Mémoire de Master 1

Master Linguistique, Informatique et Technologies du Langage

Classification et analyse de l'expression des problèmes techniques

Shaad CASSAM SULLIMAN

Sous la direction de Ludovic TANGUY et Mariame MAAROUF

Juin 2023

Remerciements

Ce travail n'aurait pas pu aboutir sans l'aide précieuse de plusieurs personnes ayant œuvré à plusieurs niveaux dans sa réalisation. J'adresse mes remerciements à ces personnes qui m'ont aidé à réaliser le travail de longue haleine décrit dans ce mémoire.

Tout d'abord, je remercie mes directeurs de mémoire, Ludovic TANGUY et Mariame MAAROUF, qui ont su me guider et répondre à mes questions tout au long de ce travail. Leurs retours précis sur les améliorations possibles à plusieurs niveaux ont permis au projet de prendre sa forme actuelle. Je les remercie également pour leurs travaux précédents et actuels sur lesquels j'ai pu m'appuyer dans la réalisation de cette étude.

Je remercie Cécile FABRE et Lydia Mai HO-DAC pour leurs enseignements tout au long de l'année, qui se sont révélés vitaux à plusieurs étapes de la réalisation de ce travail.

Je remercie Damien RAVÉ, créateur et gérant du site CommentReparer.com, pour nous avoir autorisé à utiliser les données de son site dans notre étude, et notamment dans la création du corpus CoCoRep (Corpus Comment Reparer). Ces données étant un pré-requis pour l'étude menée, celle-ci n'aurait pas pu aboutir sans son autorisation.

Je remercie également Sophie, qui a accepté de jouer le rôle de l'annotateur externe et s'est appliquée pour nous fournir une annotation de qualité sur deux échantillons de titres. L'annotation de ces deux échantillons étant au cœur de la réussite du projet de recherche, les efforts de Sophie ont également contribué à l'aboutissement du travail.

Je tiens également à remercier Solyne, qui m'a accompagné toute l'année et sans qui je n'aurais pas trouvé la force d'aller jusqu'au bout de cette année et de ce travail.

Enfin, je remercie Masha, Angélo, Vladana, et le reste de la promo de M1 LITL, pour m'avoir épaulé et contribué à un environnement de travail rempli d'entraide et de bonne humeur.

Table des matières

Introduction

L'expression des problèmes peut se trouver sous différentes formes et correspondre à différentes utilités. Un problème peut être communiqué de façon libre ou structurée, et sa finalité est souvent, au travers de différentes méthodes, la résolution du problème. Il existe un grand nombre de façons de résoudre les problèmes que nous rencontrons dans notre vie, et la résolution de problèmes peut être assistée par d'autres personnes, ou plus récemment, par des machines, si nous sommes capable de faire comprendre à cet autre agent le problème que nous avons en l'exprimant d'une manière ou d'une autre. Cela place donc l'expression du problème dans une position centrale quant à la résolution du problème en question.

Les problèmes peuvent être de natures diverses et leur résolution aura donc des implications différentes selon le type de problème et le domaine dans lequel celui-ci apparaît. Dans l'industrie, et dans le cadre de l'utilisation d'objets ou de machines, des problèmes dits « techniques » apparaissent et leur résolution peut être cruciale car le problème peut causer des dégâts humains, matériels, financiers, ou encore écologiques. La résolution de ces problèmes, exprimés grâce à la langue, peut être aidée de systèmes informatiques qui facilitent l'analyse, et donc la résolution des problèmes.

Sachant que les expressions de ces problèmes techniques désignent des états relativement similaires, c'est-à-dire des « problèmes techniques », nous pouvons supposer retrouver des similarités dans ces expressions quel que soit le problème exprimé. Cela pourrait nous permettre de dégager une typologie de l'expression des problèmes qui, par la suite, serait intéressante à utiliser pour l'analyse des problèmes tels que décrits par des utilisateurs du langage naturel.

Le projet de recherche dans lequel nous nous situons s'intitule « Classification et analyse de l'expression des problèmes techniques ». Il s'agit d'un travail dont le but est d'aboutir, en premier lieu, à un système de classification automatique des expressions de problème technique dans un corpus de fiches contenant ce type d'expressions. Ce travail s'inscrit dans un contexte de recherche accompagné par Ludovic TANGUY et Mariame MAAROUF, doctorante qui effectue une thèse sur le repérage, le typage, ainsi que la modélisation des problèmes techniques dans les fiches d'anomalies du domaine spatial. Le concept de modélisation passe par la méthodologie TRIZ développée par l'ingénieur soviétique Altshuller dès 1946, qui consiste en la représentation des problèmes en vépoles. Le « vépole », dans l'approche TRIZ, désigne une représentation d'objets concrets (ou « substances ») par des correspondants abstraits (« champs ») afin de trouver des solutions (Ilevbare et al., 2013).

La base de données sur laquelle porte ce travail de thèse est une base de données appartenant au Centre National d'Études Spatiales qui contient des fiches d'anomalies relatant des problèmes en lien avec les lanceurs spatiaux Ariane 5. Ces fiches d'anomalies font état d'écarts à la norme constatés sur les lanceurs. Cependant, cette base de données est hautement confidentielle et n'est donc pas disponible pour la présentation de travaux d'analyse.

Dans ce contexte d'utilisation limitée de ces données, la première partie de notre travail consiste en l'élaboration d'un corpus similaire, contenant l'équivalent de fiches d'anomalie qui relatent un problème quelconque. Pour ce faire, nous avons choisi de construire un corpus dont les données sont extraites d'un forum internet sur lequel les utilisateurs peuvent venir présenter un problème qu'ils ont eu avec un appareil ou un objet, et d'autres utilisateurs peuvent leur répondre et les aider à résoudre le problème. Ces données ressemblant dans une certaine mesure aux fiches d'anomalie confidentielles que nous avons évoquées, elles pourront entre autres servir à la deuxième partie de notre travail, qui consistera à repérer des structures et des schémas récurrents d'expressions de problèmes dans le corpus, pour ensuite proposer un algorithme de classification automatique des fiches selon les types d'expression de problèmes que nous aurons dégagés. Néanmoins, les différences de forme et de contenu des données, dans la visée de la rédaction des fiches par exemple, ainsi que le

média sur lequel les fiches sont rédigées, nous amèneront à adapter notre analyse à notre corpus en prenant en compte ses spécificités.

Au travers de ce projet de recherche, nous souhaitons observer les similarités entre certains types d'expression de problème technique, dégager les patrons syntaxiques et les termes utilisés parmi ces types d'expression similaires, et savoir si de telles similarités peuvent amener à la création d'un système de règles linguistiques pour catégoriser ces différents types d'expression. Aussi, nous nous penchons sur la problématique suivante : « Comment peut-on catégoriser de façon automatique les expressions d'un problème technique ? »

Dans ce mémoire, nous proposons tout d'abord un état de l'art faisant le point sur différents types d'expression de problèmes, ainsi que sur les méthodes développées et utilisées afin de rechercher et d'extraire les informations dans ces différents types de données contenant l'expression d'un problème. Nous ferons également un point sur les spécificités du langage dans le cadre de la communication médiée par les réseaux, et de la création d'un corpus contenant des données de ce type. Dans un deuxième temps, nous nous pencherons sur les données récoltées et la constitution du corpus sur lequel nous travaillerons, avant de présenter les différentes observations linguistiques et statistiques que nous avons pu effectuer sur ces données. Dans une troisième partie, nous détaillerons les catégories retenues pour la classification des expressions de problèmes, et présenterons les différentes étapes de la création du guide d'annotation qui est à la base de la catégorisation automatique. Puis, dans une quatrième partie, nous nous concentrerons sur la méthodologie utilisée pour la catégorisation automatique. Enfin, nous évaluerons les résultats de notre système de classification et discuterons de ces résultats en nous interrogeant sur les suites que peuvent donner ce travail de recherche.

1. La catégorisation des problèmes techniques et ses difficultés

Dans cette première partie, nous tâchons de dresser un état de l'art autour des formes d'expression de problèmes et plus spécifiquement des problèmes techniques. L'expression du problème apparaissant dans différents contextes pour différentes raisons et ayant des visées différentes, nous nous penchons, en lien avec le cadre de l'étude, sur deux contextes d'expression du problème, à savoir le contexte industriel avec les Fiches d'Anomalie, et l'expression sur les réseaux, dans un contexte plus général de communication médiée par les réseaux. Nous en profitons également pour aborder plusieurs problématiques fondamentales dans l'étude linguistique de ce type de données.

1.1. Les différentes formes d'expression de problèmes techniques et leur analyse

1.1.1. Les retours d'expérience et les fiches d'anomalie

Nous avons rapidement abordé la notion de fiches d'anomalies dans notre introduction. Dans cette partie, nous nous pencherons en détail sur ce que celles-ci représentent et les informations qu'elles contiennent.

Afin d'expliquer la notion de fiches d'anomalies, nous devons d'abord nous concentrer sur la notion de Retour d'Expérience. Les Retours d'Expérience (REX ou RETEX) peuvent être définis comme « un processus de signalisation, de stockage et d'analyse des événements (écarts, défaillances, incidents) dans une entreprise », mais également comme l'objet textuel qui contient la description de ces évènements (Blatter & Raynal, 2015). Il s'agit donc d'un processus durant lequel tout écart à la norme constaté par un technicien, un opérateur, ou toute autre personne impliquée, est relaté sous forme textuelle dans des fiches stockées dans une base de données interne à l'entreprise ou commune à un domaine particulier dans le but d'être ensuite analysées.

Les écarts à la norme que l'on qualifie de « problèmes techniques » sont des écarts à la norme autour ce qui concerne l'état d'un objet fabriqué de façon artificielle ou d'un appareil, son fonctionnement, son apparence, ou encore les problématiques liées au fonctionnement de plusieurs de ces objets ou appareils dans un système. Cet écart à la norme, pour constituer un « problème technique », est un écart négatif, ce qui signifie que la situation n'est pas satisfaisante et appelle à une résolution (ce que nous rappelle le terme « problème »). De cette façon, un écart à la norme ne peut pas être considéré comme un « problème technique » s'il porte par exemple sur un objet naturel. Par exemple, un arbre planté dans un parc qui s'est abattu suite à une tempête peut constituer un écart à la norme, mais il ne s'agit pas d'un problème technique, car l'arbre détérioré ou ayant changé d'apparence est un objet naturel. En revanche, un banc cassé par le passage d'une tempête dans un parc peut constituer un cas de problème technique. Cette définition nous permet d'inclure des cas typiques tels que le dysfonctionnement d'une machine, une fuite, un blocage, mais également une détérioration au niveau d'une couture d'un vêtement ou la finition d'un jouet.

Le stockage dans les bases de données de ces écarts à la norme se fait au moyen de ce qu'on peut appeler des Fiches d'Anomalie (FA). Ces fiches sont généralement constituées d'une description brève de l'écart à la norme constaté, ainsi que d'une description narrative plus détaillée, donnant accès à des informations supplémentaires sur la façon dont un écart à la norme s'est déroulé, ou encore l'état de l'environnement autour de l'endroit où a été constaté l'écart. Malgré un effort de normalisation des fiches d'anomalies au moyen de directives plus ou moins strictes en termes de structuration des fiches et des informations présentes sur ces fiches, la majorité des problèmes relatés dans ces fiches le sont

sous la forme de texte libre, changeant en fonction de l'auteur, de son style d'écriture, du temps dont il dispose ou encore de son niveau d'expertise. L'équilibre recherché en matière de normalisation sur la forme ou la structure est néanmoins un équilibre difficile à atteindre, car il est important surtout que la rédaction se fasse en premier lieu, et l'opérateur pourrait choisir de ne pas rédiger si les règles se révèlent trop contraignantes ou trop difficiles à suivre.

Le but de ces retours d'expérience est d'avoir une trace écrite des écarts à la norme constatés, afin de pouvoir par la suite effectuer des analyses de ces traces écrites pour notamment analyser les écarts négatifs, afin de repérer des tendances négatives dans le fonctionnement des appareils ou dans les services proposés. L'analyse de ces fiches d'anomalies peut ensuite permettre de déceler des problèmes potentiellement récurrents, ainsi que repérer les causes de ces problèmes récurrents, et permettre, en fonction du temps et des moyens alloués, de mettre en place des actions afin de viser ces causes et les éliminer pour corriger les problèmes causés (Gaillard, 2005).

Les problèmes ainsi relatés dans les fiches d'anomalie sont faits sous la forme d'expression de problèmes techniques, qui peuvent varier selon plusieurs facteurs, notamment le type de problème ou la personne qui s'exprime sur le problème au travers de ces fiches. Le travail parallèle de Mariame MAAROUF sur les fiches d'anomalie du domaine spatial a amené à la mise en place d'une typologie de ces expressions de problèmes techniques dans le but de repérer et de modéliser les problèmes dans les fiches d'anomalie. Cette typologie comprend 12 catégories plus où moins précises, et c'est ce degré de précision variable qui est à la base de cette typologie. Chaque description de problème dans une FA se voit attribuer une catégorie d'expression du problème grâce au repérage de marqueurs associés à des catégories. La catégorie correspondant au marqueur le plus précis parmi les marqueurs présents dans la description du problème sera attribuée à cette description de problème. La liste des catégories d'expression du problème dans l'ordre de précision peut être retrouvée dans le tableau ci-dessous. Nous proposons ensuite une présentation de chaque catégorie de cette typologie, de la même façon que ces catégories sont présentées dans le guide d'annotation utilisé pour l'annotation des données de FA du domaine spatial.

Catégorie	Etiquette
Fuite	Fuite
Signal qui s'est déclenché	Signal
Obstacle	Obstacle
Dégradation - Usure - Saleté	Dégradation
Objet manquant (dispositif, document, outil)	Absence
Hors spécification	HorsSpec
Dispositif qui ne fonctionne pas	FonctionnePas
Action difficile ou impossible	Impossible
Etat du monde	EtatDuMonde

Tableau 1 - Catégories d'expressions du problème du domaine spatial dans l'ordre de spécificité et étiquettes correspondantes

a) Fuite (Fuite) :

Présence d'une fuite, d'un écoulement d'une substance. Cette tâche ne consiste pas à rechercher la cause de la FA. Ainsi, la présence de tâches d'un liquide ou de flaques, sans marqueur qui est propre à la fuite, n'est pas considéré comme appartenant à cette catégorie. Par exemple, une FA telle que « Présence de tâches d'huile sur le sol » n'est pas du type *Fuite*. En revanche, « Présence

d'une fuite d'huile sur le sol » relève bien du type Fuite. Les marqueurs typiques de cette catégorie sont : « fuite », « fuyarde », « écoulement »...

b) Signal qui s'est déclenché (Signal) :

Rapport d'une alerte, témoin, signal, qui s'est déclenché et qui indique la présence d'un problème. La FA ne décrit pas quel est le problème, uniquement qu'un indicateur d'un problème s'est déclenché. Cela peut s'agir d'un témoin qui s'allume ou bien d'un message d'erreur qui apparaît. Nous retrouverons comme marqueurs : « alerte », « témoin », « alarme »...

c) Obstacle (Obstacle) :

Une action n'est pas réalisable dû à la présence d'un obstacle. Il y a un élément qui en bloque un autre. Les marqueurs sont : « bloque », « gêné », « empêche »...

d) Dégradation - Saleté - Usure (Dégradation) :

Constatation de dégradation, de saleté ou d'usure. La FA relève la présence d'une détérioration d'un équipement de quelque nature. Cela recouvre tous les problèmes où quelque chose est « cassé », « déchiré », mais aussi la présence de « poussière », d'« insectes » ou de « corrosion ».

e) Objet manquant - Dispositif, document, outil (Absence) :

Objet, dispositif, document manquant. Les marqueurs seront de l'ordre de : « sans », « absence », « perte »...

f) Hors spécification (HorsSpec) :

Ce type comprend toute évocation d'un état qui est décrit spécifiquement comme ne correspondant pas à l'état attendu, mais dont le type d'écart n'est pas spécifié, c'est-à-dire que nous n'avons pas de précision sur si l'objet est dégradé ou manquant. En revanche, on nous fait part d'une situation qui est décrite comme incorrecte. Les marqueurs peuvent être très variés : « hors famille », « au lieu de », « attendu... mesuré... » (dans ce cas-ci, les deux annotations doivent être reliées en chaîne), « mal »...

g) Dispositif qui ne fonctionne pas (FonctionnePas) :

Dispositif, objet qui ne fonctionne pas parce qu'il est en panne ou défectueux. C'est uniquement un constat du problème qui est exposé ici, pas d'information supplémentaire sur la raison de la panne. Les marqueurs sont souvent : « HS » (Hors Service), « défectueux », « ne fonctionne pas »...

h) Action difficile ou impossible (Impossible) :

Ce type survient lorsque la FA fait état d'une action impossible ou difficile à effectuer. La description du problème ne comprend aucun diagnostic, aucune information sur la raison pour laquelle l'action est impossible. La proposition est un simple constat de l'impossibilité d'accomplir la tâche. Les différents marqueurs peuvent être de l'ordre de : « impossibilité », « dur », « difficulté »...

i) Etat du monde (EtatDuMonde) :

Présence de quelque chose qui ne devrait pas être présent, ou pas dans cette position. Ce type est le seul pour lequel il n'y a pas de marqueur spécifique au type du problème obligatoirement présent. Dans ces cas, la notion de problème n'est pas portée par une unité lexicale ; elle est implicite et existe du fait que le corpus soit des rapports d'incidents.

