation dans la section 6.2.2.

Nous avons déjà établi que nous ne pouvions pas avoir recours à une évaluation métrique : nous devions donc choisir des évaluateurs humains. Le système au développement duquel nous avons travaillé est un système à visée commerciale qui a donc des utilisateurs potentiels clairement définis (des clients d'Exane). Cependant, dans un contexte industriel, il n'était pas possible de faire évaluer le système par ces utilisateurs. Nous avons néanmoins eu l'opportunité de faire évaluer les textes générés par des analystes et vendeu·r·se·s travaillant à Exane, qui pouvaient donc, en tant que spécialistes du domaine, évaluer les textes d'après des critères de qualité intrinsèques comme leur lisibilité ou leur exactitude. Nous pouvions donc envisager une évaluation intrinsèque semblable à celle décrite dans la première ligne de la table 2.1 [Hastie and Belz, 2014].

Le profil des personnes ayant pu répondre aux formulaires d'évaluation (cf.6.2.3) est relativement varié : certaines ont pour langue natale l'anglais, tandis que d'autres sont principalement francophones. Bien que toutes comprennent l'anglais, qui est la langue de travail pratiquement incontestée de la finance de marché, il nous semblait important de relever cette distinction et d'observer son impact potentiel sur les notes attribuées au texte, d'autant plus que notre période de stage nous a permis de noter que nos interlocuteurs anglophones offraient plus spontanément des remarques d'ordre linguistique ou grammatical que leurs homologues français.

Nous avons soumis nos textes à des personnes exerçant deux métiers distincts : des vendeu·r·se·s et des analystes. Ce choix fut fait en partie pour obtenir une nombre relativement conséquent d'évaluations ; cependant, ces deux métiers peuvent apporter une perspective intéressante à l'évaluation :

— un·e analyste est plus à même d'évaluer la justesse du raisonnement d'analyse derrière le texte
— un·e vendeu·r·se, qui, dans le cadre de son activité, est directement en contact avec les clients d'Exane, est plus à même d'évaluer l'utilité qu'aurait le texte pour un·e client·e

6.2.2 Choix des données évaluées

[Hastie and Belz, 2014] décrivent trois types d'évaluation en GAT :

1. des composants de GAT en isolation (« NLG components in isolation »)
2. des systèmes de GAT « end-to-end »
3. des systèmes de GAT dans le cadre d'un système plus large (« embedded components »)

Nous avons décidé que notre évaluation finale serait une évaluation des textes générés dans leur intégralité, ce qui correspond donc à l'option 2. En effet, nous avions déjà eu l'opportunité d'effectuer des évaluations informelles de certains composants

du système (comme la gestion de la synonymie) de façon régulière au cours du stage. Il nous était donc plus utile d'évaluer le système dans son ensemble.

Par ailleurs, nous avons aussi tenté de prendre en compte dans la rédaction du formulaire le fait que les rapport générés automatiquement étaient ensuite intégrés à des présentations plus complexes incluant souvent des données numériques.

Nous devions également choisir un nombre raisonnable de textes à évaluer, dont la variété permettrait d'obtenir une vue d'ensemble de la performance du système. Comme le souligne [Reiter, 2017], il est possible de choisir aléatoirement un certain nombre de textes ou de scénarios à évaluer, ou bien de définir des catégories auxquelles appartiennent différents textes, et de choisir des textes appartenant à diverses catégories. Pour Reiter, cette distinction est liée à celle entre le cas moyen (*average case*) et le pire cas (*worst case*) : si l'objectif est d'évaluer la performance moyenne du système, une sélection aléatoire de textes est probablement préférable. Si l'objectif est de déterminer quelle est la pire performance du système, il vaut mieux choisir les textes évalués de façon à évaluer des textes très différents.

Comme établi dans le chapitre 1, dans le cadre de la génération de rapports d'analyse, il est essentiel de ne pas induire le lecteur en erreur (volontairement ou non) : par conséquent, nous avons choisi de sélectionner manuellement les textes évalués [1]. Ce choix se justifiait d'autant plus que la longueur et le contenu des textes peuvent varier de façon assez significative dans notre corpus généré.

Nous avons finalement choisi dix entreprises, et dix textes générés associés à ces entreprises, pour lesquelles nous avons généré un formulaire d'évaluation. Ce nombre a été choisi en partie après des discussions internes sur la durée de l'évaluation, et d'autre part parce qu'il permettait de représenter une certaine variété du corpus (table 6.1).

Nom	Sous-secteur	Début du suivi par Exane	Longueur (caractères)	Notation
COMP1	Food & HPC	avant 2009	592	(=)
COMP2	Aerospace & Defence	avant 2009	791	(+)
COMP3	IT Services	avant 2009	464	(+)
COMP4	Diversified Financials	avant 2009	485	(=)
COMP5	Pharmaceuticals	avant 2009	601	(=)
COMP6	Luxury Goods	avant 2009	671	(=)
COMP7	Food & HPC	avant 2009	480	(=)
COMP8	Luxury Goods	avant 2009	593	(=)
COMP9	Food & HPC	avant 2009	466	(+)
COMP10	IT Hardware	2014	320	(+)

TABLE 6.1 – Caractéristiques des textes sélectionnés pour l'évaluation

6.2.3 Déroulement de l'évaluation

Une fois les textes à évaluer sélectionnés, nous devions créer un formulaire d'évaluation accessible permettant aux évaluateurs de travailler rapidement et efficacement. Nous décrivons d'abord son format de façon générale, puis l'échelle de notation

1. De plus, le nombre limité de textes sélectionnés ne justifiait pas l'automatisation de cette tâche.

suggérée. Enfin, nous présentons notre réflexion sur les critères d'évaluation choisis. Un exemple de grille d'évaluation remplie est présenté en annexe (A.1).

Nous avons créé un fichier Excel par entreprise, chacun suivant le même modèle :

— La première feuille du classeur Excel contient un extrait de la grille financière de Cube, extrait le jour de la génération du texte [2]
— La deuxième feuille contient le texte généré complet en isolation
— La troisième contient le formulaire d'évaluation

Nous avons donc fait le choix de donner à l'évaluateur un accès permanent aux données ayant servi à la génération du texte (en les séparant néanmoins dans différentes feuilles). D'après [Novikova et al., 2018], bien que ces données soient nécessaires pour l'évaluation de l'exactitude d'un texte, elles peuvent brouiller l'évaluation de sa lisibilité. Cependant, il faut noter que le lien entre les données de la grille et le texte généré n'est pas direct : un nombre important de calculs intermédiaires a été effectué par le moteur de génération. La présence de la grille ne signifie donc que l'évaluateur doit l'utiliser de façon approfondie, mais plutôt pour confirmer son idée générale de la situation de l'entreprise.

1	2	3	4	5
very bad or inappropriate	fairly bad or inappropriate	acceptable, neutral	fairly good or appropriate	very good or appropriate

TABLE 6.2 – Échelle de Likert employée pour l'évaluation

Nous avons choisi une seule échelle de Likert à cinq réponses pour tous les critères d'évaluation, présentée dans la table 6.2. Ce choix a en partie été inspiré par le choix de [Reiter and Belz, 2009], dont la formulation incite cependant plus explicitement une comparaison entre différents textes (par exemple, « bien, mais vous avez déjà lu un meilleur rapport »). Le nombre impair de réponses signifie que la réponse 3 peut être considérée comme neutre. La double formulation des réponses (bon/mauvais et convenable/non-convenable) est en partie due au fait que cette échelle est employée pour juger de la lisibilité, l'exactitude et l'utilité du texte.

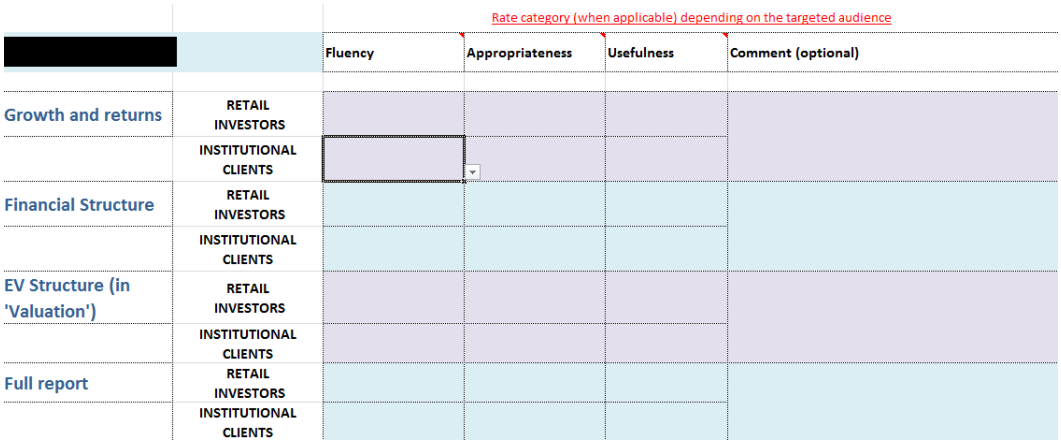


FIGURE 6.1 – Critères d'évaluation d'un texte (Excel)

2. Le but est d'éviter que les données de la grille et le texte soient désynchronisés après, par exemple, la mise à jour de la grille.