Une différenciation est à faire entre les segments de contexte qui ne comportent pas de marqueurs et ceux qui sont des « États du monde ». Pour cela, le critère déterminant est la présence d'un autre segment comportant un déclencheur dans la FA. Si un déclencheur est présent, on considère que le segment sans marqueur est un élément de contexte. Si aucun déclencheur n'est présent dans toute la FA, celle-ci est un « État du monde ». Par exemple, l'énoncé « La porte est ouverte » est considéré comme un « État du monde ». En revanche, « La porte est ouverte. Le loquet est cassé » est

un cas de type « Dégradation » et seul le marqueur « cassé » est à annoter dans la FA. Une étiquette « État du monde » est à apposer sur le premier token de la FA. Si la FA comporte plusieurs phrases, seule une étiquette est à mettre, il est inutile de mettre plusieurs étiquettes « État du monde ». À noter : les cas où la description ne fait pas état d'un problème mais du risque d'apparition d'un problème sont à considérer comme de type « État du monde ».

1.1.2. L'expression orale des problèmes techniques

Malgré le caractère pratique des retours d'expérience dans le but d'exprimer un problème et de pouvoir l'analyser, la majorité des problèmes techniques dans le monde ne s'expriment pas d'une façon aussi codifiée que celle du retour d'expérience, même si celui-ci reste tout de même relativement libre. Dans tous les domaines et pour tous les types de problèmes, mais également dans le cas de problèmes techniques dans le cadre des industries, il y a énormément de façons différentes d'exprimer un problème rencontré ou constaté. Parfois, l'expression d'un problème sous une autre forme que le REX peut même être préférable, lorsque la résolution d'un problème critique est soumise à une contrainte de temps par exemple, ou que le problème est si minime qu'il serait plus embêtant d'utiliser des ressources supplémentaires pour analyser le problème alors que sa résolution immédiate coûterait moins de temps et moins d'argent à l'organisme.

Pour ces raisons, il arrive que les problèmes rencontrés soient exprimés à l'oral, de façon non contrainte ou semi-contrainte par une norme permettant une communication rapide et fluide grâce à l'utilisation de jargon ou d'acronymes propres au domaine en question. Malgré l'absence éventuelle d'une norme d'expression lors de l'expression orale d'un problème, il est intéressant de constater que nous retrouvons tout de même des similarités entre les différentes situations d'expression de problèmes à l'oral (Condamines & Vergely, 2001).

Dans un contexte d'entreprise ou dans un contexte industriel où plusieurs parties doivent se coordonner de façon efficace afin de mener un travail à bien et garantir la sécurité des équipes présentes sur le terrain, il n'est pas rare de retrouver des expressions de problèmes dans les dialogues entre ces différentes parties, qui doivent communiquer souvent rapidement afin de trouver des solutions et permettre à la chaîne de travail de continuer son fonctionnement de façon fluide. Dans ce contexte, une étude a été effectuée afin de rendre compte des potentielles régularités dans l'expression orale des problèmes dans le domaine de l'aviation (Condamines & Vergely, 2001). La visée de cette étude est de répertorier les différentes façons d'exprimer un problème à l'oral dans ce milieu et de construire une grammaire basée sur les régularités d'expressions présentes dans les dialogues exprimant un problème entre les chefs de salle et les superviseurs de contrôle aérien, dont les missions sont entre autres de garantir la sécurité et la fluidité de l'activité aérienne et du contrôle de cette activité au moyen de résolution rapide des problèmes rencontrés. Pour ce faire, l'expression de ces problèmes contient un certain nombre de mots appartenant au jargon aérien permettant de véhiculer une information précise dans un temps moindre. Les similarités observées sont également d'ordre syntaxique, car cette étude révèle un certain nombre de patrons d'expression de problèmes auquel se conforme de façon inconsciente une grande partie de l'expression des problèmes.

En dehors de la communication interne à l'entreprise visant à une fluidité de l'activité, il existe d'autres contextes dans lesquels l'expression d'un problème peut se faire de façon orale. Nous pouvons mentionner les services de téléassistance, qui consistent à aider des particuliers ou des professionnels demandant de l'aide par téléphone à effectuer des tâches d'une grande diversité. Ces tâches peuvent aller de la réparation d'un appareil ménager ou électronique, comme un ordinateur ou un lave-vaisselle. Là où certains de ces problèmes pourraient requérir une intervention d'un ingénieur qualifié, d'autres peuvent être résolus par le locuteur directement en suivant les instructions et les conseils proposés par l'interlocuteur assistant.

1.1.3. Techniques d'analyse de similarité et recherche de signaux faibles

La récolte d'expression de problèmes au travers des REX ou sous d'autres formats variés ne constitue pas un but en soi. En effet, le stockage des informations ne sert que si ces informations sont ensuite utilisées à des fins d'étude et d'analyse afin de détecter des tendances et schémas répétés, ainsi que des problèmes à résoudre.

A cet effet, Ansoff (1975) propose l'hypothèse selon laquelle il existe dans ces textes ce qu'il appelle des « signaux faibles » (ou *weak signals* en anglais). Ces signaux faibles seraient des informations présentes dans un texte qui alertent de façon précoce d'une potentielle tendance ou d'un problème non repéré ou non référencé de façon explicite. En effet, les signaux faibles sont, par définition, « faibles », c'est-à-dire qu'ils sont difficilement reconnaissables et analysables, même par des analystes experts, car ils apparaissent dans les textes de façon implicite, cachés sous des informations de surface, ou fragmentés entre d'autres informations qui peuvent « brouiller les pistes ».

Il est pourtant très important pour les entreprises et les acteurs de l'industrie car, s'ils sont analysés, ils peuvent fournir des informations extrêmement précieuses sur le fonctionnement ou les dysfonctionnements de leur chaîne de travail, et peuvent permettre aux entreprises d'affiner leurs stratégies ou renforcer certains aspects de leur activité, tant sur le plan évident de la sécurité des employés ou de la population générale, que sur le plan financier pour assurer la pérennité de l'entreprise.

Cependant, pour effectuer des analyses pertinentes sur les fiches d'anomalies présentant des rapports d'expérience afin d'en extraire les signaux faibles éventuels et en tirer des conclusions sur la direction à suivre, des techniques avancées de traitement du langage sont nécessaires. De telles techniques sont très importantes dans un contexte de grandissement rapide du nombre de fiches d'anomalie dans tous les domaines, contexte dans lequel il devient de plus en plus pertinent de trouver de façon d'analyser et d'extraire les signaux faibles de ces fiches de façon entièrement automatisée.

Au-delà de la recherche de signaux faibles, l'analyse de REX et des informations contenues dans ceux-ci passent par une recherche de similarités entre les fiches d'anomalies contenant des informations textuelles qu'il est important d'extraire et de regrouper afin de dégager des tendances claires. Un système conçu pour l'analyse des REX devrait regrouper les fiches similaires ensemble, ce qui permettrait, en analysant les similarités des fiches dans chaque groupement, de repérer et d'analyser les tendances présentes dans les textes. Dans cette optique, plusieurs outils ont été développés dans le domaine du Traitement Automatique des Langues (TAL). Galand et al. (2018) en font un inventaire dans leur article concernant la recherche de signaux faibles dans les REX du domaine spatial.

Afin de réaliser une telle analyse, des techniques d'apprentissage automatique sont de plus en plus utilisées dans le domaine du TAL. Les techniques d'apprentissage automatique permettent à un système informatique, ou « modèle », d'apprendre de ses données sans passer par un système de règles créé par un expert du domaine. Il existe deux types d'apprentissage automatique, qui sont l'apprentissage non supervisé et l'apprentissage supervisé. Lors d'un apprentissage non supervisé, un modèle essaie d'inférer tout seul des similarités selon différentes méthodes, mais ne s'appuie pas sur un output préalable pour calculer la probabilité que tel output se retrouve en sortie lorsque tel input est fourni. Il existe également des techniques d'apprentissage supervisé qui fonctionnent quelque peu différemment. Dans l'apprentissage supervisé, le modèle a à sa disposition des données et la sortie attendue qui serait celle produite par un potentiel modèle parfait. Après un entraînement sur une partie des données, le modèle tente de deviner les valeurs en sortie associées aux valeurs en entrée. Puisque les valeurs attendues en sortie sont connues dans ce type d'apprentissage, il est possible d'évaluer le système en calculant son rappel, mesure du nombre de bonnes prédictions par rapport au nombre de bonnes réponses attendues, ainsi que sa précision, mesure du nombre de bonnes prédictions pour une valeur de sortie par rapport au nombre total de prédictions pour cette même valeur.

La réduction de dimensionnalité est une des approches envisagée dans la catégorisation automatiques des fiches d'anomalie quand il s'agit du traitement du texte. Le concept de réduction de dimensionnalité consiste à regrouper les termes similaires présents dans les textes en une seule entité permettant d'effectuer un calcul de similarité des fiches plus exact. La réduction de dimensionnalité permet également un stockage de données moins coûteux sur le plan de l'espace de stockage.

La réduction de dimensionnalité peut s'effectuer sur un texte après vectorisation. En réalité, il est beaucoup plus facile pour un système informatique de comprendre un texte lorsque celui-ci est représenté sous la forme de nombres. Un des premiers traitements communs à toute analyse de similarité entre des textes consiste en la conversion du document textuel en vecteur qui donnera ensuite des informations sur sa ou ses valeurs sémantiques. Pour ce faire, il convient de passer par plusieurs phases de traitement. D'abord les mots sont découpés (on parle de « tokenisation ») et le document est représenté sous la forme d'un sac de mots. Ensuite, chaque mot peut se voir attribuer des informations supplémentaires qui permettront un meilleur calcul de similarité, comme son lemme (il s'agit de la « lemmatisation »). Enfin, chaque mot présent dans la base de données entière constitue une dimension sur laquelle nous allons, pour chaque document présent dans la base de données, attribuer un nombre correspondant à la fréquence de ce terme dans le document. La liste de nombres représentant la fréquence de chaque mot dans chaque document constitue une matrice de vecteurs, chacun représentant un document sous forme mathématique. La fréquence de chaque mot est présente sous la forme du TF-IDF (Term Frequency - Inverse Document Frequency), qui correspond à la fréquence du terme dans un document, comparé à sa fréquence dans la base de données entière.

A la suite de cette transformation et préparation des données, il est possible d'utiliser certaines techniques mathématiques et statistiques pour tirer des conclusions sur la nature des textes. La réduction de dimensionnalité consiste donc à réduire les dimensions de la matrice pour faire émerger des tendances lexicales sous-jacentes. Une des techniques utilisables à cet effet s'intitule LSA (Latent Semantic Analysis), qui consiste à regrouper les dimensions similaires sur le plan sémantique (mots synonymes ou différentes façons d'exprimer le même concept). De cette façon, les dimensions restantes présenteront des thèmes ou des hyperonymes équivalents au regroupement de plusieurs dimensions représentées par des termes synonymes ou similaires. Cela rend le calcul de similarité plus pertinent entre les vecteurs restants.

Le LDA (Latent Dirichlet Allocation) est une autre technique de réduction de dimensionnalité, qui ne regroupe pas les termes similaires, mais crée une nouvelle matrice dans laquelle chaque dimension est un « topic » plutôt qu'un terme présent dans le texte. On parle alors de « topic modeling ». Chaque document sera donc représenté par un vecteur pour lequel chaque dimension correspond à un topic ayant une certaine probabilité d'être retrouvé dans chaque document.

D'autres techniques d'apprentissage automatique existent pour regrouper des documents, comme le clustering en k-moyennes, qui permet de regrouper les documents selon un nombre de catégories prédéfini. Le nombre de catégories, si pertinent, permet de voir apparaître des similarités décelées par le système informatique qui n'auraient pas forcément été repérées par un humain.

Ces techniques d'apprentissage non supervisé sont intéressantes lorsqu'il n'y a pas d'information de sortie à laquelle doit être couplée l'information en entrée du modèle. Lorsqu'un jeu de données comprend une information sur ce qui doit se trouver en sortie, il est plus intéressant de faire appel à des techniques d'apprentissage supervisé. Le SVM (Support Vector Machine) est une technique d'apprentissage supervisé qui a prouvé son efficacité dans la catégorisation automatique des fiches rapportant des problèmes (Tanguy et al., 2016). Ce modèle tente de « comprendre » le lien entre les données en entrée et en sortie pour construire une représentation interne de la façon dont sont catégorisés les documents, pour pouvoir ensuite catégoriser de la meilleure façon possible d'autres documents proposés en entrée au modèle en fonction des similarités trouvées avec les documents analysés par le modèle lors de la phase d'apprentissage.

Néanmoins, au-delà des techniques d'apprentissage automatique, il existe également des approches basées sur la linguistique et l'extraction d'informations dans les textes sur la base de règles linguistiques et de modèles syntaxiques ou d'informations sémantiques sur les termes contenus dans les textes afin d'extraire des termes pertinents dans les textes et d'effectuer une classification (Blatter & Raynal, 2015). Il peut s'agir de techniques se basant sur des informations existantes, comme des ressources langagières (on peut citer par exemple les terminologies qui contiennent des informations sur des termes utilisés dans un certain domaine). Dans une mesure plus large, il est également possible de se baser sur des patrons syntaxiques afin de rapprocher non pas des termes, mais des types d'expression qui peuvent également être porteuses de sens dans un contexte de REX et de rédaction plus ou moins libre de textes relatant des problèmes.

1.1.4. Utilisations dans le monde industriel

L'utilisation de techniques de TAL pour la recherche de signaux faibles dans les REX constitue déjà une pratique courante dans le monde industriel, notamment dans les secteurs d'activité où la sécurité est une composante vitale au bon déroulement des activités de l'entreprise et au bien-être des utilisateurs des services proposés par les entreprises. Tanguy et al. (2016) donnent l'exemple de l'analyse des REX dans le domaine de l'aviation, comme le faisaient Condamines et Vergely (2001) sur le plan oral, car il s'agit d'un domaine au sein duquel doivent s'échanger beaucoup d'informations de nature technique, et les activités dans ce domaine, comme dans le domaine spatial dont ils est question dans le travail de recherche de thèse adjacent à ce travail de mémoire, mettent une importance absolue dans la sécurité, car une tendance dysfonctionnelle non interceptée à temps peut avoir des conséquences désastreuses sur des vies humaines dans ces domaines.

Galand et al. (2018) parlent également de la recherche de signaux faibles dans les REX dans le domaine spatial. Comme précédemment montré, la recherche spatiale place la sécurité au premier plan, mais également d'autres aspects qui rendent l'analyse des REX et la détection des signaux faibles vitaux dans une perspective de prospérité de l'entreprise. En effet, des accidents pouvant mettre des vies en danger peuvent également être catastrophiques sur le plan financier, car un accident d'appareil peut endommager des composants qui coûtent potentiellement très cher et qui doivent donc être reconstruits ou remplacés. De telles pertes peuvent être évitées grâce à l'analyse de signaux faibles. D'autres conséquences, notamment sur le plan environnemental, pourraient se révéler catastrophiques si les causes de ces incidents ne sont pas corrigées à temps. Dans le contexte de lanceurs spatiaux, si une fusée habitée lancée depuis la Guyane française tombe dans la forêt Amazonienne et explose, en plus des conséquences tragiques de la perte de vie, et les conséquences financières dûes à l'explosion de l'appareil, la propagation des flammes dans une forêt dense et riche comme celle-ci peut se prouver dévastatrice d'un point de vue environnemental.

1.1.5. Avantages de l'extraction d'informations et de la typologie des problèmes techniques

La catégorisation des retours d'expérience et le repérage des différents types d'expression des problèmes sont intéressants en soi, car ils permettent de déceler des tendances et des similarités entre différents faits observés par analyse des signaux faibles. Néanmoins, ces analyses nous permettent également d'accéder à certaines informations contenues directement dans le texte et les extraire afin de remplir des bases de données beaucoup plus normalisées en récupérant les occurrences de mots qui se trouvent dans un certain schéma d'expression. Par exemple, si dans une fiche d'anomalie, l'auteur rapporte que l'appareil « ne s'allume plus », un système d'extraction d'informations se basant sur des patrons d'expression pourra voir que le verbe présent dans la construction « *ne ... plus* » nous renseigne sur la nature du problème, qui dans ce cas est un problème d'allumage.

Des normes d'écriture de retours d'expérience peuvent être mises en place, comme la norme ADREP utilisée dans le domaine de l'aviation (Tanguy et al., 2016), peut bien sûr permettre de formaliser l'expression des problèmes afin de rendre cette extraction d'information beaucoup plus facile. En reprenant notre exemple de problème d'allumage, s'il s'avère que le problème est mentionné dans une construction autre que « *ne ... plus* », et que notre système n'est capable que d'extraire les informations présentes dans cette construction, il ne sera plus capable de nous informer sur la nature du problème. La formalisation de l'expression des problèmes peut se faire au travers d'un set de descripteurs non-ambigus, comme c'est le cas pour le programme de REX de la NASA, le ASRS. Un langage standardisé peut donc être utilisé afin de faciliter l'analyse et l'extraction d'information a posteriori (ASRS utilise ACFT pour « aircraft », WX pour « weather », FLT pour « flight »).