La grille d'évaluation (figure 6.1) permet d'évaluer chaque paragraphe du texte (« Growth and returns », « Financial structure », et « Valuation »), s'il a été généré, ainsi que le texte dans son ensemble (« Full report »). Pour chacun de ces paragraphes, on effectue une distinction entre les *investisseurs institutionnels* (« institutional clients ») et les *investisseurs particuliers* (« retail investors »), deux types de clients potentiels : dans le cadre de notre évaluation, on retiendra principalement que les investisseurs institutionnels attendent en général une analyse plus spécifique et détaillée de la situation d'une entreprise que les investisseurs particuliers.

L'évaluateur peut donc noter huit combinaisons d'extrait de texte et de type d'investisseur selon trois critères, la **lisibilité**, l'**exactitude** et l'**utilité**. Ces critères lui sont brièvement définis de façon simple et univoque [Bachelet, 2012], et nous revenons sur ces définitions dans la section suivante. Il également possible (mais pas obligatoire) de commenter sa notation dans un champ libre, afin de fournir aux évaluateurs la possibilité d'expliciter une difficulté rencontrée dans le texte.

6.3 Critères d'évaluation

6.3.1 Lisibilité

Le formulaire d'évaluation définit, sous forme de question, la notion de lisibilité ainsi :

> Does the text sound like an analyst report ?[3]

Nous avons choisi le terme « lisibilité » pour rassembler plusieurs notions anglophones proches mais distinctes : *clarity*, *fluency*, et *fidelity*. Comme fil conducteur, nous choisissons de définir la notion par rapport à la langue de spécialité afin de fournir un critère d'évaluation simple aux évaluateurs qui lisent un nombre important de rapports et connaissent bien cette langue de spécialité.

Cependant, cette notion n'est pas nécessairement clairement définie par tous les utilisateurs qui ne sont pas des terminologues et qui ne sont donc pas nécessairement en accord sur l'emploi de certaines tournures ou de certains termes[4] : comme on l'a vu dans la section 5.4, certains rapports rédigés par des humains sont considérés comme peu clairs par d'autres spécialistes de la langue.

6.3.2 Exactitude

Le formulaire d'évaluation définit, sous forme de question, la notion d'exactitude ainsi :

> Is the text an accurate/appropriate description of the grid ?[5]

« Exactitude » recouvre ici plusieurs termes : *completeness*, *relevance*, et *non-ambiguity*.

Nous avons déjà noté qu'il est essentiel qu'un rapport ne soit jamais faux, mais il est plus difficile d'évaluer l'ambiguité d'un texte (qu'il soit généré automatiquement ou non). Par exemple, [Dale and Reiter, 1995] utilisent les maximes gricéennes comme point d'appui pour la génération d'expressions référentielles claires et non-ambigües. Nous avons par ailleurs vu dans la section 5.2 que des problèmes d'ambiguité peuvent être causées par l'utilisation de séries de données temporelles.

3. Ce texte ressemble-t-il à un rapport d'analyse ?
4. Nous avons également noté plus haut l'influence de la langue natale d'un utilisateur.
5. Ce texte est-il une représentation exacte et pertinente de la grille ?

D'autre part, la question de l'exactitude d'une phrase n'est pas entièrement détachée de la question de sa complexité linguistique. Plus une phrase est linguistiquement et rhétoriquement simple, plus le risque d'ambiguité est faible ; néanmoins, on court le risque de générer des phrases considérées comme trop simples ou trop peu précises.

Enfin, il nous faut garder à l'esprit qu'un texte peut être entièrement correct mais incomplet si certaines informations essentielles de la grille n'ont pas été transmises par le texte.

6.3.3 Utilité

Le formulaire d'évaluation définit, sous forme de question, la notion d'utilité ainsi :

> Would this text be useful to the targeted audience ? [6]

L'utilité est probablement le critère le plus variable en fonction du projet considéré, mais aussi le plus difficile à évaluer indirectement (c'est-à-dire sans évaluer l'utilisation du système par ses usagers). Même s'il est possible de faire évaluer l'utilité d'un texte par des spécialistes non-usagers comme on se propose de le faire, la définition du critère : **quel est l'usage du projet ?**

Cependant, l'intérêt de ce critère est aussi d'inciter l'évaluateur à envisager le texte dans son ensemble à fournir une analyse de sa structure, mais aussi à comparer le texte à d'autres textes générés, dans la mesure où ces rapports ne sont pas lus de façon indépendante mais utilisés pour comparer différentes entreprises.

6.4 Analyse des résultats

6.4.1 Exploitation des formulaires

Nous avons recueilli en tout vingt formulaires. Toutes les personnes ayant participé à l'évaluation n'ont pas rempli l'intégralité des questionnaires (table 6.3), en grande partie par manque de temps. Certaines nous ont transmis des commentaires généraux sur le niveau de langue des textes en dehors des formulaires.

Par ailleurs, nous avons pu nous apercevoir, en parcourant les formulaires manuellement, de certaines faiblesses dans notre modèle d'évaluation qui avaient induit les évaluateurs en erreur :

- Il aurait fallu effectuer une distinction visuelle plus claire entre l'évaluation des différents paragraphes et l'évaluation du rapport complet : dans plusieurs cas, l'évaluateur n'a pas noté le rapport dans son ensemble
- Très peu de de formulaires présentent une différence de notation en fonction du type d'investisseur ciblé [7] : il aurait probablement été plus efficace de poser la question une seule fois par formulaire
- Certains évaluateurs n'ont pas rempli le formulaire dans son intégralité, sans nécessairement justifier cette décision

Malgré ces difficultés, nous avons pu recueillir une quantité raisonnable d'évaluations. Afin d'exploiter ces données d'évaluation de façon automatisée, nous avons

6. Ce texte serait-il utile pour son lectorat ?

7. Cette différence, quand elle existe, est par ailleurs souvent mineure (un point sur l'échelle de Likert).

	EN1_AN1	EN2_AN2	EN3_SA1	EN4_AN3	FR1_SA2
COMP1	✓	✓			
COMP2	✓		✓		
COMP3	✓		✓		
COMP4	✓				
COMP5	✓	✓	✓		✓
COMP6	✓		✓		
COMP7	✓	✓	✓		✓
COMP8	✓				
COMP9	✓				
COMP10	✓				
Commentaires généraux		✓	✓	✓	

Chaque évaluat·eur·rice est identifié·e par sa langue natale (EN/FR : anglais ou français) et son métier (AN/SA : analyse ou vente).

TABLE 6.3 – Répartition des formulaires d'évaluation

```
  appraiser stock_name            section          type_investor      type_rating                                    rating
0   FR1_SA2      COMP7  Growth and returns     RETAIL\n INVESTORS          fluency                                         5
1   FR1_SA2      COMP7  Growth and returns     RETAIL\n INVESTORS  appropriateness                                         2
2   FR1_SA2      COMP7  Growth and returns     RETAIL\n INVESTORS       usefulness                                         2
3   FR1_SA2      COMP7  Growth and returns     RETAIL\n INVESTORS     free_comment  les commentaires sur la croissance organique e...
4   FR1_SA2      COMP7  Growth and returns  INSTITUTIONAL CLIENTS          fluency                                         5
(560, 6)
```

FIGURE 6.2 – Représentation des données d'évaluation recueillies (Python)

extrait les résultats des tableurs Excel via un script Python afin de les stocker dans un DataFrame `pandas` de 560 lignes (figure 6.2). Nous avons également recueilli les commentaires dans un document à part pour pouvoir les parcourir plus en détails (table A.3).

6.4.2 Présentation des résultats

	Lisibilité	Exactitude	Utilité
Moyenne	3.7	3.37	3.2
Médiane	4.0	4.0	3.0

TABLE 6.4 – Moyenne et médiane des notes attribuées pour chaque critère (« Growth and returns »)

	Lisibilité	Exactitude	Utilité
Moyenne	3.6	3.3	3.0
Médiane	3.5	3.0	3.0

TABLE 6.5 – Moyenne et médiane des notes attribuées pour chaque critère (« Financial structure »)

Comme mentionné dans la section précédente, un grand nombre d'évaluateurs n'ont pas complété la section servant à l'évaluation du rapport complet. Au vu du

	Lisibilité	Exactitude	Utilité
Moyenne	3.5	1.5	1.4
Médiane	3.5	1.5	1.0

TABLE 6.6 – Moyenne et médiane des notes attribuées pour chaque critère (« EV structure/Valuation »)

nombre réduit de résultats , nous avons donc dû nous rabattre sur une analyse de notes de chaque paragraphe (tables 6.4,6.5, et 6.6). Nous avions formulé deux hypothèses générales :

— la lisibilité serait le critère où les résultats seraient les meilleurs
— la dernière section (« Valuation ») étant la moins développée, elle obtiendrait les pires résultats

Ces deux hypothèses sont dans l'ensemble confirmées. On notera que l'utilité est souvent le critère sur lequel les textes ont été le moins bien notés, ce qui peut être en partie dû au fait que c'est le critère le plus difficile à définir et évaluer. Les mauvais résultats de la section « Valuation » (table 6.6) sont explicables par le fait que cette partie n'était pas encore complétée au moment de l'évaluation (une remarque qui se retrouve dans les commentaires libres). Comme le précise le nom complet de la section, seul le paragraphe sur l'EV était généré au moment de l'évaluation, alors que la notion de « valuation » recouvre un champ beaucoup plus large. Ceci peut expliquer les mauvais résultats en exactitude et en utilité. D'autre part, cette section est la section qui nécessite le plus de comparaisons avec d'autres entreprises et l'explicitation d'informations stratégiques, ce qui peut aboutir à la génération de phrases moins satisfaisantes (cf. section 5.2.3).

Nous avons ensuité analysé les résultats en fonction des évaluateurs, afin de vérifier si certains avaient tendance à attribuer des notes plus hautes ou plus basses que la moyenne. En effectuant ce travail (table 6.7), nous avons pu nous apercevoir qu'un des évaluateurs attribuait les mêmes notes aux trois critères (cf. figure 6.3), d'où les résultats obtenus dans la table. Nous pouvons confronter ces résultats aux hypothèses formulées en 6.2.1 (en gardant à l'esprit que tous les évaluateurs n'ont pas fourni le même nombre de données) :

— il semblerait que les choix personnels de notation de l'évaluateur soient plus significatifs que son métier, au vu de l'écart entre les notes attribuées par EN1_AN1 et EN2_AN2
— l'évaluateur francophone (FR1_SA2) a bien attribué des notes de lisibilité relativement hautes, mais il est difficile d'en tirer des conclusions définitives au vu du nombre limité de données recueillies

	EN1_AN1	EN2_AN2	EN3_SA1	FR1_SA2
Lisibilité	3.4	2.5	4.3	4.3
Exactitude	3.4	2.3		2.6
Utilité	3.4	2.2	2	2.6

Un des évaluateurs n'a jamais évalué l'exactitude du texte, mais n'a pas fourni d'explication à ce choix.

TABLE 6.7 – Moyenne des notes attribuées par chaque évaluateur

		Fluency	Appropriateness	Usefulness
Growth and returns	RETAIL INVESTORS	4	4	4
	INSTITUTIONAL CLIENTS	4	4	4
Financial Structure	RETAIL INVESTORS	3	3	3
	INSTITUTIONAL CLIENTS	3	3	3
EV Structure (in 'Valuation')	RETAIL INVESTORS	2	2	2
	INSTITUTIONAL CLIENTS	2	2	2

FIGURE 6.3 – Exemple d'un formulaire rempli par EN1_AN1

Enfin, en observant les notes obtenues par chaque rapport, nous pouvons relever que certaines entreprises sont généralement mieux notés, tandis que d'autres sont généralement considérés comme insuffisants, comme par exemple le rapport sur COMP10, ce qui s'explique en partie par le fait que COMP10 est une entreprise relativement récente dont le traitement par Exane n'a commencé qu'en 2014 (cf. table 6.1). Ce manque de données peut expliquer les difficultés du système à produire un rapport satisfaisant.

	Lisibilité	Exactitude	Utilité
COMP1	2.8	2.7	2.6
COMP2	3.6	3.5	3.1
COMP3	4.1	3.25	2.3
COMP4	3.0	3.0	3.0
COMP5	3.4	2.9	2.7
COMP6	4.2	4.0	2.8
COMP7	4.2	2.7	2.5
COMP8	4.0	4.0	4.0
COMP9	3.3	3.3	3.3
COMP10	2.0	2.0	2.0

TABLE 6.8 – Moyenne des notes attribuées à chaque entreprise

6.4.3 Que retenir de l'évaluation ?

En reprenant les résultats observés dans la section précédente et en les accompagnant d'une lecture des commentaires libres, nous pouvons relever plusieurs grandes lignes d'amélioration possibles :

— nous avons reçu plusieurs commentaires extrêmement spécifiques de lisibilité (par exemple à propos de termes comme « go up », qui n'était pas considéré comme un synonyme approprié de monter), ce qui renforce notre analyse précédente sur l'importance du choix lexical pour la lisibilité du texte
— beaucoup de commentaires reprochaient au texte de ne pas assez expliciter le raisonnement derrière les phrases générées, ce qui les rend moins crédibles pour le lecteur
— ce problème était souvent associé à l'aspect « générique » des textes : certains commentaires soulignaient que certaines phrases étaient tellement génériques que leur apport sémantique était minimal

D'un point de vue technique, une solution couramment employée par des systèmes de GAT commerciale et qui pourrait être une bonne solution pour la gestion des problèmes de lisibilité est la correction des textes post-génération par un humain spécialiste du domaine, comme le suggèrent [Reiter et al., 2005].

Il est également possible d'apporter des améliorations au protocole d'évaluation proposé. Idéalement, comme on l'a vu, il faudrait pouvoir faire évaluer les textes par des utilisateurs. Pour s'approcher de cette situation idéale, nous pourrions envisager de reproduire au plus près le format final dans lequel ces rapports sont présentés, c'est-à-dire entourés par une sélection de la grille financière, mais aussi parfois par du texte rédigé par des analystes. Ce format nous permettrait notamment de nous intéresser au rapport entre le texte et l'image dans le cadre de la GAT, un sujet déjà abordé sous plusieurs aspects :

— la capacité de la GAT à générer des textes plus synthétiques que des graphes [van der Meulen et al., 2010]
— la gestion des sous-entendus involontaires dans une présentation textuelle et graphique [Marks and Reiter, 1990]
— l'utilité de la GAT pour l'aide à la prise de décisions [Gkatzia et al., 2016])

Enfin, nous pouvons aussi envisager des types d'évaluation fondées sur un principe entièrement différent, comme [Sambaraju et al., 2011] qui favorise une analyse rhétorique des textes générés, ou [Wang et al., 2018] qui se propose d'évaluer l'exactitude d'un texte en reconstruisant la base de données ayant servi à la générer.

6.5 Conclusion

Nous avons décrit et critiqué le processus d'évaluation humaine de notre système de génération que nous avons mis en place. En décrivant ce processus, nous avons tenté de répondre aux questions suggérées par [Mellish and Pan, 2008] :

— quelles sont les données d'entrée ?
— quelles sont les données de sortie ?
— quels sont les critères d'évaluation ?
— comment comparer les résultats ?
— comment choisir nos données ?
— comment évaluer les différences de jugement entre évaluateurs ?

Nous avons traité et présenté les résultats de cette évaluation et proposé des améliorations qui pourraient permettre d'obtenir des résultats plus significatifs.

Conclusion finale

Dans ce mémoire, nous avons décrit le développement, le fonctionnement et l'évaluation d'un système de GAT produisant des rapports simplifiés d'analyse financière.

Nous avons défini un état de l'art de la recherche en GAT, que nous avons relié à notre projet de recherche autour de trois grandes problématiques essentielles pour un système de GAT dans un domaine spécialisé comme la finance de marché :

— l'exploitation de données non-textuelles complexes et variées qui doivent être modélisées de façon exploitable par le système
— la division du processus de génération en différents modules plus ou moins indépendants
— les critères et méthodes d'évaluation des textes générés, qui doivent permettre d'évaluer leur cohérence et leur visée discursives

Notre recherche nous a amenée à définir trois aspects spécifiques du projet ayant une influence sur la rédaction de règles de génération et sur le système en général.

D'une part, nous avons établi qu'un domaine comme la finance de marché possède une langue spécialisée dont il faut comprendre les caractéristiques pour pouvoir la reproduire pour les lecteurs des textes générés.

Par ailleurs, nous avons insisté sur l'importance de la notion de cohérence temporelle, qui est une part fondamentale de la cohérence générale du texte, et que nos règles de génération doivent conserver, particulièrement dans un domaine comme la finance qui définit des marqueurs temporels spécifiques.