La difficulté à l'adoption et à l'utilisation de normes d'écriture pour les retours d'expérience réside dans une contrainte propre à la complexité de la norme. En effet, plus la complexité de la norme augmente, plus elle permet d'encoder des informations précises de façon détaillée, ce qui permet en retour d'avoir une meilleure compréhension des causes des écarts rapportés dans les fiches d'anomalie. En revanche, plus la complexité de la norme augmente, moins elle est facile d'accès et d'utilisation, et il y a donc un risque que la norme soit utilisée de façon erronée lors de la rédaction de fiches, ce qui fausserait les analyses effectuées en aval et diminue la pertinence d'utilisation de la norme. De cette façon, il arrive très souvent que les normes soient appliquées a posteriori par un analyste, plutôt que lors de la rédaction de la fiche, qui se fait donc de façon plus ou moins libre.

Le but des techniques de TAL comme celles mentionnées ci-dessus est de permettre de catégoriser et d'extraire des informations pertinentes dans le texte libre sans forcément avoir besoin de s'appuyer sur la qualité de l'encodage du rapport. Il est tout de même intéressant de noter que des techniques de TAL peuvent également être utilisées lors de la rédaction de ces rapports afin de standardiser le langage utilisé, les structures syntaxiques permises, et la façon dont les informations sont structurées dans le texte, dans le but d'avoir des algorithmes de catégorisation et d'extraction d'information plus faciles à mettre en place, et potentiellement plus rapides et également moins coûteux, dans le cadre d'une analyse a posteriori.

1.2. La Communication Médiée par les Réseaux et la demande d'aide sur les réseaux

1.2.1. La Communication médiée par les Réseaux

Les nouvelles technologies et l'avènement d'internet, ainsi que sa popularisation autour du globe, ont donné place à de nouvelles formes de communication, basées ou non sur des formats de communication analogues déjà existants. En effet, grâce aux opportunités de communication que nous donnent internet et les nouvelles technologies, nous avons assisté à la naissance des différentes formes de communication en ligne, que l'on peut regrouper sous l'appellation de Communication Médiée par les Réseaux ou CMR en français, ou bien Computer Mediated Communication en anglais, souvent noté sous l'acronyme CMC (Poudat et al., 2020). Avec la proéminence de l'anglais dans ces nouvelles formes de communication, il arrive également que les études françaises utilisent l'acronyme CMC pour désigner ces types de communication.

Les différents types de communication regroupés sous l'appellation de Communication Médiée par les Réseaux présentent les avantages de la communication écrite, à savoir notamment la nature permanente de la trace écrite qui pourrait ensuite servir à une analyse, ainsi que la facilité dans le traitement des informations contenues dans cette trace écrite. Cependant, la CMR présente

également les avantages associés à la communication telle que l'on peut la retrouver dans des contextes d'expression orale, comme la spontanéité de l'expression ou la fluidité du transfert d'information (Romiszowski & Mason, 2004). Une des raisons pour laquelle la CMR présente autant d'avantages est la diversité des modes de communication qui sont regroupés sous l'appellation CMR. En effet, la CMR ne se cantonne pas au domaine écrit, mais admet également de la communication sous forme orale ou même sous format vidéo.

De la même façon, la CMR présente également une diversité sur le plan des temps de communication (Nejad et al., 2021). La communication médiée par les réseaux peut se faire de façon synchrone, c'est-à-dire que les locuteurs communiquent et voient les réponses de leurs interlocuteurs en temps réel, mais elle peut également se faire de façon asynchrone, c'est-à-dire que l'expression est destinée à être enregistrée et vue ou lue en différé dans le temps.

Alors que l'expression écrite classique est par nature destinée à être lue de façon différée, ce qui en fait un type de communication asynchrone, la CMR permet une communication écrite synchrone sous la forme de messagerie instantanée par exemple. Celle-ci permet également une communication écrite qui s'inscrit plutôt dans une asynchronicité au travers des emails ou des forums de discussion.

D'un autre côté, alors que l'expression orale est habituellement synchrone lors de discussions en face à face par exemple, celle-ci peut s'adapter à une communication asynchrone sous la forme de messages vocaux ou de messages vidéo envoyés par email. Cependant la communication orale synchrone est également facilitée par la CMR, via les messages vocaux sur les plateformes de messagerie instantanée, ou même directement via les applications de conversation (ou *chat*) vidéo comme Skype ou Zoom dans un cadre professionnel ou académique.

Malgré cette dualité opposant apparemment deux pôles aux fonctionnements différents, la communication médiée par les réseaux est en réalité hybride dans beaucoup de cas. En effet, les messages envoyés par messagerie instantanée ne sont pas forcément lus instantanément par l'interlocuteur, et à l'inverse les discussions sur les forums ou les échanges d'information par email peuvent se faire de façon quasi-instantanée, ce qui floute la frontière hypothétique entre les deux temps de communication. De la même façon, la CMR est hybride du point de vue des modes de communication. Les services de messagerie instantanée sont encore une fois un bon exemple, car il est possible d'envoyer à la fois des messages textuels et des messages vocaux, ainsi que d'alterner entre les deux types de communication, au sein d'une même conversation.

1.2.2. La CMR dans l'étude linguistique

La Communication Médiée par les Réseaux apporte une nouvelle source de matière langagière à étudier dans un contexte linguistique. Dû à la nature unique mais plurielle de la CMR, celle-ci fait émerger des types d'expression et des comportements spécifiques au moyen (ou moyens) d'expression. La nature informatique des données permet également une étude linguistique basée sur des outils informatiques qui s'impose comme une norme dans une époque où les données langagières sont nombreuses, volumineuses, et pas toujours facilement traitables manuellement.

L'étude de Mondada (1999) présente les débuts de l'analyse linguistique du contenu en communication asynchrone sur les réseaux, à la fin d'une décennie de diffusion massive des technologies de l'information et de la communication, notamment avec l'essor de la communication par Internet. C'est en effet dans ces années que l'utilisation d'Internet se diversifie pour ne plus être concentrée sur la recherche d'informations, mais également pour constituer un moyen d'entretenir un lien social à distance. On voit donc, dès la fin du XXe siècle, des schémas conversationnels propres à la CMR apparaître notamment dans les courriels, dans lesquels nous voyons par exemple dans une réponse à un courriel, une partie du courriel précédent qui est repris (ex: « vous suggérez que … ») afin de répondre à certaines parties des courriels de façon claire.

La communication médiée par les réseaux est également marquée par des rituels de politesse qui peuvent également être analysés sous un angle linguistique, même s'il est également intéressant d'étudier les CMR dans le cadre de la sociologie. Amato et Boutin (2013) s'intéressent à ces rituels linguistiques présents dans les CMR, et notamment dans la communication par forums, dans lesquels les messages se suivent sous forme de fil de discussion, et chaque message peut-être adressé à (et lu par) tous les usagers du forum. Au-delà d'une certaine liberté dans la forme des messages ou dans la créativité linguistique (présence de verlan, d'argot), l'observation de marqueurs de politesse dans les messages est intéressante, compte tenu du sentiment d'anonymat qui peut accompagner la communication sur les réseaux. L'étude remarque une diminution des marqueurs de politesse sur le forum Usenet entre 2000 et 2010, même si celles-ci restent quand même souvent présentes. Dans certaines chartes de bonne conduite de forums, il n'est d'ailleurs pas rare d'observer un encouragement à la civilité, et au contraire un découragement de comportements langagiers injurieux ou hostiles.

Malgré quelques études sur d'autres langues, les ressources linguistiques en termes de CMR sont essentiellement en anglais. Cependant, certaines initiatives apparaissent de plus en plus pour les études de comportement linguistique sur les réseaux. C'est le cas du corpus CoMeRe, qui est un corpus regroupant des conversations de chat et de SMS en français avec une couverture géographique variée (Chanier et al., 2014). De tels projets s'inscrivent dans une optique de passage de la linguistique manuelle au Traitement Automatique des Langues, et dans le cas de CoMeRe, il s'agit également d'un effort de création de corpus de référence en langue française qui n'existe pas jusqu'alors. En effet, avec l'augmentation des données et l'amélioration des techniques d'analyses computationnelles, le TAL devient un domaine de plus en plus important dans la recherche, mais est surtout très bien adapté à des données de CMR qui sont créées directement au format numérique et qui sont plus rapidement et plus facilement accessibles pour une analyse avec des outils de plus en plus performants, à l'inverse d'autres corpus qui ont dû être numérisés.

Lors de la création d'un corpus de CMR, nous devons être conscients des choix que nous faisons, car à chaque étape il est possible de faire des choix qui viendront biaiser les données et nous donner une image différente de la réalité lors des analyses. Ces choix peuvent être de plusieurs ordres, comme le choix des données, le choix de l'équilibre et de la représentation des échantillons par rapport à la population étudiée, mais également au niveau de l'encodage des données qui permettent plus ou moins facilement d'exporter, de partager, et de réutiliser les données, et aussi au niveau légal, dans la distribution des données et leur utilisation dans le domaine de la recherche. Ces questions ne sont pas triviales et influencent grandement l'étape de récupération de données dans le cadre de la création d'un corpus et de l'étude de ces données.

L'encodage des données, et également des métadonnées, peut se révéler important pour la facilité d'accès et d'analyse des données. Ces métadonnées sont facilement accessibles dans le cadre des CMR et il est important de les inclure dans les corpus, notamment pour les informations presque systématiquement présentes comme l'horodatage ou l'auteur du message. L'encodage du mode de communication de départ est également important pour une étude sur les CMR qui présentent une variété de modes et de temps de communication comme évoqué précédemment (Poudat et al., 2020). Le cas de CoMeRe donne un exemple d'encodage des données sous la norme TEI, qui est une norme devenue standard dans l'élaboration de corpus en Sciences Humaines et Sociales, et sur laquelle nous nous attarderons plus tard lors de la création de notre corpus.

L'utilisation des données de CMR dans le cadre de la création d'un corpus amène également à se poser des questions d'un point de vue éthique et légal. En effet, les données récupérées sur Internet ou par SMS par exemple peuvent être soumises à des contraintes auxquelles ne sont pas soumises d'autres textes disponibles de façon publique comme dans les journaux ou les romans. Tout d'abord, ces données, si hébergées sur Internet, peuvent l'être sous une licence qui ne permet pas de les collecter et de les réutiliser. Dans le cadre des forums, la personne ou l'organisme qui gère le forum peut décider de l'utilisation qui pourra être faite des informations que les utilisateurs décident de poster. Ainsi, pour protéger les utilisateurs, mais aussi pour éviter certaines répercussions légales ou éthiques dûes à des manquements au niveau du site, beaucoup de sites décident d'interdire la

réutilisation des données.

La sensibilité des données joue également un rôle dans la réflexion sur l'utilisation des données de CMR. Ghliss et André (2017) expliquent que les données sensibles, définies comme étant des données laissant apparaître l'appartenance d'un personne à quelconque groupe, organisation, idéologie, ethnie ou façon de fonctionner par la Commission Nationale de l'Informatique et des Libertés (CNIL), sont difficilement traitables telles quelles et demandent un certain niveau de prétraitement. Il en va de même pour les données à caractère personnel, qui sont des indices pouvant permettre d'identifier une personne. Dans le cadre de la protection de la vie privée des personnes, la CNIL demande aux chercheurs manipulant ce genre de données de faire en sorte de ne pas mettre en danger les participants aux études ou encore de préserver leur vie privée. De cette façon, il est important en tant que chercheur lors de la création d'un corpus, de passer par des étapes qui garantiront ces aspects, comme une anonymisation des auteurs, un remplacement des informations sensibles par des marqueurs vagues, mais de faire en sorte également de garder les données utilisables. Ainsi, les informations de genre dans les textes pourront être gardées par exemple, car celles-ci ne sont pas assez spécifiques pour permettre une identification du locuteur ou de l'interlocuteur.

1.2.3. Un cas de CMR : la communication par forums

En ce qui concerne la communication par forums, essentiellement asynchrone, celle-ci peut se faire dans un grand nombre de contextes et de domaines très variés. En effet, alors que les premiers forums servaient surtout à discuter de problématiques liées à l'informatique et de problèmes techniques, car ils étaient destinés aux rares utilisateurs qui pouvaient et savaient manipuler un ordinateur, des forums pour tous types de domaines ont vu le jour avec la démocratisation de l'ordinateur et d'internet. De nos jours, nous pouvons retrouver des forums d'aide aux jeux vidéo, des forums de critique de films, ou encore des forums d'aide aux problèmes médicaux ou de santé.

Dans ce genre de contexte de communication sur les forums, les utilisateurs ont notamment l'habitude d'utiliser ce genre de plateforme afin de demander de l'aide par rapport à un sujet particulier, que ce soit pour résoudre un problème d'ordre mécanique, médical, ou pour savoir comment terminer un niveau particulièrement difficile dans un jeu vidéo. L'attente de ces utilisateurs exprimant un problème est que d'autres utilisateurs pourront leur venir en aide et communiquer avec eux afin de trouver une solution au problème. Si certains utilisateurs trouvent qu'une question provoque trop de gêne pour pouvoir en parler à ses proches, ils peuvent utiliser le principe d'anonymat que permet internet afin de poser leur question en toute discrétion.

Au-delà des forums de discussion classiques, de plus en plus de sites qui s'apparentent à des réseaux sociaux où à des communautés en ligne permettent d'exprimer un problème et de demander des solutions à des étrangers de façon anonyme ou non. Le site internet Reddit est un bon exemple de ce genre de sites sur lesquels des communautés se créent afin de parler de sujets extrêmement variés. Les utilisateurs du site Reddit se regroupent en communautés appelées *subreddits* afin de converser, à la manière d'un forum, sur des sujets divers dans la limite de ce qui est permis par les conditions d'utilisation générales du site.

Des sites de forum plus particuliers ne s'attardant que sur un seul domaine existent également, c'est le cas par exemple de forums de bricolage et de réparation comme le site commentreparer.com. Sur ce site, les utilisateurs sont invités à poster un message (ou *post*) détaillant un problème qu'ils peuvent rencontrer concernant un appareil ou un objet particulier, qu'il soit électronique ou qu'il fasse plutôt partie du mobilier de maison. Ils sont invités à le faire de façon succincte sous forme de titre, ainsi que de façon plus détaillée sous forme de description du problème. Nous proposons ci-dessous un exemple d'un tel post, sous forme de capture d'écran d'une page du forum.

Figure 1 - Capture d'écran d'un post avec son titre, son contenu, et ses métadonnées

D'autres utilisateurs peuvent ensuite proposer des pistes pour la résolution du problème, ou demander des informations supplémentaires pour aider à la résolution du problème. Ce genre de site correspond relativement bien au type de données dont nous avons besoin pour simuler des fiches d'anomalies, et nous détaillerons leur utilisation plus tard dans ce dossier.

Nous avons pu faire un tour des problématiques liées à l'étude de l'expression des problèmes techniques dans différents contextes dans cette première partie. Les pistes de réflexion apportées par notre état de l'art nous permettent ensuite d'avancer de façon réfléchie dans les différentes étapes constituant la démarche d'étude que nous entreprenons. En effet, dans le cadre de notre étude, nous nous penchons sur l'expression des problèmes techniques, mais nous devons nous adapter également aux spécificités linguistiques des données de CMR. De plus, nos données amènent des questionnements sur les aspects techniques et légaux de l'utilisation de telles données, pour lesquels nos lectures ont permis d'apporter certains éléments de réponse.

2. Données de travail et visée de l'étude

Dans cette deuxième partie, il est question d'utiliser de nouvelles données afin de mener à bien notre étude. Ces données n'étant pas directement disponibles sous forme de corpus, nous décrivons ici le processus de constitution du corpus dans ses différents aspects, en prenant en compte les difficultés techniques et les questionnements soulevés par notre état de l'art. Nous présentons ensuite les données récoltées et compilées sous forme de corpus, afin d'avoir une vision globale de ce que contient le corpus, et de rapprocher nos observations à l'apport théorique de l'état de l'art sur le sujet. Cette présentation du corpus nous permet également d'observer ses spécificités et de dégager des composants à prendre en compte pour la typologie.

2.1. Constitution du corpus

2.1.1. Réflexion sur les sources de données utilisables

Comme il a été mentionné dans la partie Introduction de ce dossier, nous n'avons pas la possibilité d'utiliser les données de fiches d'anomalie du CNES, ce qui nous amène à devoir compiler un corpus de fiches contenant des expressions de problèmes plus ou moins équivalentes à celles présentes dans la base de données du CNES. Dans cette partie, nous allons donc présenter le cheminement de notre réflexion quant aux sources de données utilisables pour constituer un tel corpus.