Enfin, nous avons montré que nos règles de génération et de structuration de textes étaient influencées par différentes classifications et hiérarchisations d'entreprises qui définissent le contexte dans lequel le texte est lu et compris.

Parmi les points difficiles abordés lors de notre travail de recherche, nous pouvons citer la gestion de la génération d'expressions numériques, d'autant plus dans des textes financiers qui en contiennent beaucoup. Nous avons également abordé la notion de synonymie et son rapport à la notion de terme d'un domaine spécialisé. De plus, nous avons observé que la structuration d'un texte, particulièrement les paragraphes composant un texte sont fortement hétérogènes, est une facteur essentiel de la cohérence du texte.

Nous avons également proposé et décrit un protocole d'évaluation humaine complet, où plusieurs spécialistes du domaine financier ont à la fois noté et commenté de façon libre des textes générés. Nous avons recueilli ces réponses et analysé ces résultats pour repérer des points d'amélioration et d'expansion possibles de notre système. Nous avons aussi proposé des façons d'améliorer le protocole d'évaluation lui-même après avoir observé ses faiblesses.

Perspectives

Si nous avions l'opportunité d'approfondir ce travail de recherche, outre la poursuite de notre recherche sur les difficultés mentionnées dans cette conclusion, nous

aimerions en particulier confronter notre travail sur l'évaluation humaine au nouveaux articles de recherche publiés après la douzième International Conference on Natural Language Generation de début novembre 2019 [8], qui présentent des suggestions de bonnes pratiques pour l'évaluation humaine en GAT [9], et l'usage des échelles de Likert dans les évaluations de GAT [10].

8. www.inlg2019.com.
9. www.inlg2019.com/assets/papers/98_Paper.pdf.
10. www.inlg2019.com/assets/papers/57_Paper.pdf.

BIBLIOGRAPHIE

[Androutsopoulos et al., 2013] Androutsopoulos, I., Lampouras, G., and Galanis, D. (2013). Generating Natural Language Descriptions from OWL Ontologies: the NaturalOWL System. *Journal of Artificial Intelligence Research*, 48:671–715. – Cité page 24.

[Aoki et al., 2018] Aoki, T., Miyazawa, A., Ishigaki, T., Goshima, K., Aoki, K., Kobayashi, I., Takamura, H., and Miyao, Y. (2018). Generating Market Comments Referring to External Resources. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 135–139, Tilburg University, The Netherlands. Association for Computational Linguistics. – Cité pages 25, 34 et 35.

[Appelt, 1985] Appelt, D. E. (1985). *Planning English Sentences*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK. – Cité page 27.

[Avenel et al., 2017] Avenel, J.-D., du Castel, V., Jablanczy, A., Kelly, M., and Vesterman, C. (2017). *Finance de marché français-anglais*. Gualino. – Cité page 11.

[Bachelet, 2012] Bachelet, R. (2012). ABC de la méthodologie de la recherche : cours de recueil, analyse et traitement des données. http://rb.ec-lille.fr/Cours_de_recueil_analyse_et_traitement_de_donnees.htm. – Cité page 73.

[Bateman and Zock, 2014] Bateman, J. and Zock, M. (2014). Natural Language Generation. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*, pages 284–304. Oxford University Press, Oxford, United Kingdom, 2nd edition. – Cité page 23.

[Belz and Gatt, 2008] Belz, A. and Gatt, A. (2008). Intrinsic vs. Extrinsic Evaluation Measures for Referring Expression Generation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 197–200, Columbus, USA. Association for Computational Linguistics. – Cité pages 31 et 32.

[Binsted and Ritchie, 1994] Binsted, K. and Ritchie, G. (1994). An Implemented Model of Punning Riddles. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1)*, AAAI '94, pages 633–638, Menlo Park, USA. American Association for Artificial Intelligence. – Cité page 24.

[Bouayad-Agha et al., 2011] Bouayad-Agha, N., Casamayor, G., and Wanner, L. (2011). Content selection from an ontology-based knowledge base for the generation of football summaries. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 72–81, Nancy, France. Association for Computational Linguistics. – Cité page 25.

[Bourbeau et al., 1990] Bourbeau, L., Carcagno, D., Goldberg, E., Kittredge, R., and Polguère, A. (1990). Bilingual Generation of Weather Forecasts in an Operations Environment. In *Proceedings of the 13th Conference on Computational Linguistics*

- *Volume 1*, COLING '90, pages 90–92, Helsinki, Finland. Association for Computational Linguistics. – Cité page 25.

[Braun et al., 2018] Braun, D., Reiter, E., and Siddharthan, A. (2018). SaferDrive: An NLG-based behaviour change support system for drivers. *Natural Language Engineering*, 24(4):551–588. – Cité page 25.

[Castro Ferreira et al., 2018] Castro Ferreira, T., Moussallem, D., Kádár, A., Wubben, S., and Krahmer, E. (2018). NeuralREG: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1969, Melbourne, Australia. Association for Computational Linguistics. – Cité page 28.

[Castro Ferreira et al., 2019] Castro Ferreira, T., van der Lee, C., van Miltenburg, E., and Krahmer, E. (2019). Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 552–562, Hong Kong, China. Association for Computational Linguistics. – Cité page 28.

[Chang and Manning, 2014] Chang, A. X. and Manning, C. D. (2014). TokensRegex: Defining cascaded regular expressions over tokens. Technical Report CSTR 2014-02, Department of Computer Science, Stanford University. – Cité pages 65 et 89.

[Cimiano et al., 2013] Cimiano, P., Lüker, J., Nagel, D., and Unger, C. (2013). Exploiting Ontology Lexica for Generating Natural Language Texts from RDF Data. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 10–19, Sofia, Bulgaria. Association for Computational Linguistics. – Cité page 26.

[Dale and Reiter, 1995] Dale, R. and Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263. – Cité page 73.

[Danlos and Roussarie, 2000] Danlos, L. and Roussarie, L. (2000). La génération automatique de textes. In Pierrel, J.-M., editor, *Ingénierie des Langues*, Traité IC2 (Information, communication et commande). Hermès, Paris, France. – Cité pages 26 et 38.

[Dušek et al., 2020] Dušek, O., Novikova, J., and Rieser, V. (2020). Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2e NLG challenge. *Computer Speech & Language*, 59:123–156. – Cité pages 6, 28 et 30.

[Evans et al., 2002] Evans, R., Piwek, P., and Cahill, L. (2002). What is NLG? In *Proceedings of the International Natural Language Generation Conference*, pages 144–151, Harriman, USA. Association for Computational Linguistics. – Cité page 23.

[Fisher et al., 2016] Fisher, I. E., Garnsey, M. R., and Hughes, M. E. (2016). Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research. *Intelligent Systems in Accounting, Finance and Management*, 23:157–214. – Cité page 34.

[Friedman et al., 2002] Friedman, C., Kra, P., and Rzhetsky, A. (2002). Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235. – Cité page 38.

[Gao et al., 2019] Gao, J., Galley, M., and Li, L. (2019). Neural Approaches to Conversational AI. *Foundations and Trends in Information Retrieval*, 13(2-3):127–298. – Cité page 28.

[Garbe, 2019] Garbe, J. (2019). StoryAssembler: An Engine for Generating Dynamic Choice-Driven Narratives. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, volume 24 of *FDG '19*, pages 1–10, San Luis Obispo, USA. Association for Computing Machinery. – Cité page 25.

[Gardent et al., 2017] Gardent, C., Shimorina, A., Narayan, S., and Perez-Beltrachini, L. (2017). The WebNLG Challenge: Generating Text from RDF Data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics. – Cité page 26.

[Gatt and Krahmer, 2018] Gatt, A. and Krahmer, E. (2018). Survey of the State of the Art in Natural Language Generation: Core Tasks, Applications and Evaluation. *Journal of Artificial Intelligence Research*, 61:65–170. – Cité pages 25, 26, 29, 31 et 55.

[Gatt and Reiter, 2009] Gatt, A. and Reiter, E. (2009). SimpleNLG: A Realisation Engine for Practical Applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93, Athens, Greece. Association for Computational Linguistics. – Cité pages 21 et 89.

[Gkatzia et al., 2016] Gkatzia, D., Lemon, O., and Rieser, V. (2016). Natural Language Generation Enhances Human Decision-Making with Uncertain Information. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 264–268, Berlin, Germany. Association for Computational Linguistics. – Cité page 78.

[Gkatzia and Mahamood, 2015] Gkatzia, D. and Mahamood, S. (2015). A Snapshot of NLG Evaluation Practices 2005 - 2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 57–60, Brighton, United Kingdom. Association for Computational Linguistics. – Cité page 29.

[Hastie and Belz, 2014] Hastie, H. and Belz, A. (2014). A Comparative Evaluation Methodology for NLG in Interactive Systems. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 4004–4011, Reykjavik, Iceland. European Language Resources Association. – Cité pages 6, 29, 30 et 70.

[Hunter et al., 2012] Hunter, J., Freer, Y., Gatt, A., Reiter, E., Sripada, S., and Sykes, C. (2012). Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial Intelligence in Medicine*, 56(3):157–172. – Cité pages 24, 25, 32 et 59.

[Hunter, 2007] Hunter, J. D. (2007). Matplotlib: A 2d Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95. – Cité pages 89 et 97.

[Jurafsky and Martin, 2009] Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Education, Inc., Upper Saddle River, USA, 2nd edition. – Cité page 64.

[Kittredge et al., 1994] Kittredge, R., Kim, M., Goldberg, E., and Polguère, A. (1994). Sublanguage Engineering in the FoG System. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, ANLC '94, pages 215–216, Stuttgart, Germany. Association for Computational Linguistics. – Cité pages 38 et 63.

[Kukich, 1983] Kukich, K. (1983). Design of a Knowledge-based Report Generator. In *Proceedings of the 21st Annual Meeting on Association for Computational Linguistics*, ACL '83, pages 145–150, Cambridge, USA. Association for Computational Linguistics. – Cité pages 25 et 34.

[Langkilde and Knight, 1998] Langkilde, I. and Knight, K. (1998). Generation that Exploits Corpus-Based Statistical Knowledge. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 704–710, Montréal, Canada. Association for Computational Linguistics. – Cité page 28.

[L'Homme, 2018] L'Homme, M.-C. (2018). *La terminologie : principes et techniques*. Paramètres. Presses de l'Université de Montréal, Montréal, Canada. – Cité pages 25, 36 et 37.

[Lin, 2004] Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. – Cité page 31.

[Liu et al., 2016] Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, USA. Association for Computational Linguistics. – Cité pages 6 et 32.

[Lopez, 2013] Lopez, S. (2013). *Norme(s) et usage(s) langagiers : le cas des communications pilote-contrôleur en anglais*. Thèse, Université Toulouse 2, Toulouse, France. – Cité pages 36, 37 et 38.

[Mahamood and Reiter, 2011] Mahamood, S. and Reiter, E. (2011). Generating Affective Natural Language for Parents of Neonatal Infants. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 12–21, Nancy, France. Association for Computational Linguistics. – Cité page 25.

[Manning et al., 2014] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60. – Cité pages 64 et 89.

[Manor and Li, 2019] Manor, L. and Li, J. J. (2019). Plain English Summarization of Contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11, Minneapolis, USA. Association for Computational Linguistics. – Cité page 24.

[Marks and Reiter, 1990] Marks, J. and Reiter, E. (1990). Avoiding Unwanted Conversational Implicatures in Text and Graphics. In *Proceedings of the Eighth National Conference on Artificial Intelligence - Volume 1*, AAAI'90, pages 450–456, Boston, USA. AAAI Press. – Cité page 78.

[Mason et al., 2019] Mason, S., Stagg, C., and Wardrip-Fruin, N. (2019). Lume: A System for Procedural Story Generation. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, FDG '19, pages 15:1–15:9, San Luis Obispo, USA. Association for Computing Machinery. – Cité page 25.

[Mauldin, 1984] Mauldin, M. L. (1984). Semantic Rule Based Text Generation. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 376–380, Stanford, USA. Association for Computational Linguistics. – Cité page 26.

[Mellish and Dale, 1998] Mellish, C. and Dale, R. (1998). Evaluation in the context of natural language generation. *Computer Speech & Language*, 12(4):349–373. – Cité pages 29 et 31.

[Mellish and Pan, 2008] Mellish, C. and Pan, J. Z. (2008). Natural Language Directed Inference From Ontologies. *Artificial Intelligence*, 172(10):1285–1315. – Cité page 78.

[Meteer, 1991] Meteer, M. W. (1991). Bridging the generation gap between text planning and linguistic realization. *Computational Intelligence*, 7(4):296–304. – Cité page 26.

[Moeschler and Auchlin, 2018] Moeschler, J. and Auchlin, A. (2018). *Introduction à la linguistique contemporaine*. Armand Colin, Malakoff, France. – Cité page 11.

[Moryossef et al., 2019] Moryossef, A., Goldberg, Y., and Dagan, I. (2019). Step-by-Step: Separating Planning from Realization in Neural Data-to-Text Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2267–2277, Minneapolis, USA. Association for Computational Linguistics. – Cité page 28.

[Murakami et al., 2017] Murakami, S., Watanabe, A., Miyazawa, A., Goshima, K., Yanase, T., Takamura, H., and Miyao, Y. (2017). Learning to Generate Market Comments from Stock Prices. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1 (Long Papers), pages 1374–1384, Vancouver, Canada. Association for Computational Linguistics. – Cité pages 25, 34 et 35.

[Nguyen, 2013] Nguyen, T. A. T. (2013). *Generating Natural Language Explanations For Entailments In Ontologies*. Thèse, The Open University, United Kingdom. – Cité page 24.

[Novikova et al., 2017] Novikova, J., Dušek, O., and Rieser, V. (2017). Data-driven Natural Language Generation: Paving the Road to Success. *arXiv:1706.09433 [cs]*. WiNLP workshop at ACL 2017. – Cité page 28.

[Novikova et al., 2018] Novikova, J., Dušek, O., and Rieser, V. (2018). RankME: Reliable Human Ratings for Natural Language Generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, USA. Association for Computational Linguistics. – Cité page 72.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*,

pages 311–318, Philadelphia, USA. Association for Computational Linguistics. – Cité page 31.

[Portet et al., 2009] Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., and Sykes, C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7):789–816. – Cité pages 24, 40, 58 et 63.

[Power and Williams, 2011] Power, R. and Williams, S. (2011). Generating Numerical Approximations. *Computational Linguistics*, 38(1):113–134. – Cité page 68.

[Puzikov and Gurevych, 2018] Puzikov, Y. and Gurevych, I. (2018). E2e NLG Challenge: Neural Models vs. Templates. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 463–471, Tilburg University, The Netherlands. Association for Computational Linguistics. – Cité page 28.

[Reiter, 2017] Reiter, E. (2017). How to do an NLG Evaluation: Human Ratings in Artificial Context. http://ehudreiter.com/2017/01/09/human-ratings-nlg-evaluation. – Cité page 71.

[Reiter, 2018a] Reiter, E. (2018a). Hallucination in Neural NLG. http://ehudreiter.com/2018/11/12/hallucination-in-neural-nlg. – Cité pages 6 et 30.

[Reiter, 2018b] Reiter, E. (2018b). A Structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401. – Cité page 32.

[Reiter and Belz, 2009] Reiter, E. and Belz, A. (2009). An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*, 35(4):529–558. – Cité pages 31 et 72.

[Reiter and Dale, 1997] Reiter, E. and Dale, R. (1997). Building Applied Natural Language Generation Systems. *Natural Language Engineering*, 3(1):57–87. – Cité pages 23 et 26.

[Reiter and Dale, 2000] Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, United Kingdom. – Cité pages 5, 25, 26 et 27.

[Reiter et al., 2003] Reiter, E., Robertson, R., and Osman, L. M. (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1):41–58. – Cité pages 25 et 32.

[Reiter et al., 2005] Reiter, E., Sripada, S., Hunter, J., Yu, J., and Davy, I. (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1):137–169. – Cité pages 25, 28, 37, 51, 67 et 78.

[Sambaraju et al., 2011] Sambaraju, R., Reiter, E., Logie, R., McKinlay, A., McVittie, C., Gatt, A., and Sykes, C. (2011). What is in a text and what does it do: Qualitative Evaluations of an NLG system – the BT-Nurse – using content analysis and discourse analysis. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 22–31, Nancy, France. Association for Computational Linguistics. – Cité pages 59 et 78.

[Smiley et al., 2016] Smiley, C., Plachouras, V., Schilder, F., Bretz, H., Leidner, J., and Song, D. (2016). When to Plummet and When to Soar: Corpus Based Verb

Selection for Natural Language Generation. In *Proceedings of the 9th International Natural Language Generation conference*, pages 36–39, Edinburgh, United Kingdom. Association for Computational Linguistics. – Cité pages 35, 51, 64, 65 et 67.

[Smiley et al., 2017] Smiley, C., Schilder, F., Plachouras, V., and Leidner, J. L. (2017). Say the Right Thing Right: Ethics Issues in Natural Language Generation Systems. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 103–108, Valencia, Spain. Association for Computational Linguistics. – Cité pages 25 et 59.

[Smith et al., 2016] Smith, A. M., Katz, D. S., and Niemeyer, K. E. (2016). Software Citation Principles. *PeerJ Computer Science*, 2:e86. – Cité page 89.