Avant tout, il faut s'assurer que les équivalents de fiches d'anomalie correspondent un minimum au type de fiches que nous pourrions retrouver dans la base de données du CNES concernant les lanceurs Ariane. De cette façon, notre corpus doit contenir des fiches contenant l'expression de problèmes liés à des appareils ou des machines, car le côté technique est supposément un pilier central des fiches d'anomalie du CNES. Cependant, puisque les fiches d'anomalie ne sont pas uniquement centrées sur les écarts à la norme constatés sur des machines, il est également possible d'inclure dans notre corpus des fiches ayant moins de rapport avec des machines, tant que celles-ci expriment un problème. Il faut tout de même souligner le fait que, peu importe les données que nous obtenons et utilisons pour la création du corpus, celles-ci n'auront qu'une correspondance minimale avec le type de données très spécialisées présentes dans les bases de données du CNES, ce qui signifie qu'il ne sera de toute façon pas pertinent de chercher les données les plus fidèles. Ce qui nous intéresse étant la forme de l'expression du problème technique plutôt que le fond, des données ressemblantes sur la forme restent tout de même intéressantes à étudier.

Comme évoqué dans la partie 1.2.3. de ce mémoire, les forums d'aide sont une source d'information facilement utilisable lorsque l'on a besoin d'analyser des façons d'exprimer un problème. En effet, comme mentionné, les utilisateurs postent un message sur ces forums car ils ont en général besoin d'aide dans la résolution d'un problème, ce qui correspond au type de données dont nous avons besoin. Il est tout de même important de garder à l'esprit le fait que malgré les nombreuses similarités entre ces données de forum et les Fiches d'Anomalie, la visée d'une FA est néanmoins différente, car la résolution d'un problème ou la demande d'aide ne représentent pas le but premier et immédiat du rédacteur de la FA, même si la résolution du problème se place au coeur du processus de REX qui amène la rédaction de ces FA.

La création d'un corpus nous amène naturellement à nous poser un certain nombre de questions, dont la première vise les types de données et les informations qui doivent y figurer. Dans le cadre de notre étude, le corpus dont nous avons besoin est un corpus contenant ce qui pourrait s'apparenter à des fiches d'anomalies, c'est-à-dire de descriptions de problèmes techniques. Il est donc admis que nous sélectionnons comme sites candidats des forums d'aide centrés sur les problèmes techniques (forums de bricolage, forums de dépannage, forums de SAV etc.).

Afin de construire un corpus à partir de telles données, nous devons être capables d'extraire ces données afin de les encoder dans un format qui nous permettra par la suite d'assurer la facilité d'utilisation et d'échange de ces données. L'accessibilité des données nous pose donc problème, car afin de permettre à des moteurs de recherche et autres systèmes d'accéder aux données d'un site de façon automatique, les sites peuvent disposer d'une sorte de carte du site (autrement dit, une *sitemap* en anglais) qui permet à un programme censé récupérer les données du site de le faire facilement en accédant aux pages suivant les liens disponibles sur cette sitemap. Les données peuvent également être structurées sur les pages du site afin de faciliter leur extraction.

Les contraintes de structure de la page constituent un premier obstacle technique à la constitution d'un corpus. En effet, alors que la présence d'une telle structure sur une page permet une facilité dans l'extraction des données, il arrive souvent que les forums ne soient pas très structurés dans les balises qui constituent leur site, et il arrive également souvent que le site ne proposent pas de sitemap, ou que celle-ci, si elle existe, ne permet pas d'accéder à une hiérarchie d'informations comme présentée sur le site.

Pour l'extraction des données, et pour avoir des métadonnées intéressantes et réutilisables, nous filtrons également nos sites au travers de la disponibilité de certaines données sur les appareils ou objets présentant un problème. Si ceux-ci sont annotés séparément, ils sont ensuite plus facilement reconnaissables dans le texte et il est également possible de résoudre certaines inférences qui pourront être faites. De plus, l'absence d'espaces prédéfinis pour le renseignement de la marque et du modèle de l'appareil par exemple, peuvent amener certains posts à ne pas faire de mention explicite de la source du problème rencontré, ce qui rend le traitement automatique plus compliqué. D'autres métadonnées, comme la catégorie ou le type d'appareil, permettent une classification par type d'appareil et une analyse différente par catégorie le cas échéant. Nous sélectionnons donc les sites pour lesquels ces informations sont demandées et ensuite facilement accessibles sur la page du site.

Cette facilité dans l'accessibilité constitue un point important à considérer dans le processus technique de récupération des données lors de la création du corpus. En effet, la structure des pages à partir desquelles les informations sont récupérées puis stockées, permet de guider le programme lors de la récupération des données. Les balises HTML utilisées sur les pages peuvent indiquer le type d'information qui sont contenues dans ces balises, mais ce n'est pas une obligation, et il arrive souvent que les balises ne donnent pas d'informations supplémentaires que ce qui permet d'afficher correctement les informations sur la page pour les utilisateurs humains. Un utilisateur humain peut facilement comprendre, si la page dispose d'un texte affichant la marque et le modèle de l'objet, que le texte compris dans cette partie de la page correspond aux informations qui doivent être extraites et doivent exister dans les métadonnées du post sous ces identifiants. Un programme informatique, en revanche, ne pourra pas inférer ces informations et l'importance de ces informations sans avoir une indication supplémentaire dans la balise HTML définissant l'identité de l'information textuelle qui y est comprise.

Le contenu réel des messages sur ces forums constitue un autre obstacle à la récupération de données intéressantes. Même si les forums sur lesquels nous nous penchons sont des forums dans lesquels il est question de problèmes sur des machines, c'est-à-dire des **problèmes techniques**, il arrive souvent que les forums de discussion présentent beaucoup de messages ne relatant pas un problème, mais demandant plutôt explicitement de l'aide, ou n'existant que pour exprimer un avis positif ou négatif sur un appareil. Il s'agit donc également d'un aspect à prendre en compte dans notre recherche de source pour notre corpus. Nous cherchons donc à filtrer les sites candidats en ne retenant que ceux qui sont axés sur la recherche d'aide, et dans lesquels les fiches ou posts dédiés à la discussion d'un sujet, l'échange d'idées, ou d'autres posts qui ne présentent pas l'expression d'un problème technique, soient inexistantes ou gardées à un minimum.

Cependant, le plus gros obstacle réside dans la disponibilité des données dans le but de créer un corpus libre d'utilisation, ou au moins dans le but d'analyse au sein de ce projet de recherche et dans la sphère scientifique. En effet, il s'agit d'un obstacle considérable car sans l'autorisation des sites

et des organismes auxquels appartiennent les données, il est impossible d'utiliser et de partager les données sans risquer des poursuites judiciaires. Après notre recherche, nous nous rendons compte malheureusement qu'il est explicitement écrit sur la majorité des sites que nous pensions utiliser, qu'il était interdit d'utiliser leurs données.

Comme abordé dans la partie 1.2.2, il est très important de considérer les aspects éthiques et légaux de la récupération de données sur des sites internet pour la recherche. Krotov et al. (2020) soulignent bien l'importance de ces réflexions sur des plans qui, si trop peu ou pas considérés, peuvent amener à des conséquences légales dans différents domaines comme le copyright ou encore la protection des individus et de leurs données. Les lois sur la protection des données de la CNIL représentent les lois les plus proches à la manipulation de ce type de données, car il n'existe pas de lois spécifiques à la récupération des données de façon automatique sur des sites internet. Les questionnements sur lesquels il faut donc se pencher dans le contexte de création de corpus de cette façon sont les questions d'utilisation des données et si celle-ci, ainsi que l'analyse de ces données, est permise par les conditions générales d'utilisation du site, mais également si la récupération de ces données ne viole pas de libertés individuelles et ne met pas en danger les personnes. Il est important de réfléchir également à l'acte d'extraction des données, car celui-ci peut constituer une charge importante pour les serveurs sur lesquels est hébergé le site, et il convient donc de faire en sorte que l'extraction des données n'endommage pas le fonctionnement du site et de ces serveurs. Même si ces derniers questionnements ne sont pas explicités par la loi de la même façon dans tous les pays et dans tous les domaines, les questionnements d'ordre éthique sont tout aussi importants pour la création de telles données. Krotov et al. (2020) concluent tout de même que si les conditions générales d'utilisation font mention ou non de la possibilité ou l'interdiction du prélèvement des données du site, il est tout de même possible de s'entretenir avec le propriétaire du site ou de l'hébergeur des données afin de demander directement une confirmation ou infirmation de la possibilité d'extraction et d'utilisation des données.

Puisque nous avions déjà notre oeil sur le site commentreparer.com, et que celui-ci ne présente pas d'interdiction explicite d'utilisation de ses données, nous avons demandé au propriétaire du site si les données de son site étaient utilisables dans le cadre de la création d'un corpus à but d'analyse linguistique, ce à quoi le propriétaire nous a donné une autorisation écrite par email. Nous indiquons dans le tableau 1 ci-dessous les différentes facilités et inconvénients que présente chaque site, sachant que les données sont inutilisables sans accord d'utilisation avec le propriétaire de la base de données. C'est pour cette raison que nous avons choisi les données du site commentreparer.com afin de constituer notre corpus.

Nom du site	URL du site	Que des fiches de problème et pannes	Catégories clairement délimitées	Marque et modèle facilement accessibles	Présence d'un sitemap.xml	Autorisation d'utilisation	Solution clairement dénotée
CommentReparer.com	https://www.commentreparer.com/	Y	Y	Y	N	Y	Y
Forum Futura Sciences Dépannage	https://forums.futura-sciences.com/depannage/	Y	N	N	N	?	N
Communauté SAV Darty	https://sav.darty.com/	N	N	Y	N	N	Y
Forum Système D - Equipement de la maison	https://www.systemed.fr/forum-bricolage/equipement-de-la-maison-f76.html	N	N	N	N	N	N
Forum ADEPEM	https://forum.adepem.com/categories	N	Y	N	Y	N	N
Bricolage L'internaute Electroménager	https://bricolage.linternaute.com/forum/electromenager-14	N	Y	N	Y	N	Y
Bricoleur du Dimanche : Forum bricolage	https://www.bricoleurdudimanche.com/forums/forums-bricolage.html	N	Y	N	Y	N	N
Forum électroménager	http://www.forum-electromenager.com/	Y	Y	N	Y	N	N
Forum Spareka	https://forum.spareka.fr/	N	Y	Y	Y	?	N

Tableau 2 - Résumé des recherches de sites internet dans le cadre de la construction d'un corpus

2.1.2. Réflexion sur les données

Les données présentes sur le site commentreparer.com sont des données quelque peu différentes de ce que l'on peut trouver dans des fiches d'anomalie d'entreprises, pour différentes raisons. Tout d'abord, les textes mêmes ne sont pas uniformes, tant sur le style d'écriture, que sur la longueur, ainsi que sur le format et la disposition des informations sur la page. Certaines informations peuvent parfois apparaître mais pas à l'endroit prévu à cet effet par le site (marque et modèle de l'appareil qui devraient être indiquées en métadonnées ne se trouvant que dans le corps du texte), et certaines informations requises pour comprendre le problème sont parfois totalement absentes du message. Toutes ces contraintes rendent une potentielle analyse complexe, car il faut prendre en compte la multitude de possibilités de représentation de l'information afin de pouvoir analyser les textes.

De plus, le format n'encourage pas l'utilisation d'une typologie détaillée permettant de mieux classer et d'encoder les informations de façon efficace. En effet, il ne s'agit pas du but des posts de forum qui, à la différence des fiches d'anomalie, sont rédigés par des personnes potentiellement non-expertes du domaine, et destinés à être lus également par des humains pas forcément experts du domaine pour être compris, donc les informations doivent être accessibles aux non-experts pour que le message remplisse sa fonction. Néanmoins, Tanguy et al. (2016) montrent les avantages d'une typologie dans le processus d'analyse et de catégorisation des fiches en aval. Il existe tout de même des champs différents en fonction des informations à fournir lors de la création d'un post sur le forum, comme un espace dédié pour renseigner la marque et le modèle de l'appareil dont il est question. De plus, la construction du site donne lieu à certaines constantes textuelles influencées par la façon dont les champs d'entrée sont agencés. Nous pouvons citer l'exemple de *Bonjour* qui apparaît au début d'un grand nombre de messages, simplement car ce mot est déjà pré-rempli dans le champ d'entrée de la description détaillée du problème lors de la création d'un post. Ce type d'occurrences fait écho aux attentes en matière de politesse discutées dans la partie 1.2.2. Nous montrons ci-dessous un aperçu de l'interface de saisie d'un post et de ses différents champs comme décrits ci-dessus. Malgré les indications visibles sur cette image, et donc visible par les auteurs de posts, il arrive que ces indications ne soient pas respectées (exemple : pas de nom d'appareil dans « Erreur f03 » malgré l'indication « Votre question doit comporter l'appareil concerné [...] »)

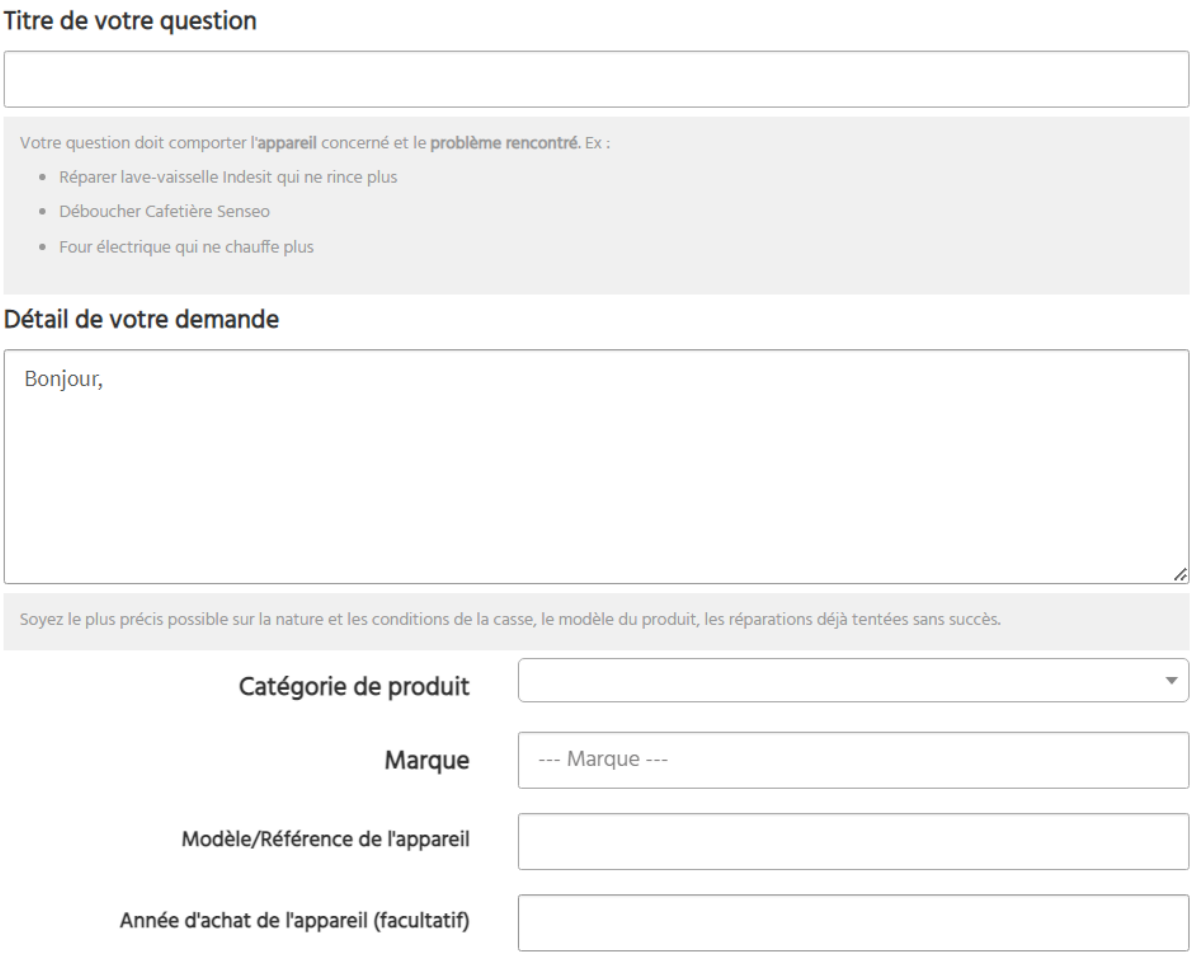

Figure 2 - Capture d'écran de l'interface d'écriture d'un post sur le forum commentreparer.com

L'analyse des données textuelles est également rendue difficile par une autre différence avec les fiches d'anomalie rédigées par des experts dans un domaine, qui est que les messages de forum sont souvent rédigés par des personnes non expertes qui peuvent utiliser du vocabulaire non précis et des tournures de phrases vagues afin de décrire leur problème (exemple : « [...] Je dois vous préciser par ailleurs qu'il manquait visiblement sur la « tête » du bras (**le terme n'est sans doute pas le bon je n'y connais pas grand chose**) [...] »). Alors que la description vague peut être comprise par les humains grâce à leur capacité d'inférence, ces informations ne sont pas présentes directement dans le texte et cette imprécision constitue donc une difficulté supplémentaire à l'analyse.