[Spärck Jones and Galliers, 1995] Spärck Jones, K. and Galliers, J. R. (1995). *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer. – Cité page 31.

[Tintarev et al., 2016] Tintarev, N., Reiter, E., Black, R., Waller, A., and Reddington, J. (2016). Personal storytelling: Using Natural Language Generation for children with complex communication needs, in the wild.... *International Journal of Human-Computer Studies*, 92-93:1–16. – Cité pages 24 et 26.

[Turing, 1950] Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(October):433–60. – Cité page 24.

[van Deemter et al., 2005] van Deemter, K., Krahmer, E., and Theune, M. (2005). Squibs and Discussions: Real versus Template-Based Natural Language Generation: A False Opposition? *Computational Linguistics*, 31(1):15–24. – Cité page 28.

[van Deemter and Reiter, 2018] van Deemter, K. and Reiter, E. (2018). Lying and Computational Linguistics. In Meibauer, J., editor, *The Oxford Handbook of Lying*, pages 420–435. Oxford University Press, Oxford, United Kingdom. – Cité page 25.

[van der Lee et al., 2017] van der Lee, C., Krahmer, E., and Wubben, S. (2017). PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104, Santiago de Compostela, Spain. Association for Computational Linguistics. – Cité pages 21, 24 et 25.

[van der Lee et al., 2018] van der Lee, C., Krahmer, E., and 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 Conference on Natural Language Generation*, pages 35–45, Tilburg University, The Netherlands. Association for Computational Linguistics. – Cité page 28.

[van der Meulen et al., 2010] van der Meulen, M., Logie, R. H., Freer, Y., Sykes, C., McIntosh, N., and Hunter, J. (2010). When a graph is poorer than 100 words: A comparison of computerised natural language generation, human generated descriptions and graphical displays in neonatal intensive care. *Applied Cognitive Psychology*, 24(1):77–89. – Cité page 78.

[Wang et al., 2018] Wang, Q., Pan, X., Huang, L., Zhang, B., Jiang, Z., Ji, H., and Knight, K. (2018). Describing a Knowledge Base. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 10–21, Tilburg

University, The Netherlands. Association for Computational Linguistics. – Cité page 78.

[Weizenbaum, 1966] Weizenbaum, J. (1966). ELIZA—a Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 9(1):36–45. – Cité page 24.

[White, 2005] White, M. (2005). Designing an Extensible API for Integrating Language Modeling and Realization. In *Proceedings of Workshop on Software*, pages 47–64, Ann Arbor, Michigan. Association for Computational Linguistics. – Cité page 21.

[Williams and Power, 2013] Williams, S. and Power, R. (2013). Hedging and rounding in numerical expressions. *Pragmatics & Cognition*, 21(1):193–223. – Cité page 68.

[Xie, 2017] Xie, Z. (2017). Neural Text Generation: A Practical Guide. *arXiv:1711.09534 [cs, stat]*. – Cité page 28.

[Xing et al., 2018] Xing, F. Z., Cambria, E., and Welsch, R. E. (2018). Natural Language Based Financial Forecasting: A Survey. *Artificial Intelligence Review*, 50(1):49–73. – Cité pages 34 et 35.

[Yngve, 1961] Yngve, V. (1961). Random Generation of English Sentences. In *1961 International Conference on Machine Translation of Languages and Applied Language Analysis*, pages 66–80, Teddington, United Kingdom. – Cité page 24.

[Yu et al., 2007] Yu, J., Reiter, E., Hunter, J., and Mellish, C. (2007). Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13(1):25–49. – Cité pages 24 et 26.

[Zock and Sabah, 1992] Zock, M. and Sabah, G. (1992). La génération automatique de textes : trente ans déjà, ou presque. *Langages*, 26(106):8–35. – Cité pages 23 et 24.

ANNEXE A

DOCUMENTATION ET RÉFÉRENCES

A.1 Langages et bibliothèques

A.1.1 Langages de programmation

Ce mémoire a été mis en page avec LaTeX.

Nom	Version
Excel	2007
Java	8
Python	3.7

TABLE A.1 – Logiciels et langages utilisés dans le cadre de ce mémoire

A.1.2 Bibliothèques et APIs [Smith et al., 2016]

Nom	Version	Référence
matplotlib	3.1.1	[Hunter, 2007]
numpy	1.17	`numpy.org`
pandas	0.25	`www.pandas.pydata.org`
pickle		`docs.python.org/3/library/pickle.html`
SimpleNLG	3	[Gatt and Reiter, 2009]
StanfordCoreNLP	3.8	[Manning et al., 2014]
TokensRegex		[Chang and Manning, 2014]
xlrd	1.2	`www.python-excel.org/`

TABLE A.2 – Bibliothèques utilisées dans le cadre de ce mémoire

A.2 Évaluation

La figure A.1 est une capture d'écran du formulaire d'évaluation décrit dans le chapitre 6 [1]. Il est présenté dans un tableur Excel qui contient également le texte généré (dans une autre feuille du tableur), ainsi que la grille de résultats de l'entreprise concernée extraite de Cube. L'évaluat·eur·ice peut choisir entre 1 et 5 dans un menu

1. Le nom de l'évaluateur et de l'entreprise évaluée ont été caviardés.

déroulant dans chacune des cellules. Une description plus complète des catégories (« fluency », etc.) est visible en survolant la cellule en question. Tous les formulaires remis n'étaient pas remplis complètement, soit parce que le texte ne s'y prêtait guère, soit parce que le formulaire manquait de clarté.

La table A.3 recense les commentaires libres recueillis via ces formulaires.

Commentaire
les commentaires sur la croissance organique et la croissance des EPS sont clés (projections vs moyenne historique).
Il manque le commentaire sur le niveau de la dette et le ratio ND/Ebitda
Qui a-t-il derrière les réévaluations d'actifs ? Le montant est significatif. Donner les ratio de valo eg PE actuel vs PE historique avec décote/prime vs historique aurait du sens
Les commentaires sont trop succincts. Ils ne permettent pas de se faire une idée sur la société et d'avoir une opinion
- les commentaires historiques sur la conversion de cash sont bons mais je rajouterais notre projection sur les 3 prochaines années . Je mettrais un commentaire systematique sur la croissance organique (passée vs projectée) et sur les croissances des EPS (passée vs projetée)
Je rajouterais des commentaires sur le ratio Net Debt/Ebitda et son évolution
Peut-être faudrait -il rappeler ce qui intervient de le calcul de l'EV. Dans le cas de COMP il y a bcp d'interêts minoritaires D'ailleurs pourquoi ne pas rajouter un commentaire sur la valo vs peers 15.9xPE19 vs peers at XXXX en indiquant la decote ou prime moyenne historique vs peers
Ca manque de précision dans les commentaires. Normalement on lisant les commentaires, on doit être capable de déviner le secteur et les spécificités de la société. Les commentaires sont trop génériques
ROCE was at 15% - when ?... Expect a declining EPS in 2019 - why ?... FCF should remain flat - why ?
"Especially indebted" seems poor English. Also 0.8£ doesn't scan well.
Above it was especially indebted. Here it appears to have limited financial leverage. I would like metrics to back-up statements. . . . No valuation provided.
The growth and returns section was one of the more complete / useful of those I've seen across the examples given to me. This was because it gave comments on FCF, returns and EPS.
What about returns and EPS ?... The English doesn't quite scan correctly in my view - for ex. "mainly driven by growth in EBITDA" might read better.
"Decidedly solid" seems odd English to me. . . . Should we talk about what the level of net debt : ebitda is expected to fall to rather than just the spot year level ?
Should it read 'predominantly pensions' not 'pensions predominantly'.
Financial Structure comments seem ok here.
Missing views on returns and earnings. . . Should we consider the absolute FCF growth rather than the CAGR ?
This is a useful comment on dividends. Is it really a 'financial structure' comment though ?
Information of very limited use.
The dividend comment stand-alone was a good standard.

FCF is absolute growth or CAGR best to use?... On EPS the I like the shape of the comment but it doesn't read as well as it could. Specifically the sentence which starts "However".... ROE comment also doesn't read perfectly... .Like the point about ROCE not being that relevant in the commentary.
Is the dividend at Eur5.2 good or bad in terms of yield / relative to the past etc etc.... Should it not be Eur5.2 not 5.2Eur.... There is a lack of other essential financial information in this section eg net debt : ebitda.
Lacks any detail
What is driving the EPS turnaround... What about returns?.. .Should we only consider EPS growth over 1 year?
Dividend increase is from what level to Eur2?.... Need to have stats next to commentary like "the financial situation" is improving - for example Net Debt : EBITDA yr 1 and yr 3.
Needs more than just a financial leverage comment.
Language : 1) better to say increase, rise, progress instead of 'go up'; 2) 'should' indicates an obligation. Better to say 'is expected to', or 'we expect / estimate/ forecast ...'
Language 'went up' see comment above; Content : 1) divis increased every year (the 'maintained' is not necessary here), 2) payout ration above 68% in 2015 –> the precise number sounds odd anyway, would be better to use 'rounded' figures (eg 70%); 3) sentence on estimated dividend is clumsy 'we expect this trend to continue' while 10% marks an acceleration vs 4% pa on average in the past
Makes no sense
1) Capex comment is true for 2015/16 but not for 2017/18 when capex normalized. 2) EPS comment wrong. 2019 accelerates vs 2018 and vs average growth trend (c.10% since 2011 ie excluding the >40% growth in 2010). Even incl. 2011 in the average, the 2019 growth marks not a slowdown but is in line.
1) Comment is too vague (financial situation is improving); not clear whether this is based on last few years (ND/EBITDA looks fairly stable though) or estimated years, 2) dividend comment : first sentence on maintaining/growing dividend is not necessary given later comment of av dividend growth. Payout ratio can be rounded
Not meaningful
Language : 1) due to 'high' capex levels. 2) 'Should' indicates an obligation –> better to use we expect / we forecast / we anticipate / we estimate, 3)EPS : instead of 'going up' it is better use 'increase', grow, progress. 4) 'expect a declining EPS' –> better to say 'we expect EPS to decline' or '..a decline in EPS', 5) ROCE comment sounds a bit clumsy. We could say 'COMP boasts a ROCE of XXX, significantly above ...'
Comment on indebtedness not partlicularly meaningful (particularly indebted compared to peers? / compared to its historic debt level? Based on ND/EBITDA or ND/MC? Indebted as a result of acquisitions?
Not meaningful
Need to contextualise why we think it can maintain a dividend when leverage is so high. What is leverage? What is payout?
Not sure this is particularly value add
Comments on capex/sales historically would have been helpful to contextualise the FCF increase

Very good
Not sure this has added much value
Some commentary on forward expectations may be helpful
Contextualise leverage and payout ratio
Very clear, succinct and accurrate
NA - no real comment on EV
NA - no full report section
Like commentary on payout
Perfect
Commentary on the dividend growth and the yield would be helpful
Some more context of capex/sales would be useful and what is preventing FCF to grow. But generally ok.
Some context on scale of leverage and dividend payout would be useful.
Clear
Fairly generic commentary
A little better than Growth and returns, especially for retail investors.
Na - no real section here
Not enough commentary
I only did three but the main issues are fairly repetitive ... so I think it gives you an idea.
Looking briefly at the outstanding list of companies they would receive similar comments. In short the fluency is not far off. The detail is nonetheless lacking in my view with many comments unsubstantiated.
Go up/going up is too colloquial. It is better to say “increase/will increase”
Financial “position” rather than “situation”
“Has” limited financial leverage – rather than “uses”
[company] remains “heavily” indebted... to me, especially/particularly seem appropriate in the context of comparing one company to another

TABLE A.3 – Liste des commentaires libres recueillis lors de l'évaluation

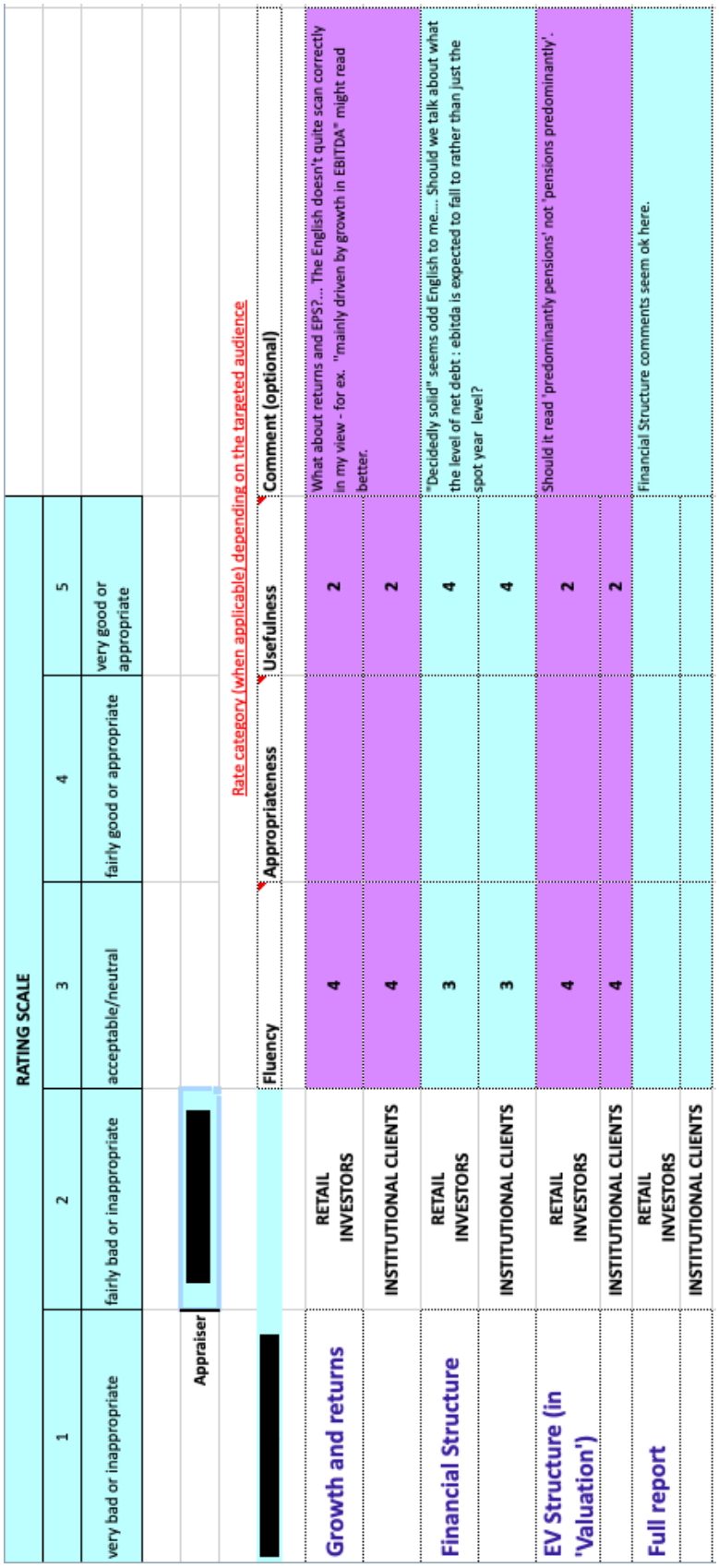
A.3 Classification financière

La table A.4 présente, à titre de comparaison avec la classification Exane présentée dans la figure 3.5, un extraction de la classification GICS ®[2] du secteur « Industrials ».

2. www.msci.com/gics.

Industrie	Secteur	Sous-secteur
Industrials	Capital Goods	Aerospace & Defense
		Building Products
		Construction & Engineering
		Electrical Equipment
		Industrial Conglomerates
		Machinery
		Trading Companies & Distributors
	Commercial & Professional Services	Commercial Services & Supplies
		Professional Services
	Transportation	Air Freight & Logistics
		Airlines
		Marine
		Road & Rail
		Transportation Infrastructure

TABLE A.4 – Classification GICS ® d'un secteur (Industrials) en sous-secteurs



RATING SCALE				
1	2	3	4	5
very bad or inappropriate	fairly bad or inappropriate	acceptable/neutral	fairly good or appropriate	very good or appropriate

Appraiser [redacted]

Rate category (when applicable) depending on the targeted audience

[redacted]		Fluency	Appropriateness	Usefulness	Comment (optional)
Growth and returns	RETAIL INVESTORS	4		2	What about returns and EPS?... The English doesn't quite scan correctly in my view - for ex. "mainly driven by growth in EBITDA" might read better.
	INSTITUTIONAL CLIENTS	4		2	
Financial Structure	RETAIL INVESTORS	3		4	"Decidedly solid" seems odd English to me.... Should we talk about what the level of net debt : ebitda is expected to fall to rather than just the spot year level?
	INSTITUTIONAL CLIENTS	3		4	
EV Structure (in 'Valuation')	RETAIL INVESTORS	4		2	Should it read 'predominantly pensions' not 'pensions predominantly'.
	INSTITUTIONAL CLIENTS	4		2	
Full report	RETAIL INVESTORS				Financial Structure comments seem ok here.
	INSTITUTIONAL CLIENTS				

FIGURE A.1 – Exemple de formulaire utilisé pour l'évaluation des textes (Excel)