Même lorsque les mots précis décrivant un problème sont présents dans le texte, nous rencontrons un autre problème de taille lors de l'exploitation des textes, qui est que beaucoup d'utilisateurs écrivent en faisant un certain de nombre de fautes d'orthographe plus ou moins sévères, ce qui empêche la reconnaissance et la lemmatisation des mots, et rend difficile l'exploitation des données (exemple : « Plus aces avec télécommande » pour signifier « plus d'accès avec télécommande »).

Enfin, les données du forum ne portent pas toujours sur des problèmes absolument similaires aux problèmes rencontrés dans l'industrie, car le site propose également des catégories qui s'éloignent de l'expression de problèmes comme ce que nous pouvons retrouver dans un corpus industriel autour de machines. En effet, les catégories « Linge, vêtement, bijoux » ou encore « Mobilier, maison » s'éloignent des fiches d'anomalie que l'on peut trouver concernant les fusées Ariane. Cela découle notamment du fait que l'environnement dans lequel apparaissent les problèmes (maison ou usine par

exemple) différent en problématiques et surtout en objets présents et risquant de causer des problèmes. Il s'agit donc également d'un aspect à prendre en compte lors de l'extrapolation des résultats à la suite des analyses et catégorisations que nous effectuerons.

2.1.3. Mise en forme d'un corpus

Comme nous avons pu le voir, la constitution d'un corpus de CMR n'est pas une tâche facile, car il faut faire des choix à chaque étape de la création pour s'assurer d'avoir les données les plus intéressantes et les plus représentatives possibles, tout en naviguant entre les différentes difficultés posées par les données et les droits d'accès à ces données. De plus, les problématiques sur lesquelles il convient de se pencher lors de la création d'un corpus de CMR sont des problématiques nouvelles par rapport à ce qui est habituel lors de la création d'un corpus écrit classique (Poudat et al., 2020).

Il y a également autant de façons de créer un corpus que de types de données dont on dispose et de ce que l'on veut en faire. On pourra faciliter l'accès à certaines informations dans nos données ou structurer celles-ci de façon à rendre certaines analyses plus faciles en fonction des analyses visées par exemple. La CMR se présentant sous multiples formes, écrite, audio, vidéo, synchrone ou non, chaque format et chaque visée d'étude appelle à réfléchir à des problématiques différentes et spécifiques à la visée de notre corpus.

Dans le cadre d'une étude sur l'expression des problèmes techniques dans des messages de forum comme nous proposons d'effectuer dans ce projet de recherche, il peut être intéressant d'analyser les fils de discussion afin d'y trouver des patrons typiques de reformulation de problème et des informations supplémentaires non communiquées dans le message initial, mais cela se révèle relativement compliqué dans le cadre de ce mémoire, et le corpus prendrait une place trop importante par rapport à ce que nous proposons de faire avec les données. Nous resterons donc, dans notre première version du corpus, à la sauvegarde du premier message de chaque fil de discussion, qui présente l'expression initiale du problème, afin de n'analyser que celui-ci. Il pourrait être envisageable plus tard cependant, d'extraire la totalité des messages pour tous les fils de conversation, ainsi que tous les types de messages présents sur chaque conversation, afin d'effectuer des analyses plus approfondies sur les fils de discussion.

Afin d'encoder nos données de façon structurée, nous utilisons le langage de balisage XML. XML est un langage de structuration des données dérivé de SGML, qui nous permet d'annoter des données de telle sorte que celles-ci soient plus facile d'accès par des programmes informatiques. Néanmoins, le langage XML nous donne beaucoup de libertés, ce qui en retour ne permet pas un encodage efficace des données. Si nous ne suivons pas une norme d'encodage, nos données pourraient correspondre à n'importe quelle information, et la nature de l'information et du texte, ainsi que les informations autour du texte que l'on appelle **métadonnées**, seraient difficilement distinguables les unes des autres.

Pour pallier ce problème, une norme internationale est apparue dans le domaine des Sciences Humaines et Sociales, il s'agit de la norme TEI. Selon Romary et Hudrisier (2002), la norme TEI (Text Encoding Initiative) « est une norme de balisage, de notation et d'échange de corpus des documents électroniques fondée sur le SGML ». Il s'agit en clair d'une norme permettant l'encodage électronique de textes dans le but de créer des corpus qui pourront ensuite être facilement partagés et analysés. Nous utiliserons donc la version TEI-P5 dans notre corpus pour structurer les données.

Pour l'encodage d'informations dans le corpus, nous utilisons deux balises très importantes qui délimitent les données pour chaque texte. Il s'agit des balises <*teiHeader*> et <*text*>. La balise <*teiHeader*> contiendra les métadonnées associées au corpus entier comme l'auteur, l'organisme prenant en charge la construction du corpus, mais également les métadonnées associées au texte, comme l'auteur du message ou le lien de la page d'où vient le message. La balise <*text*> contient les informations présentes sur la page, c'est-à-dire le texte détaillé ainsi que son titre, mais également

l'horodatage du message et les catégories auxquelles ce message appartient.

La norme TEI présente des avantages considérables pour le partage ainsi que la réutilisation future des données d'un corpus. En effet, son but est de proposer un encodage « simple, clair et concret » (Romary & Hudrisier, 2002) des textes afin de faciliter leur traitement grâce notamment à une lisibilité humaine, mais également à une compatibilité avec les standards logiciels existants. Beaucoup de logiciels et de plateformes d'analyse de corpus utilisent en effet la norme TEI pour le stockage et l'analyse des données, ce qui en fait un choix judicieux lors de la création d'un nouveau corpus.

Cependant, l'utilisation de la norme TEI pose certains problèmes lors de la structuration de données de CMR. La CMR contient de nouveaux modes de communication, et à ce titre, les normes pour l'encodage de ces types de communications sont encore instables. En attendant la stabilisation des normes pour l'encodage de ces types de données, il faut se résigner à tenter d'utiliser des balises correspondant à ce qui nous semble être la catégorisation la plus proche des informations que nous pouvons trouver dans des corpus relatant des modes de communication plus classiques (Poudat et al., 2020).

Enfin, la norme TEI se veut claire et concise, mais elle se veut également détaillée, et comme nous l'avons vu précédemment, une augmentation du degré de détail attendu dans une norme rend souvent cette norme difficile à appréhender, surtout par un utilisateur débutant. Il en résulte que la documentation de cette norme est relativement lourde et difficile à lire en entier pour avoir un aperçu de tous les outils mis à disposition dans cette norme. Les façons d'encoder sont donc également floues car les balises se veulent capables de contenir toute une multitude de types d'informations différentes, et il est d'autant plus difficile de savoir comment encoder des données venant de modes de communication plus récents comme la CMR.

2.1.4. Sélection des informations et automatisation de la création du corpus

Lors de la création d'un corpus, nous devons être conscients des choix que nous faisons, car à chaque étape il est possible de faire des choix qui viendront biaiser les données et nous donner une image différente de la réalité lors des analyses. Pour cette raison, et afin de rester le plus neutre possible et englober le plus de données, nous choisissons de prélever le premier message et ses métadonnées pour chacun des fils de discussion présent dans chacune des neuf catégories du site commentreparer.com, à condition d'avoir accès au fil de discussion. Ce premier message est également susceptible de contenir des images ou des fichiers donnant plus de détails sur la nature du problème, que nous choisissons également de référencer. Enfin, certaines informations sont absentes de quelques messages alors qu'elles sont présentes dans la plupart des autres messages. Dans ce cas, le message est tout de même extrait, et les informations inexistantes donneront lieu à des balises vides dans le corpus.

Dans cette première version du corpus, le premier message est le seul à être extrait. Nous avons choisi d'inclure, pour chaque message, son titre, le corps du message, l'horodatage, le lien vers la page, la catégorie et la sous-catégorie auxquelles appartient le message, les fichiers supplémentaires éventuels, ainsi que les informations concernant la marque, le modèle, et éventuellement l'année de commercialisation de l'objet ou de l'appareil dont il est question. Voici ci-dessous un aperçu des informations que nous extrayons des posts du forum dans cette version du corpus, où les informations extraites sont encadrées en rouge. D'autres messages accompagnés de métadonnées similaires pourront être également captés dans une version future du corpus.

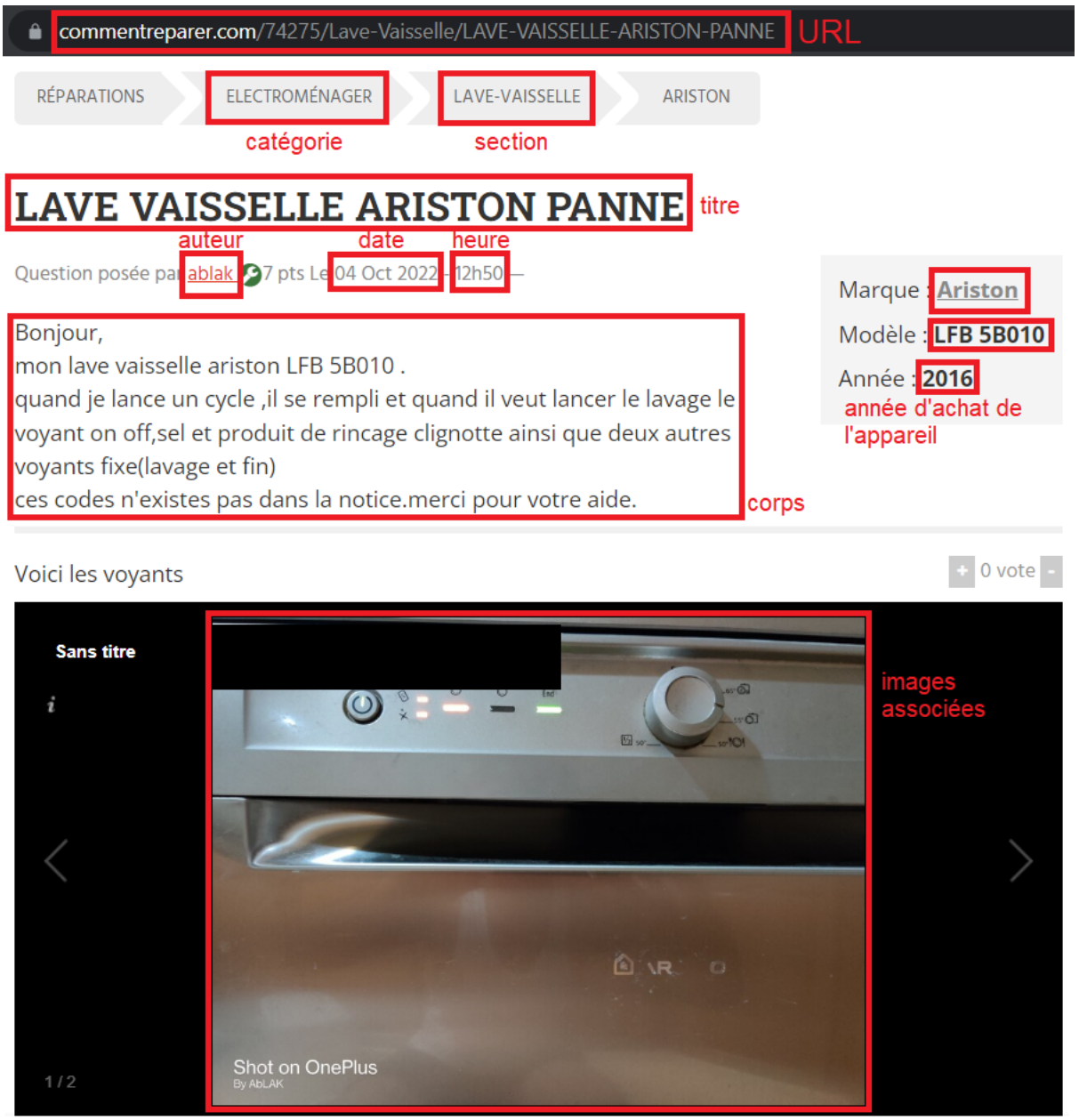

Figure 3 - Informations extraites des posts sur commentreparer.com pour la constitution du corpus

Pour extraire les données et constituer le corpus de façon automatique, nous créons et utilisons un « webcrawler » codé à l'aide du langage de programmation Python. Le webcrawler est un programme qui a pour but de scanner un site internet afin d'en extraire les informations dont nous avons besoin. Nous stockons ensuite ces informations dans des fichiers XML en suivant la norme TEI.

2.2. Le corpus CoCoRep

2.2.1. Présentation du corpus

Dans cette partie, nous proposons une description générale du corpus obtenu grâce aux étapes détaillées dans les parties précédentes de ce dossier.

Le corpus obtenu, baptisé **CoCoRep** (pour **Co**rpus **Co**mment **Rep**arer), est un corpus contenant des textes exprimant un problème, extraits du site français commentreparer.com. Le corpus CoCoRep contient 61269 posts d'utilisateurs (5 010 346 tokens au total, dont 511 505 tokens pour les titres uniquement, et 4 498 841 tokens pour les corps de posts uniquement) exprimant un problème, et chaque post est présent dans le corpus sous la forme d'un fichier XML suivant la norme TEI-P5. Il correspond à la totalité des posts présents sur le site commentreparer.com à la date de la récupération des données, le 27 novembre 2022. Les posts sont rangés par catégories et par sous-catégories, lesquelles apparaissent également dans la partie des fichiers contenant les métadonnées. Neuf catégories générales sont présentes dans le corpus, qui correspondent aux neuf catégories du site commentreparer.com. Il s'agit des catégories suivantes, rangées par ordre de numéro de catégorie sur le site :

1) Mobilier, Maison;
2) Electronique, informatique;
3) Electroménager;
4) Auto-moto;
5) Audio-vidéo;
6) Vêtements, linge, bijoux;
7) Jardinage, bricolage;
8) Réparations diverses;
9) Plomberie-Chauffage.

Chaque catégorie contient des sous-catégories permettant de classer de façon plus pertinente les différentes fiches que nous avons obtenues après extraction des informations du site source. Il s'agit, dans la majorité des cas, de sous-catégories correspondant à un type de produit ou d'objet dont il est question dans les fiches. Nous pouvons citer comme exemple des sous-catégories comme « Télévision », « Chaîne hifi », ou encore « Vidéoprojecteur » dans la catégorie « Audio-vidéo ».

Il est très intéressant de noter que, sur le site commentreparer.com, chaque catégorie, chaque sous-catégorie, ainsi que chaque post, possède un identifiant numérique, ce qui nous permet de catégoriser et de stocker plus facilement nos fiches dans les fichiers XML correspondant. En effet, chaque fichier dans notre corpus est nommé selon le motif suivant :
CR-[numéro de sous-catégorie]-[numéro de post]

Dans ces noms de fichiers, la partie « CR » correspond au nom du site : CommentReparer.com. Les parties suivantes correspondent respectivement aux identifiants de sous-catégorie et de post comme présents dans les liens correspondant à ces différentes entités sur le site. Les identifiants de catégorie n'ont qu'un chiffre (de 1 à 9), les identifiants de sous-catégories ont de deux à trois chiffres, et les identifiants de post ont en général cinq chiffres.

Concernant la hiérarchisation des fichiers et des dossiers du corpus, elle imite la hiérarchisation des données sur le site, c'est-à-dire que le corpus contient neuf dossiers, chacun étant associé à une catégorie, et chacun contenant des dossiers correspondant aux sous-catégories. Chacun de ces dossiers de sous-catégorie sont nommés comme suit : **[numéro de sous-catégorie]-[nom de sous-catégorie**]. Cela peut permettre de mieux se repérer lors de l'automatisation d'une analyse en aval. Enfin, chacun de ces sous-dossiers contient les fichiers associés à tous les posts présents sur le site dans la sous-catégorie correspondante.

Dans chaque fichier, en plus du titre du post et de la description plus détaillée du problème postée par l'utilisateur, nous avons également des informations complémentaires sous forme de métadonnées, comme le nom d'utilisateur de l'auteur du post, le lien du post, ou encore un horodatage qui pourra éventuellement permettre d'effectuer une analyse diachronique de l'expression des problèmes.

Il semble également pertinent de mentionner ici que malgré la diversité des problèmes que nous avons abordée au début de cette partie, le corpus contient un nombre largement plus grand de posts dans les catégories « techniques » que dans les autres catégories. Pour reprendre les exemples évoqués précédemment, la catégorie « Electroménager » contient 34711 posts, ce qui correspond à plus de la moitié du corpus, tandis que la catégorie « Mobilier, maison » contient 773 posts. Il semble donc important de le remarquer pour ne pas être surpris lors d'une analyse quantitative qui révèlerait une forte tendance à l'utilisation de mots typiquement associés à des problèmes plus techniques. Voici un tableau détaillant cette répartition de posts par catégorie d'appareils, ainsi que d'autres informations selon cette même distinction, comme le nombre de tokens ou la moyenne de tokens par catégorie.

Catégorie	Nombre de posts	Total tokens dans les titres	Total tokens dans les posts	Moyenne des tokens dans les titres	Moyenne des tokens dans les posts
Electroménager	34711	303033	2685242	8,73	77,36
Audio-vidéo	11095	87478	805171	7,88	72,57
Electronique, informatique	7424	57915	417494	7,8	56,24
Réparations diverses	2567	20925	190154	8,15	74,08
Jardinage, bricolage	2440	18876	177177	7,74	72,61
Plomberie-Chauffage	1341	10652	108897	7,94	81,21
Auto-moto	872	6507	56388	7,46	64,67
Mobilier, Maison	773	5809	55525	7,51	71,83
Vêtements, linge, bijoux	46	310	2793	6,74	60,72
Total	**61269**	**511505**	**4498841**	**8,35**	**73,43**

Tableau 3 - Répartition des posts et tokens par catégorie d'appareil dans le corpus CoCoRep

Ci-dessous, nous présentons la représentation du nombre de posts par longueur en tokens dans tout le corpus, avec une distinction entre les titres et les corps de textes, qui amènent tout de même à un résultat similaire. En effet, nous pouvons voir, en appliquant une échelle logarithmique sur ces fréquences, qu'il y a beaucoup plus de posts contenant peu de tokens, et que plus les posts s'allongent, plus leur fréquence diminue. Nous avons notamment pour les titres, une médiane à 8 tokens pour des titres allant de 1 à 76 tokens, et une médiane à 58 tokens pour les corps de posts allant de 0 à 1348 tokens.

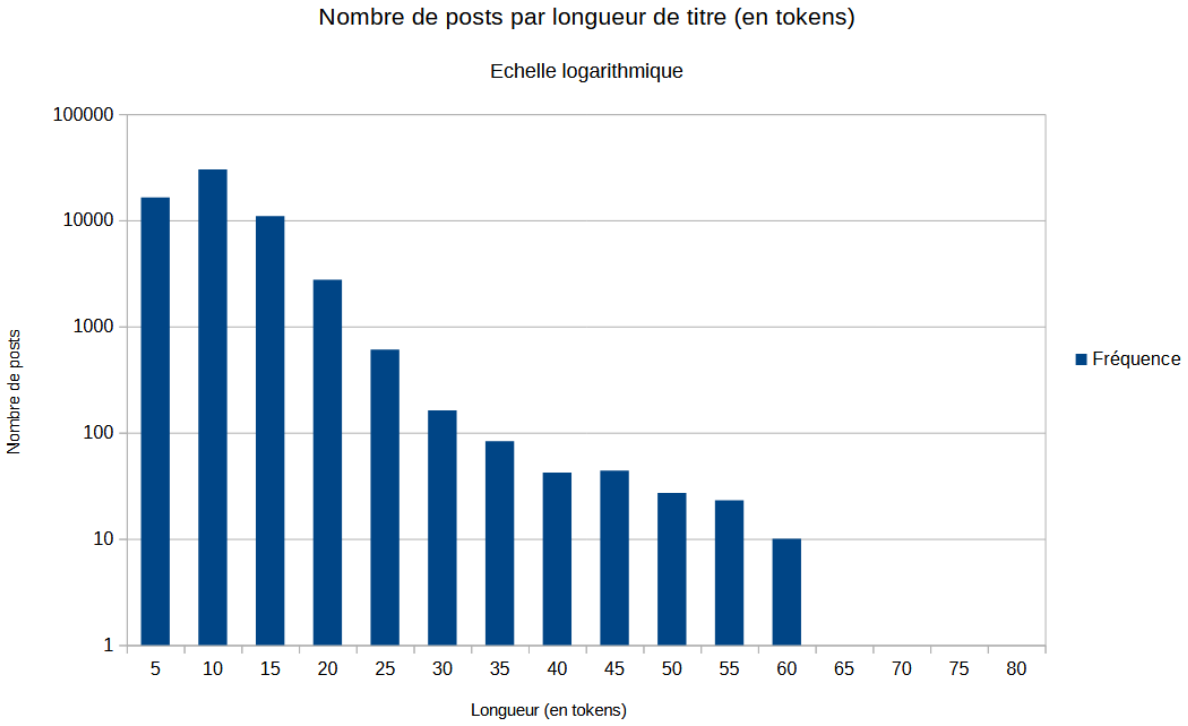

Figure 4 - Diagramme en barre représentant le nombre de posts en fonction de la longueur de leur titre en tokens, avec une échelle logarithmique

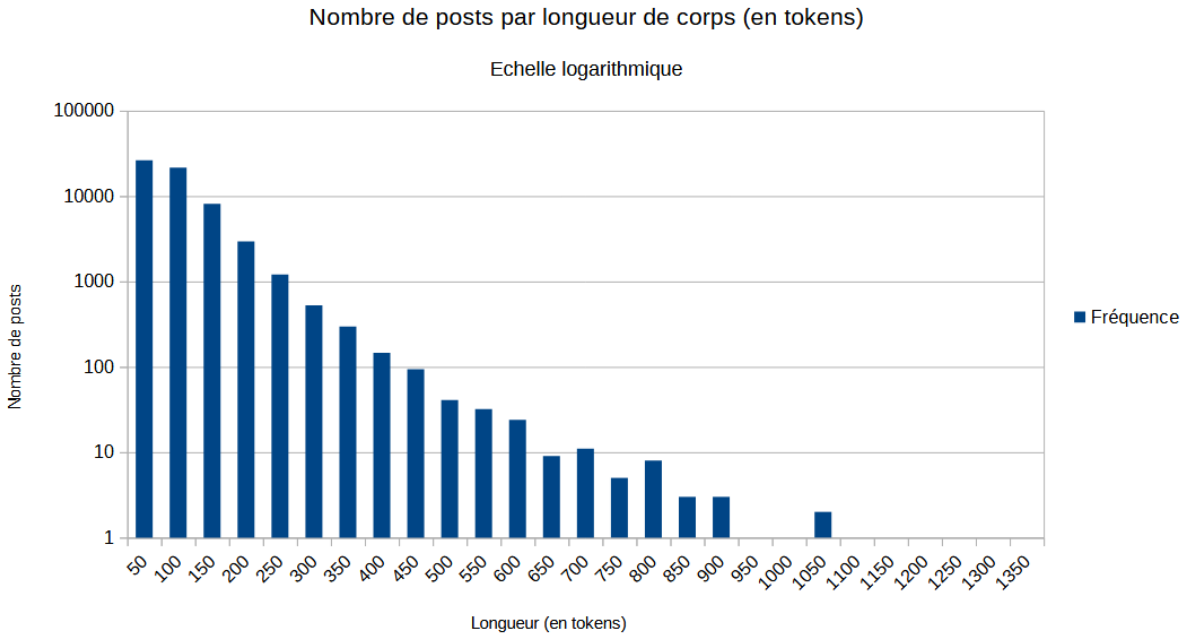

Figure 5 - Diagramme en barre représentant le nombre de posts en fonction de la longueur de leur contenu descriptif en tokens, avec une échelle logarithmique

2.2.2. Statistiques textuelles

Dans cette partie nous utiliserons une combinaison de programmes Python que nous créons, et du logiciel de textométrie TXM, afin d'effectuer des analyses quantitatives un peu plus détaillées susceptibles de nous révéler des tendances présentes dans les titres des messages enregistrés dans le corpus.

TXM est un logiciel qui représente la culmination d'un effort de mutualisation des programmes informatiques de statistiques et de traitement automatique des langues. Ce projet est porté par l'université de Lyon depuis les années 2000 et financé par l'Agence Nationale de la Recherche française (Heiden et al., 2010). Il s'agit d'une plateforme logicielle faisant office de boîte à outils pour l'analyse statistique et linguistique de corpus électroniques.

TXM utilise par défaut le format TEI pour la structuration des données de chaque corpus présent dans la base de données, mais le logiciel met également à disposition un certain nombre de modules d'importation qui permettront à l'utilisateur de charger un corpus sous un format différent, comme un format de texte brut, qui sera ensuite converti par le module d'importation en XML TEI dans un format utilisable par TXM.

Notre corpus est, comme expliqué précédemment, formaté sous la norme TEI-P5, ce qui devrait permettre un import dans TXM sans avoir la nécessité de passer par un module de conversion des données en format TEI. Cependant, le corpus en l'état serait beaucoup trop volumineux pour un chargement et une analyse par TXM. En effet, lors du chargement d'un corpus, TXM utilise TreeTagger afin de tokeniser, lemmatiser, et appliquer des informations comme la catégorie grammaticale aux données, ce qui prend un temps considérable. Ce processus étant inévitable pour une analyse linguistique pertinente des données du corpus, nous choisissons de ne charger que les titres dans un premier temps pour effectuer leur analyse et se rendre compte de leurs régularités. En somme, nous créons une partition du corpus en extrayant le titre de chaque fichier du corpus, afin de tous les concaténer dans un fichier texte, qui sera ensuite importé dans TXM afin d'effectuer les analyses.

Nous commençons par utiliser TXM pour obtenir la table lexicale du corpus constitué des titres de notre corpus CoCoRep, dont nous proposons un extrait ci-dessous.

frlemma	Fréquence
de	16908
@card@	16348
ne	13992
le	12637
plus	11675
réparer	8859
se	7239
qui	6856
mon	6205
un	5882
sur	5010
pas	4858
à	4704
.	4550
lave-linge	4065
allumer	3926
et	3836
linge	3760

Tableau 4 - Table lexicale des lemmes des titres du corpus CoCoRep réalisée avec TXM

Les adverbes « *ne* » et « *plus* » apparaissent très haut dans la liste des lemmes du corpus, donc de façon extrêmement fréquente, juste derrière quelques prépositions et déterminants, ainsi que derrière le lemme « @card@ » qui dénote la présence d'un nombre dans le texte, ce qui correspond souvent à la référence de l'appareil dont il est question, mais peut correspondre également à d'autres informations numériques comme le nombre de fois qu'un signal retentit, ou encore la température de lavage pour un lave-linge ou un lave-vaisselle.

Dans cette même table lexicale, le premier lemme correspondant à un mot lexical est le verbe « *réparer* » qui se trouve également dans le nom du site source. Le lemme « *réparer* » apparaît dans environ 89% des cas au tout début du titre, mais est également souvent précédé de l'adverbe « *comment* », rappelant encore une fois le nom du site source, à hauteur de 6% des occurrences.

En regardant la table lexicale des catégories grammaticales, on se rend compte du fait que la catégorie Nom commun est très représentée, car elle représente 30% de la totalité des occurrences du corpus. Les noms communs sont utilisés dans le corpus pour une variété de fonctions, que ce soit pour la catégorie de l'appareil, la description du problème, ou d'autres informations supplémentaires concernant le problème. Les noms propres, qui sont souvent la marque du produit ou son modèle, constituent la deuxième catégorie la plus fréquente des occurrences à hauteur de 13%.

frpos	Fréquence
NOM	145382
NAM	65266
ADV	37170
PRP	35411
VER:pres	34388
ADJ	32039
VER:infi	19365
DET:ART	18251
NUM	17334
PUN	11657
PRO:PER	11043
ABR	9274
VER:pper	7904
KON	7528
PRO:REL	7512
SENT	7247
DET:POS	7080
PRP:det	4131
VER:simp	2453

Tableau 5 - Table lexicale des catégories grammaticales des titres de CoCoRep réalisée avec TXM

Pour les titres, il est intéressant d'observer qu'un certain nombre de posts ont des titres totalement égaux. Certains de ces titres sont relativement génériques et il est donc attendu de retrouver plusieurs fois le même titre car un problème similaire peut apparaître plusieurs fois et les combinaisons d'informations à mettre dans le titre ainsi que leur taille sont limitées. Nous pouvons donner comme exemple les titres génériques tels que « Réparation », ou un peu plus spécifiquement « Four électrique qui ne chauffe plus ». D'autres titres, néanmoins, très spécifiques dans leur forme et le problème qu'ils expriment, apparaissent plusieurs fois sans que l'on puisse facilement en faire raison. Il s'agit de titres comme « Reprogrammer un portable android qui s'allume mais ne permet pas d'acceder au menu ni afaire des appels », qui malgré les fautes d'orthographe et la spécificité du problème, apparaît 26 fois dans notre corpus. Il ne s'agit d'ailleurs pas de doublons car les auteurs sont tous différents et les corps de post sont également tous différents. Nous mettons à disposition la table de fréquence des 20 premiers groupes de titres ci-dessous, calculés avec un programme Python.

Titre	Fréquence
Four électrique qui ne chauffe plus	51
Comment réparer mon portable qui s'allume plus	38
Réparation	32
Reparation	27
Moyen de passer l'erreur B200 sur imprimante Canon	26
Reprogrammer un portable android qui s'allume mais ne permet pas d'acceder au menu ni afaire des appels	26
Réparer lave-linge	20
Réparer sèche-linge	19
Réparer lave-linge Arthur Martin Electrolux qui s'allume mais ne démarre pas les programmes	17
Réparer lave-vaisselle	17
Lave-linge	15
Lave-vaisselle Brandt erreur D07	15
Réparer imprimante HP Photosmart	14
TV Samsung ne démarre plus	13
Code erreur F8 E1 lave-vaisselle Whirlpool	12
Lave-vaisselle	12
Changer une ampoule grillée de micro ondes	11
Panne	11
Plus d'alimentation sur Sèche linge Electrolux ADC67556W	11
Réparer télévision	11

Tableau 6 -Table de fréquence des titres entiers (20 premiers groupes)

Enfin, il est intéressant de voir que les titres commencent souvent avec un certain mot qui peut donner une indication sur le type de problème, et également dans notre cas sur la catégorie dans laquelle le titre sera classé. Nous donnons ci-dessous la table de fréquence des titres comprenant chaque premier lemme. Notons que, parmi les titres commençant par « panne », « problème », ou le type d'appareil présentant un problème comme « Four » ou « Lave-linge », le lemme « réparer » est de loin le plus présent au début des titres du corpus, avec une fréquence de 7895 qui est largement supérieure au reste. Les auteurs de posts sont souvent dans la recherche d'une façon de réparer l'objet défectueux, ce qui peut expliquer que tant de titres de posts commencent par ce lemme. Notons également que cette catégorie représente 12.89% des titres du corpus.

Premier lemme du titre	Fréquence
réparer	7895
son	2460
problème	1901
Four	1624
comment	1604
panne	1452
Lave-linge	1418
Machine	1339
Lave	1266
Lave-vaisselle	1226
le	1151
réparation	990
TV	809
changer	688
Tv	685
sec	634
reparer	525
code	503
lave	462
Probleme	438

Tableau 7 - Table de fréquence des premiers lemmes pour les titres (20 premiers groupes)

Il nous est également possible d'observer les spécificités de chaque catégorie de messages grâce à l'outil « Spécificités » de TXM appliqué à une partition de notre corpus de titres par catégories suivant les catégories présentes sur le site source. En observant les spécificités des occurrences taguées comme Nom commun par TreeTagger, il apparaît que chaque catégorie présente une surreprésentation de noms d'objet ou d'appareil en rapport avec cette catégorie. Nous pouvons par exemple citer « Lave-linge » qui est le lemme le plus surreprésenté dans la catégorie « Électroménager », ou bien « Canapé » qui est le lemme le plus surreprésenté dans la catégorie « Mobilier, maison ».

Si l'on observe les spécificités des verbes selon la même partition par catégories, on peut faire une observation similaire quant à la répartition des verbes. En effet, les lemmes les plus surreprésentés dans chaque catégorie correspondent à des verbes ayant des liens lexicaux avec la catégorie, mais nous pouvons plus précisément voir deux types de verbes, à savoir ceux qui qualifient ce que l'appareil doit faire d'un côté (exemple : « laver » pour « Electroménager »), et de l'autre côté les verbes qui qualifient l'action nécessaire ou entreprise par l'utilisateur afin de résoudre le problème ou de réparer l'appareil (« reprogrammer » pour « Electronique, informatique »).

2.2.3. Repérage d'informations dans les titres et de structures

Les titres du corpus CoCoRep, de par leur contexte d'expression et leur format court, présentent des similarités que l'on peut regrouper en catégories de titres en fonction des patrons d'expression et des motifs lexicaux et syntaxiques que l'on y retrouve. En effet, même si les titres peuvent être écrits de façon totalement libre et ne sont influencés que par des suggestions de rédaction lors de la création d'un message sur le site source, nous pouvons apercevoir une régularité remarquable dans la rédaction des titres, et ce peu importe la catégorie et la sous-catégorie. Malgré les différences lexicales que nous retrouvons forcément entre les différentes catégories dûes aux spécificités de chaque catégorie, nous nous efforçons ici de proposer des catégories d'expression relativement abstraites qui peuvent s'appliquer à toutes les catégories d'appareil et à tous les domaines présents dans le corpus.

Afin d'effectuer un repérage préliminaire des différentes catégories de titres que nous détaillons ci-dessous, nous observons un échantillon de titres et tâchons de repérer les différentes informations présentes dans les titres. Ces informations semblent s'agencer de quelques façons distinctes qui pourraient donc constituer des catégories d'expression d'un problème. Les informations présentes sont de l'ordre du nom de l'appareil, la marque de l'appareil ou encore les verbes utilisés entre autres. Nous interprétons donc chaque titre de notre échantillon en repérant les types plus abstraits d'information auxquels chaque occurrence correspond, ce qui nous permet de dégager 8 catégories préliminaires parmi les titres.