ANNEXE B

EXTRAITS DE SCRIPTS

B.1 Scripts liés au corpus de référence

Le script de prétraitement du corpus de référence mentionné dans le chapitre 4 utilise les expressions régulières de la figure 4.5 pour prétraiter chaque ligne du fichier CSV et reconstituer les articles. Le script B.1 définit la fonction et la classe utilisées pour reconstituer les articles.

```
1  def get_text(text):
2      """
3      removes figures, disclosure agreements and other metrics
4      :param text: a list of lines
5      :return: a string
6      """
7      do_print = True
8      trimmed_text = []
9      for line in text:
10         if re_figure.search(line) or "Key␣valuation␣metrics" in line:
11             do_print = False
12         elif "Source:" in line or re_valuation.search(line):
13             do_print = True
14             if "Source:" in line:
15                 line = line.partition('.')[2]
16         if do_print:
17             trimmed_text.append(line.replace(';', ',').replace('Â?‘', '\'').replace('?‘',
                   '\'').replace('\0',''))
18     if 'DISCLOSURE␣APPENDIX' in trimmed_text:
19         return '␣'.join(trimmed_text[:trimmed_text.index('DISCLOSURE␣APPENDIX')])
20     else:
21         return '␣'.join(trimmed_text)
22
23 class Article:
24     def __init__(self, x):
25         self.id = x.pop(0).rstrip(';')
26         x[-1], sep, self.date = x[-1].rpartition(';')
27         self.xml = get_text(x)
28
29 class Corpus:
30     first_line = ['DIID','DIXML','DIDATE']
31     def __init__(self):
32         self.arts = []
33     def add(self, article):
34         self.arts.append(article)
35     def create_corpus(self, corpus):
36         for a in split_corpus(corpus):
37             article = Article(a)
38             self.add(article)
39     def print_corpus(self, writer):
40         writer.writerow(articles.first_line)
41         for a in self.arts:
42             writer.writerow([a.id, a.xml, a.date])
```

FIGURE B.1 – Extrait du script de prétraitement (Python)

B.2 Scripts liés aux modules de génération

```java
package com.nlptools.corenlp;
import java.io.IOException;
import java.io.PrintWriter;
import java.util.List;
import java.util.Properties;
import edu.stanford.nlp.io.IOUtils;
import edu.stanford.nlp.ling.CoreAnnotations;
import edu.stanford.nlp.ling.CoreLabel;
import edu.stanford.nlp.pipeline.Annotation;
import edu.stanford.nlp.pipeline.StanfordCoreNLP;
import edu.stanford.nlp.util.CoreMap;

public class TokensRegexAnnotator {

        public static void main(String[] args) throws IOException {
                PrintWriter out;
                String rules;
                if (args.length > 0) {
                        rules = args[0];
                } else {
                        rules = "edu/stanford/nlp/ling/tokensregex/demo/rules/expr.rules.txt";
                }
                if (args.length > 2) {
                        out = new PrintWriter(args[2]);
                } else {
                        out = new PrintWriter(System.out);
                }

                Properties properties = new Properties();
                properties.setProperty("annotators", "tokenize,ssplit,pos,lemma,tokensregex");
                properties.setProperty("customAnnotatorClass.tokensregex",
                    "edu.stanford.nlp.pipeline.TokensRegexAnnotator");
                properties.setProperty("tokensregex.rules", rules);
                StanfordCoreNLP pipeline = new StanfordCoreNLP(properties);
                Annotation annotation;

                if (args.length > 1) {
                        annotation = new Annotation(IOUtils.slurpFileNoExceptions(args[1]));
                } else {
                        annotation = new Annotation("oops");
                }

                pipeline.annotate(annotation);
                out.println("VERBE;POURCENT");
                out.println();

                List<CoreMap> sentences = annotation.get(CoreAnnotations.SentencesAnnotation.class);
                for (CoreMap sentence : sentences) {
                        for (CoreLabel token : sentence.get(CoreAnnotations.TokensAnnotation.class))
                            {
                               String ne =
                                    token.get(CoreAnnotations.NamedEntityTagAnnotation.class);
                               if (ne != null) {
                                 if (ne.equals("VERBE")) {
                                   String lemma = token.get(CoreAnnotations.LemmaAnnotation.class);
                                   out.print(lemma + " ");
                                 }
                                 else if (ne.contentEquals("POURCENT")) {
                                         String lemma =
                                              token.get(CoreAnnotations.LemmaAnnotation.class);
                              out.println(";" +lemma);
                                 }
                               }
                        }
                }
                out.flush();
        }
}
```

FIGURE B.2 – Extrait du script d'extraction de motifs (Java)

Pour effectuer l'expérience décrite dans la section 5.4.2, nous avons utilisé le script B.2 sur le corpus post-traitement afin de générer un fichier CSV. Ce fichier CSV a ensuité exploité en Python pour générer les diagrammes avec

`matplotlib` [Hunter, 2007].

B.3 Scripts liés à l'évaluation

Nous avons extrait les résultats de formulaires Excel avec Python (script B.3), puis nous les avons exploités, également en Python (script B.4).

```
1  import xlrd
2  import pandas
3  import glob
4  import pickle
5  import matplotlib
6
7  def extract_form(filepath):
8      """
9      extract form from excel file
10     :param filepath: string
11     :return:
12     """
13     global evals, companies, type_ratings
14     appraiser, sep, stock = filepath.partition('/')
15     for comp in companies:
16         if comp in stock:
17             stock_name = companies[comp]
18
19     wb = xlrd.open_workbook(filepath)
20     eval_sheet = wb.sheet_by_index(2) # extract evaluation sheet
21
22     for i in range(8,15):
23         row = eval_sheet.row_slice(i, start_colx=0, end_colx=9)
24         if row[0].value == '':
25             section = eval_sheet.cell(i-1,0).value
26         else:
27             section = row[0].value
28         type_investor = row[1].value
29
30         for type_rating, rating in zip(type_ratings,row[2:]):
31             evals = evals.append([{'appraiser':
                   appraiser,'stock_name':stock_name,'section':section,'type_investor':type_investor,
32             'type_rating':type_rating,'rating':rating.value}], ignore_index=True)
33
34 def find_comments(liste):
35     """
36     sort free comments
37     :param liste: one column of the dataframe
38     :return: list
39     """
40     comments = []
41     for rating in liste:
42         if isinstance(rating, str) and len(rating) > 3:
43             comments.append(rating)
44     return comments
45
46 companies = { # dict contains a list of the evaluated companies mapped to their abbreviated name}
47 type_ratings = ['fluency','appropriateness','usefulness','free_comment']
48 evals = pandas.DataFrame(columns=['appraiser','stock_name',
       'section','type_investor','type_rating','rating'])
49
50 for file in glob.glob('*/*.xls'):
51     print(file)
52     extract_form(file)
53
54 assert evals.shape == (560,6) # 20 evaluation sheets with 28 lines each
55 pickle.dump(evals,open('evals.pkl','wb'))
```

FIGURE B.3 – Extrait du script d'extraction des résultats d'évaluation (Python)

```python
import pandas

evals = pandas.read_pickle('evals.pkl')

def resultats(column, name, print_median=True):
    df = evals.loc[evals[column] == name]
    fluency = df.loc[df['type_rating'] == 'fluency']
    accuracy = df.loc[df['type_rating'] == 'appropriateness']
    utility = df.loc[df['type_rating'] == 'usefulness']

    print(name)

    if print_median:
        print("F-mean {}\nF-median {}".format(pandas.to_numeric(fluency['rating']).mean(),
        pandas.to_numeric(fluency['rating']).median()))
        print("A-mean {}\nA-median {}".format(pandas.to_numeric(accuracy['rating']).mean(),
        pandas.to_numeric(accuracy['rating']).median()))
        print("U-mean {}\nU-median {}".format(pandas.to_numeric(utility['rating']).mean(),
        pandas.to_numeric(utility['rating']).median()))
    else:
        print("F-mean {}".format(pandas.to_numeric(fluency['rating']).mean()))
        print("A-mean {}".format(pandas.to_numeric(accuracy['rating']).mean()))
        print("U-mean {}".format(pandas.to_numeric(utility['rating']).mean()))

companies = ['COMP1', 'COMP2','COMP3','COMP4','COMP5','COMP6','COMP7','COMP8','COMP9','COMP10']
appraisers = ['EN1_AN1','EN2_AN2','EN3_SA1''FR1_SA2']
sections = ['Growth and returns','Financial Structure','EV Structure (in \'Valuation\')']

for company in companies:
    resultats('stock_name', company, print_median=False)

for appraiser in appraisers:
    resultats('appraiser', appraiser, print_median=False)

for section in sections:
    resultats('section', section)
```

FIGURE B.4 – Extrait du script de traitement des résultats (Python)

INDEX