Les 8 catégories de titres repérées lors de ce premier survol des données sont les suivantes : le nom de l'appareil ou de l'objet seul, le nom de l'appareil ou objet accompagné du composant défectueux ou problématique, le nom de l'appareil et du composant défectueux accompagnés d'une description du problème rencontré, les informations de l'appareil accompagnées de l'indication d'un signal, une indication d'un problème sur un appareil sans décrire le problème, une demande explicite d'aide pour la réalisation d'une tâche contribuant à la résolution du problème, une demande moins explicite d'aide à la réparation, et une catégorie comprenant les titres « expressifs » avec un style narratif s'éloignant d'un simple résumé du problème.

1) Objet seul :

Tout d'abord, nous pouvons trouver un nombre conséquent de titres contenant le nom de l'appareil ou de l'objet posant problème, sans mention du problème posé. Un exemple de ce type de titre peut-être : « Balance Tanita TBF 300 ». Parfois, la référence n'est pas non plus disponible dans le titre : « Balance Terraillon ». Nous pouvons émettre l'hypothèse, pour les posts dont le titre correspond à ce genre de construction, que le problème sera explicité dans la partie descriptive du corps du message et que l'auteur n'a donc pas jugé pertinent de faire allusion au problème à nouveau dans le titre. Dans notre corpus, ce type de titre ne nous donne pas beaucoup d'informations puisque le nom de l'appareil, ainsi que sa référence, sont également accessibles dans les métadonnées de chaque post, ce qui rend leur mention dans le titre redondante du point de vue de l'extraction d'information.

2) Objet et composant :

De façon similaire, beaucoup de titres comprennent les mêmes informations (nom et référence de l'appareil), mais y ajoutent le nom du composant ou de la partie de l'appareil ou de l'objet qui présente un écart à la norme ou une défaillance. « Tige gonfleur pneu arrière trotinette e300 » est un exemple de titre correspondant à ce type. Ce type de titre ne nous indique pas la nature du problème, mais il nous donne tout de même plus d'informations que le précédent type car nous avons ici un indice sur la localisation précise du problème, ce qui peut permettre de circonscrire le problème à un certain nombre de problèmes correspondant uniquement à cette partie de l'objet ou de l'appareil.

3) Description brève du problème :

Le type de titre suivant contient le plus d'informations nécessaires à la compréhension du problème tout en ayant l'avantage d'être concis. Il s'agit de l'expression du nom de l'appareil, suivie

éventuellement de précisions comme la partie dysfonctionnelle, suivie d'une description très brève du problème rencontré, en un ou quelques mots. L'ordre des mots n'est évidemment pas toujours le même, ce qui peut amener à des difficultés de repérage d'entités par la suite. Néanmoins, ce type de titre nous donne assez d'informations pour pouvoir aborder le problème et pour les autres utilisateurs du site source, il permet de savoir si la résolution de ce problème se trouve dans leurs capacités. Dans ce type de titre, le problème peut être exprimé de plusieurs façons. Il arrive souvent que le problème soit exprimé sous la forme négative « *ne* + verbe + *plus* » ou associés à d'autres adverbes de négation comme « *pas* » (exemple : « Samsung ue40d5003bw ne sallume plus », « TV Led Continental Edison ne démarre pas »). Il est intéressant de noter une différence dans ces verbes quant à leur niveau de spécificité de description du problème. En effet, les verbes utilisés peuvent être plutôt génériques (exemple : « Philips 974 mark 2 ne fonctionne plus ») ou plus spécifiques sur la fonctionnalité qui fait défaut (exemple : « Bras de lavage de lave-vaisselle Bosch ne tourne plus »). Le problème peut également être exprimé avec un verbe à la forme affirmative, ou être accompagné du pronom relatif « *qui* » comme dans les exemples : « Téléviseur qui ronfle », « écran telé qui devient jaune », ou encore « Tv qui ne s'allume plus ». Enfin, le problème peut être exprimé au moyen d'un nom déverbal à la place d'un verbe, ou encore d'un adjectif ou d'un participe passé correspondant à son état après quelconque dysfonctionnement par exemple (exemple : « Blocage du tambour », « Fuite bac à lessive candy smart 10kg cs 13102d3/47 »).

4) Information machine :

Certains titres contiennent, en plus d'un potentiel nom d'appareil ou d'autres informations sur le problème, une indication de la machine permettant d'en savoir un peu plus sur le problème. En effet, dans le cas de certaines machines, un système intégré informe l'utilisateur de la présence d'un problème et éventuellement de la nature de celui-ci via une alarme, une lumière d'une couleur spécifique, ou encore un code d'erreur affiché sur un système d'affichage. Dans ce type de titres, l'information donnée sur le problème n'est pas directement accessible, car il y a une nécessité de savoir à quoi correspond le code d'erreur ou le son qui retentit, ce qui est renseigné, en règle générale, dans le manuel d'utilisation de l'appareil. Néanmoins, si les informations concernant la raison d'apparition de ce signal est disponible, ou si l'accès au manuel d'utilisation est possible, le signal mentionné peut nous donner des informations très précises sur la nature du problème et les solutions possibles. Certains exemples comprennent « Code erreur E8, lave-linge Daewoo DWD UD1213 » ou encore « Lave vaisselle rosieres RLF912E intégré bip en continue. ».

5) Titre générique :

Sachant que, sur le site source, le message entier devra être lu pour comprendre le problème, certains utilisateurs ne se contentent que d'un titre générique qui ne permet pas d'avoir d'informations précises sur la nature du problème. Ces titres ne sont parfois constitués que d'un mot, comme « problème » ou « panne », ou encore « réparation », mais il arrive également que l'utilisateur choisisse de mentionner tout de même la catégorie de l'appareil (exemple : « Pb lave vaisselle bosch », « Réparer TV NASCO »).

6) Aide dans le processus de réparation :

Au contraire, certains utilisateurs ne se contentent pas de décrire leur problème, mais ont déjà entamé une démarche de compréhension du problème et cherchent de l'aide dans le but d'effectuer une ou une série de tâches qui leur permettrait d'aboutir à la résolution du problème. Le problème n'est pas directement mentionné dans ce genre de cas, car il s'agit plus d'une demande d'aide ou d'une question, plutôt que d'une simple déclaration du problème. Nous pouvons mentionner des titres comme « Comment atteindre la résistance de mon four Whirlpool AKZM 770 ? » ou bien « Démontage pompe de relevage sèche linge Brandt bwd381T ».

7) Demande d'aide à la réparation :

D'autres utilisateurs posent des questions moins précises en termes d'informations contenues dans le titre. Ce type de titre peut commencer par un verbe comme « *réparer* », précédé ou non de l'adverbe interrogatif « *comment* », à l'instar du type de titre précédent, mais à l'inverse de celui-ci, ce type ne donne pas plus d'information que le nom de l'objet à réparer, et éventuellement sa référence,

ce qui l'apparente à la première catégorie de titres mentionnés dans cette partie. Quelques exemples seraient : « Réparer mon sèche-linge hotpoint Ariston TCDG51XB », « Réparer un lave-linge electrolux awf 14480 w », « Comment réparer lave-linge Ariston WMD 942 message F08 ». Ce dernier exemple nous permet de rappeler que les types proposés ne sont pas figés, et que certaines informations peuvent s'ajouter de façon libre dans le titre, comme c'est le cas pour le message d'erreur ici.

8) Description narratives ou autres informations :

Enfin, une partie des utilisateurs utilisent le champ « Titre », lors de la création de leur message, pour rédiger une description détaillée de leur problème, comme on s'attend à retrouver dans le corps du texte. Ces titres narratifs, bien qu'ils ne servent pas leur fonction attendue de résumé du problème, présentent en général beaucoup plus d'informations et de subtilités que les autres titres, de par sa longueur et son aspect narratif. Ceux-ci présentent des caractéristiques que nous retrouverons dans la partie suivante lors de l'analyse des descriptions détaillées des problèmes, comme les salutations « Bonjour » au début du texte, ou encore les questions. Nous citerons un exemple ici : « Bonjour l'appareil photo de mon téléphone tecno spark k7 s'est cassé et depuis lors le selfie ne marche plus solution svp ». Il peut également y avoir une présence de ces aspects uniquement dans le titre, qui ne sera alors constitué que de salutations ou autre, mais ne contiendra aucune information sur le problème (exemple : « S'il vous plaît ! »).

Parmi ces catégories de titre, nous retrouvons souvent des informations similaires placées dans un ordre quelconque, qui nous permettent de les placer dans une catégorie de titre en particulier. Ces éléments récurrents nous permettront ensuite d'effectuer une catégorisation des titres sur la base des éléments qui se trouvent, ou non, dans les titres. Les éléments en question sont des types suivants :
- Type d'appareil (exemple : « lave-vaisselle »)
- Marque de l'appareil (exemple : « Candy »)
- Modèle de l'appareil (exemple : « CDP 3560 »)
- Composant (exemple : « électrovanne »)
- Code ou signal d'erreur (exemple : « Erreur f03 »)
- Fonctionnalité dégradée (exemple : « ne tourne plus »)
- Comportement indésirable ou inattendu (exemple : « s'allume »)
- Action à réaliser (exemple : « réparer »)
- Marque interrogative (exemple : « comment »)
- Cause du problème (exemple : « l'appareil photo s'est cassé »)
- Action entreprise dans le but de résoudre le problème (exemple : « j'ai démonté toutes les vis qu'il y a en dessous »)
- Demande d'aide (exemple : « solution svp »)
- Marqueur de politesse (exemple : « Bonjour »)
- Marqueur expressif (exemple : « URGENT »)

A partir des informations relevées dans les titres et des structures repérées, nous proposons un rapprochement de ces catégories à la typologie de Mariame MAAROUF. La catégorie 3 correspond directement aux catégories correspondantes dans la typologie puisqu'il s'agit d'une description du problème, qui peut contenir des marqueurs de Fuite, Dispositif qui ne fonctionne pas, Hors spécification etc. La catégorie 4 correspond directement à la catégorie Signal de la typologie de base. La catégorie 5, correspondant à un titre générique, correspondra également à la catégorie représentée par ses marqueurs. Si le titre correspond à « Problème », il sera catégorisé comme « Hors spécification », alors que « Panne » sera catégorisé comme « Dispositif qui ne fonctionne pas ».

Les autres catégories sont spécifiques à notre corpus, ce qui nous amène à envisager la création de nouvelles catégories d'expressions qui correspondent à nos données. Les catégories 1 et 2, décrivent toutes deux un objet ou un appareil, sans donner d'informations supplémentaires sur le type de problème, ce qui pourra justifier la création d'une catégorie pour les annonces d'objet. Une expression d'une action faisant partie des fonctionnalités normales de l'objet ou de l'appareil ne

pouvant pas être classée dans les autres catégories, et donc correspondra à État du monde ou à sa propre catégorie différente de la catégorie pour les objets uniquement.

L'information d'une action à réaliser ou d'une question, dans les catégories 6 et 7, mais également partiellement dans la catégorie 5, qui n'apparaissent pas dans les données des FA du domaine spatial par leur contexte et leur visée, peuvent également donner lieu à une nouvelle catégorie concernant la demande d'aide.

Enfin les descriptions narratives de problème de la catégorie 8 peuvent correspondre à leur catégorie respective dans la typologie. Cependant, la présence d'un marqueur de politesse ou autre information n'étant pas liée au problème devra être classée séparément et donnera lieu à une nouvelle catégorie d'expression.

2.2.4. Aperçu des descriptions détaillées

L'analyse des descriptions détaillées des problèmes s'avère être beaucoup plus compliquée que l'analyse des simples titres. En effet, les descriptions présentent énormément de variations tant sur la longueur que sur les structures employés, et même sur le plan du déroulement discursif. Pour analyser ce genre de textes de façon pertinente, il paraît indispensable de recourir à des outils de TAL plus complexes et à des techniques statistiques.

Cependant, un aperçu des descriptions détaillées n'est pas complètement inutile, car les textes du corpus, en dépit de toutes leurs différences, présentent quelques similarités que nous mentionnerons dans cette partie, sans entrer dans une analyse détaillée de ces similarités, qui pourra faire l'objet d'une ouverture à ce travail de recherche.

La première similarité immédiatement visible entre une grande majorité de messages est que ceux-ci commencent de façon quasi systématique par « Bonjour, ». Au-delà du souci de politesse de la part des auteurs que l'on pourrait supposer en voyant ces données, cette présence quasi systématique d'un « Bonjour, » en début de message est facilement expliquée par la façon dont la rédaction d'un message se fait sur le site source. En effet, lors de la création d'un post sur commentreparer.com, l'encart destiné à la rédaction de la description détaillée du problème est déjà pré-rempli avec la chaîne de caractères « Bonjour, ». Il est possible d'effacer cette chaîne de caractères, mais la raison évoquée précédemment de soucis de politesse incite probablement les auteurs de messages à ne pas supprimer cette chaîne.

Les différents messages varient en longueur en fonction de la nature du problème, des éléments à disposition de l'auteur lors de la rédaction du message, mais également du style narratif de l'auteur du message. Cependant, un grand nombre de messages ont l'air de suivre un schéma discursif similaire qui peut se résumer de la façon suivante : après un éventuel rappel succinct du problème, l'auteur propose un contexte dans lequel il retrace les événements menant à l'occurrence du problème. Après cette mise en contexte, le problème est explicité et un certain nombre de détails sont éventuellement fournis afin de permettre aux lecteurs de comprendre plus précisément le problème. Enfin, l'auteur explique éventuellement ce qu'il a déjà entrepris afin de remédier au problème, et il termine par l'énonciation d'une question comme « Que dois-je faire pour réparer mon écran ? » ou plus vague comme « Pouvez-vous m'aider s'il vous plaît ? ».

Il est intéressant de noter qu'un grand nombre de messages se termine également par une marque de politesse comme un simple « merci », ou une formule de politesse un peu plus longue comme « Je vous remercie d'avance pour votre réponse ». Cette façon de terminer le message n'est pas sans rappeler la rédaction d'un email, si l'on reste dans cadre des CMR, ou même d'une lettre si on s'éloigne de ce cadre. Il peut sembler intéressant de faire des liens, dans une version ultérieure de ce dossier de recherche, entre ce nouveau type de communication et des schémas de communication déjà établis dans des contextes plus longuement établis.

Pour finir, nous rappelons qu'il conviendra d'utiliser des techniques plus poussées afin d'avoir un aperçu plus global et plus complet des données et effectuer les analyses plus pertinentes sur les schémas discursifs entre autres. Néanmoins, ces analyses semblent être trop complexes et relèvent d'un niveau supérieur de traitement automatique des langues, car les analyses devraient se faire sur les plans discursifs et pragmatiques, ce qui semble compliqué dans le cadre du projet de recherche pour ce mémoire. Nous nous contenterons donc d'analyser dans le reste de ce mémoire les titres seuls des posts plutôt que leurs descriptions détaillées. Ces textes plus longs et plus complexes pourront faire l'objet d'autres études sur un sujet similaire.

Dans cette deuxième partie, nous avons utilisé une panoplie variée d'outils et de techniques afin de constituer un corpus et d'avoir un aperçu global des données que nous utiliserons ensuite. Cette observation des données nous a permis de dégager des composants présents dans les titres du corpus que nous avons tâché de regrouper en catégories. Ces catégories deviendront essentielles dans la construction d'une typologie adaptée dans le cadre de notre problématique portant sur la classification de l'expression des problèmes.

3. Les catégories d'expression d'un problème

Dans cette troisième partie, nous formalisons les types d'expressions présents dans notre corpus, et adaptons la typologie utilisée dans le domaine spatial en prenant pour appui les composants principalement présents dans les titres de nos données, dégagés dans la partie précédente. Cette nouvelle typologie donne ensuite lieu à un guide d'annotation qui permet de spécifier la définition des nouvelles catégories, ainsi que la distinction entre celles-ci. La guide d'annotation est utilisé afin de déléguer l'annotation à un annotateur externe lors de la phase de création d'un système automatique de classification des titres en fonction de notre typologie. Nous décrivons donc également le processus de création et d'évaluation du guide et de l'annotateur choisi afin de s'assurer d'une représentativité de l'annotation assez forte pour donner des résultats fiables lors de la classification automatique.

3.1. Typologie de l'expression d'un problème dans les CMR

3.1.1. L'expression d'un problème dans les CMR et ses spécificités

Dans la première partie de ce mémoire, nous avons abordé l'expression du problème technique dans les Fiches d'Anomalie dans le cadre du processus de Retour d'Expérience. Cependant, même dans un cadre global similaire d'expression du problème technique, il est important de noter les différences entre le Retour d'Expérience dans un cadre industriel et l'expression d'un problème dans le cadre de la CMR. En effet, chaque cadre et chaque contexte amène avec lui certaines spécificités dans les types d'expression, le but de l'expression du problème, mais également des spécificités linguistiques sur la forme utilisée, le registre de langue, les habitudes langagières, en fonction notamment du public visé par ces différentes instances d'expression d'un problème.

Alors que l'expression du problème dans le cadre du Retour d'Expérience se place généralement dans une démarche de recensement des problèmes dans le but d'améliorer les services, de réduire certains coûts, ou encore de renforcer la sécurité des travailleurs sur le terrain, l'expression du problème dans la CMR est souvent destinée à des utilisateurs auxquels on demande de l'aide. Le Retour d'Expérience industriel se place donc dans une démarche plutôt informative qui pourra amener une analyse *a posteriori* des problèmes via des techniques de traitement automatique des langues, là où l'expression du problème dans la CMR se place dans une démarche active de recherche d'aide et de solutions à un problème. Cette différence dans le but de la démarche amène une différence de typologie du problème ainsi qu'une multitude de différences sur plusieurs plans linguistiques qu'il sera important de prendre en considération.

La demande d'aide par rapport à un problème dans le cadre de la CMR s'inscrit donc dans une démarche différente du recensement d'un problème dans les Fiches d'Anomalie que l'on pourrait par exemple retrouver dans le domaine spatial. La CMR amène notamment un nouvel environnement sociolinguistique dans lequel l'expression du problème se place dans un contexte particulier qui est celui de la demande. Ce contexte particulier se reflète dans la présence de marques de politesse (Amato & Boutin, 2013), mais également dans d'autres marqueurs comme la présence de formes verbales spécifiques à la demande plus ou moins explicite, comme l'impératif ou l'infinitif.

Cette particularité de démarche dans la CMR amène également une spécificité du contenu des fiches d'expression d'un problème, car au delà de la différence de type d'appareils et donc de types de problèmes rencontrés d'un côté par des particuliers et de l'autre par des professionnels de l'industrie dans différents domaines, d'autres différences sur le type de problèmes exprimés peuvent apparaître. Une différence clairement identifiable liée à la différence de démarche se trouve dans le fait que les utilisateurs d'une plateforme de CMR, comme un forum d'aide, ne rédigent pas de fiche exprimant un problème s'ils ne cherchent pas d'aide et s'ils n'ont pas besoin d'aide pour effectuer un action ou résoudre un problème. Cela constitue une différence dans le type de problèmes exprimés car dans le

REX, tout écart à la norme, même minime et même facilement réparable, sera exprimé, la différence étant que l'auteur d'une Fiche d'Anomalie portant sur un problème facilement résolu ou dont la solution est claire n'aura pas besoin d'aide extérieure pour résoudre le problème, mais l'écart à la norme amènera quand même à la rédaction d'une FA, ce qui n'est pas le cas sur des forums de demande d'aide.

Il peut également y avoir des différences dans la forme de l'expression d'un problème dans la CMR en fonction de la plateforme utilisée pour formuler son problème, car certaines plateformes, certains sites, ou certains forums, demanderont d'expliciter certaines choses qui ne seront pas demandées par d'autres, et des exemples sur la façon d'exprimer un problème ou des emplacements particuliers pour indiquer certaines informations sur le problème changeront le type et la structure de l'expression du problème. Dans une fiche ou un post sur un forum, certaines informations peuvent être implicites par rapport au contexte dans lequel cette fiche ou post a été rédigée, comme le site sur lequel le post est rédigé ou la catégorie dans laquelle la fiche exprimant un problème est publiée. Les métadonnées associées au post peuvent également donner des informations supplémentaires au lecteur et le guider dans sa compréhension du problème, comme la marque ou le modèle de l'appareil défectueux. Ces informations présentes dans le contexte entourant le texte exprimant un problème sont à risque de ne pas être exprimées dans le texte lui-même, par principe d'économie, ce qui change la forme de l'expression et peut notamment poser certains problèmes lors de l'analyse et du traitement automatique des données textuelles. Le principe d'économie peut notamment amener à l'omission de verbes de demande d'aide dans les posts si la demande est implicitement comprise, à l'omission du nom de l'appareil ou du type d'appareil si le post est présent dans une catégorie dédiée à ce type d'appareil ou même s'il est publié sur un forum consacré uniquement à ce type d'appareil, ou encore à l'omission d'autres informations sur l'état de la machine par exemple, qui peuvent être présentes dans d'autres champs du post ou dans les commentaires suivants dans le cadre d'un fil de discussion, ce qui est courant dans des forums.

Dans la forme de l'expression dans la CMR, il est assez courant que les textes contiennent des abréviations, du langage SMS ou un style télégraphique (Chanier et al., 2014), mais cette particularité de la CMR n'est pas forcément spécifique à celle-ci. En effet, dans le cadre du REX industriel, il est également possible de voir un principe d'économie similaire en action par un style télégraphique et des abréviations. Des abréviations codifiées comme on peut en retrouver dans les exemples de FA du domaine de l'aviation, dans les exemples de fiches de l'ASRS où « FLT » correspondra à « flight » ou « vol » en français (Tanguy et al., 2016), sont également possibles. La spécificité des abréviations et de ces expressions économiques dans la CMR vient cependant de sa variabilité et de la difficulté à prédire la correspondance entre une forme abrégée et son correspondant entier dans le cadre du traitement automatique. Ces abréviations ne sont pas codifiées et peuvent signifier plusieurs choses différentes, la signification de l'abréviation pouvant être demandée par un autre utilisateur et clarifiée plus loin dans le fil de discussion dans le cadre d'un forum par exemple, information à laquelle nous n'avons pas accès lors de l'analyse de la fiche seule.

Au-delà des abréviations et des omissions dues à un style télégraphique, les variations orthographiques sont une autre spécificité des CMR, qui peuvent représenter un obstacle assez important au traitement automatique. En effet, des variations orthographiques non prévues, effectuées sur une base phonétique (« casser » au lieu de « cassé ») peuvent poser problème lors de l'analyse d'un texte par un système entraîné sur une langue homogène sans variations de ce type. Ces variations orthographiques peuvent également rendre la compréhension humaine difficile, car les lecteurs humains peuvent également avoir des difficultés à la compréhension de certaines informations sur le problème rencontré par l'auteur du post si l'expression du problème contient des variations tellement grandes qu'elles ne permettent plus assez facilement de faire correspondre la forme exprimée avec la forme normalisée du mot (c'est le cas par exemple dans nos données avec « plus aces » qui signifie « plus d'accès », ce qui est compris uniquement grâce à d'autres champs et aux métadonnées).

La terminologie utilisée dans la CMR, comme sur des forums de demande d'aide, montre également des différences avec la terminologie utilisée dans le cadre du Retour d'Expérience. En

effet, le contexte professionnel du Retour d'Expérience dans une entreprise et les enjeux du REX conduisent naturellement à une terminologie plus spécifique qui permettra de cibler le problème et de le décrire de façon précise et concise. Dans la CMR, et dans le cadre de la demande d'aide sur les forums par rapport à un problème technique, les auteurs sont souvent des personnes non expertes dans le domaine de l'appareil défaillant, et de ce fait utilisent une terminologie plus vague pour décrire leur problème. Cette terminologie plus vague peut amener notamment à une interprétation différente selon le lecteur, et un lecteur expert dans le domaine en question pourrait avoir du mal à aiguiller l'auteur du post sur le problème et la démarche à suivre pour le résoudre.

Comme il a été mentionné plus haut, un problème exprimé sur un forum d'aide peut être rédigé dans un formulaire contenant plusieurs champs distincts pour plusieurs informations distinctes, ce qui peut amener à une économie de l'expression caractéristique de la langue. Dans ce mémoire, nous nous concentrons sur un champ en particulier, qui est celui du titre du post. En effet, le titre du post sur un forum de demande d'aide à la réparation constitue un parallèle intéressant aux expressions de problème rencontrées dans les FA et notamment dans les données sur lesquelles sont basées la typologie dont nous nous inspirons dans ce projet. Le titre d'un post est une façon compacte et concise d'exprimer un problème, et c'est donc là que les informations importantes sur la nature du problème seront exprimées. Néanmoins, il est important de prendre en compte le fait que certaines informations peuvent être omises car le format encourage l'économie de l'expression. C'est cependant le champ le plus intéressant à étudier dans un contexte de typologie d'expression d'un problème du fait de sa simplicité attendue dans l'expression qui amène les auteurs de posts à formuler de façon concise leur problème.

3.1.2. Adaptation de la typologie de base et création de nouvelles catégories

Le projet sur lequel nous travaillons se base sur un travail de typologie de l'expression des problèmes dans le domaine spatial. Dans le cadre de ce travail sur lequel nous nous basons, une typologie de l'expression des problèmes a été imaginée pour correspondre aux types d'expressions d'un problème retrouvés dans les Fiches d'Anomalie autour de lanceurs Ariane (voir partie 1.1.1). Puisque nous avons choisi nos données en partie pour leur similarité avec ces Fiches d'Anomalie, nous nous basons également sur la typologie utilisée pour ces Fiches d'Anomalie dans notre présente étude. Les problèmes évoqués peuvent être effectivement exprimés de façon similaire, malgré les spécificités de nos données, qui ont été détaillées notamment dans la partie précédente.

Il serait néanmoins inadéquat d'appliquer directement la typologie développée pour les données des Fiches d'Anomalie à nos données de forum d'aide à la réparation sans adapter cette typologie à nos données et à leurs spécificités. En effet, malgré des ressemblances et des similarités entre les deux contextes d'expression du problème, la démarche dans laquelle ils s'inscrivent, les acteurs qui y jouent un rôle, le mode d'expression, ou encore le but recherché par l'action de l'expression du problème diffèrent assez pour justifier un aménagement de la typologie à des types d'expression qui peuvent exister dans un contexte mais pas dans l'autre, et vice-versa.

Parmi les catégories présentes dans la typologie de base, beaucoup de catégories peuvent correspondre à nos données. Les catégories Fuite ou encore Signal sont notamment très présentes dans nos données malgré leur caractère spécifique à un certain type de contexte ou un certain type d'appareil. Certaines catégories sont néanmoins superflues, car elles ne reflètent pas de réalité présente dans nos données permettant de justifier leur conservation. C'est le cas de la catégorie Obstacle, qui comme elle est décrite dans la typologie de base, se rapporte à un problème où un composant en bloque un autre, où une supposée synergie de plusieurs objets ou appareils, ou encore composants, est entravée par le positionnement ou le dysfonctionnement d'un ou plusieurs des composants. Ce type de problème n'existe pas dans nos données pour plusieurs raisons. Une des raisons apparaît dans le contexte physique du dysfonctionnement, car là où dans un contexte

professionnel, le problème peut apparaître dans une usine où plusieurs appareils coexistent, un particulier demandant de l'aide pour réparer un appareil chez lui n'aura pour problème que celui qui apparaît sur l'appareil en question, et non pas par rapport à une multitude d'appareil, parmi lesquels la position d'un constituerait un obstacle à un autre. Si ce cas se produisait, la personne en charge de l'appareil pourrait facilement déplacer l'appareil en question sans rédiger de post dans le cadre de la CMR, alors que dans le cadre du REX, la rédaction d'une fiche serait justifiée.

La catégorie EtatduMonde présente dans la typologie de base a également un rôle spécial qui ne correspond pas aux données que nous avons récoltées sur le forum d'aide à la réparation. En effet, la catégorie EtatduMonde est une catégorie dans laquelle est exprimé sur la fiche un écart à la norme facilement résolu. Comme nous l'avons indiqué précédemment, ce type de fiches apparaît dans le contexte de REX car tout écart à la norme doit être recensé. Dans le contexte d'une demande d'aide sur un forum, un écart à la norme facilement résolu et ne posant pas de problème particulier ne constitue en général pas une raison qui pousserait un utilisateur à rédiger un post demandant de l'aide. Un écart comme « la porte du lanceur est ouverte » est justifié dans un REX industriel, mais dans le cas d'un problème similaire chez un utilisateur potentiel d'un forum d'aide, il sera plus facile pour l'utilisateur de fermer la porte lui-même, ou de choisir de la laisser ouverte, sans rédiger de fiche demandant de l'aide.

Nous apercevons tout de même dans nos données deux types d'expression du problème qui n'appartiennent pas aux autres types et qui tout de même correspondent à un problème sans l'exprimer. Ces catégories peuvent être considérées comme découlant de la catégorie EtatDuMonde dans leur absence de l'expression de quelconque problème potentiellement réparable. Une de ces deux catégories est la catégorie Objet, dans laquelle s'inscrivent les expressions ne contenant que le nom de l'appareil supposément dysfonctionnel ou présentant un écart à la norme, accompagné de la marque ou d'un composant de l'objet, sans expliciter le problème rencontré, ni même qu'un problème est survenu. Cette catégorie correspond à des titres comme « Ecran TV », « Lave Linge LG » ou encore « Mon clavier d'ordinateur ». Dans notre repérage des informations contenues dans les titres (voir partie 2.2.3), il nous est apparu clairement que la mention de l'objet avait un rôle proéminent dans l'expression du problème dans nos données.

L'autre catégorie est la catégorie Action, correspondant aux expressions contenant une action effectuée par l'appareil et qui ne correspond pas forcément à l'expression d'un écart à la norme, car l'action exprimée correspond aux actions normales de l'appareil ou l'objet en question. Malgré le fait que l'on puisse imaginer que l'action décrite est soit perçue comme anormale par l'auteur du post, ou qu'au contraire l'auteur demande de l'aide dans le but d'effectuer cette action, les informations correspondant à ces types d'expression ne sont pas présentes dans le titre du post. Un titre comme « Mise à jour android » peut correspondre à cette catégorie, car la mise à jour est une action normale effectuée par un système Android.

Avec les spécificités de la CMR que nous avons évoqué précédemment, un titre de post étant normalement destiné à l'expression concise d'un problème peut être utilisé également pour quelque chose de différent. Un utilisateur peut choisir par exemple de s'adresser aux utilisateurs qui verront le titre, en essayant de les encourager à cliquer sur le lien avec un « À l'aide » ou un « Urgent » afin de voir le reste du post et des informations concernant le problème rencontré. Les formules de politesse seules, qui sont normalement, en contexte en tout cas, typiques des CMR, peuvent également apparaître dans un titre. Afin de pouvoir faire correspondre une catégorie à ces occurrences et ne pas les ignorer, nous créons une catégorie Autre, à laquelle appartiennent tous les titres ne correspondant à aucune autre catégorie d'expression d'un problème, et ne contenant pas non plus d'information sur l'action entreprise ou à entreprendre, ou sur l'objet ou l'appareil défaillant. Cette catégorie Autre est donc représentée par toutes les informations n'entrant pas dans la spécification d'un problème, comme les marqueurs de politesse et les marqueurs expressifs que nous avons repérés dans la partie 2.2.3.

Cependant, la spécificité la plus flagrante de nos données est le fait que les utilisateurs expriment leur problème dans le but d'obtenir de l'aide. Cette demande d'aide est supposée dans tous

les posts mais celle-ci peut être explicitée grâce à des verbes indiquant la réparation envisagée par l'utilisateur, ou encore sous forme de question qui, une fois répondue, permettra d'accomplir le but du post, à savoir la résolution du problème. Il s'agit des catégories 6 et 7 comme décrites dans la partie 2.2.3. La demande d'aide est souvent accompagnée d'une indication sur la nature du problème à résoudre, auquel cas le titre sera classé en fonction de la façon dont le problème en question est exprimé. Néanmoins, un bon nombre de titres ne présentent d'autres informations que la demande d'aide et éventuellement l'appareil dysfonctionnel ou l'objet présentant un écart à la norme. Pour ces cas précis, dans lesquels le problème n'est pas exprimé mais le besoin de résoudre un problème l'est, nous créons la catégorie « Demande d'aide à la réparation » (DemandeAide). Des titres comme « Comment réparer mon lave linge » ou « Changer une roue voiture » correspondent à cette catégorie car il ne s'agit pas d'une expression du problème rencontré par l'appareil, mais il y a bien un problème technique et un besoin explicite d'obtenir de l'aide pour résoudre ce problème, ce qui s'inscrit dans la démarche de l'utilisation de la CMR pour l'expression d'un problème.

3.2. Le guide d'annotation

3.2.1. Création du guide d'annotation

Nous avons, à partir de la typologie de base sur laquelle notre étude se base, créé une nouvelle typologie adaptée à nos données et qui prend en compte les spécificités de l'expression d'un problème dans la CMR. A partir de notre typologie, nous créons un guide d'annotation (*annexe 1*) qui permet à des annotateurs externes d'annoter nos donner, c'est-à-dire de placer chaque titre dans une catégorie d'expression du problème. Puisque nos données ne sont pas annotées en catégorie de titre, il serait impossible de calculer l'efficacité d'un système de catégorisation automatique sans pouvoir comparer ses résultats avec les « bonnes réponses ». D'un autre côté, si nous annotons manuellement les données avec lesquelles nous évaluerons l'efficacité de notre système de catégorisation automatique plus tard, le système risque d'être basé uniquement sur les spécificités de notre échantillon et donner de très bons résultats pour celui-ci, mais de mauvais résultats pour un autre échantillon de données sur lequel il n'aurait pas été « entraîné ». Pour cette raison, nous avons besoin de déléguer l'annotation de l'échantillon à partir duquel nous évaluerons le système de catégorisation automatique à un annotateur externe, et pour que cet annotateur externe sache de quelle façon il convient d'annoter les titres et sur quelles informations se baser pour effectuer une annotation correspondant à notre typologie, un guide d'annotation est nécessaire.

Un guide d'annotation doit donc définir chaque catégorie de la façon la plus claire possible pour éviter les confusions et pour obtenir une annotation reflétant le mieux possible la typologie. Il convient donc d'avoir pour chaque catégorie une définition concise, accompagnée d'exemples concrets, éventuellement tirés des données du corpus sur la base duquel nous avons créé la typologie. Après la définition de la catégorie qui se fait au niveau de la création de la typologie, nous cherchons dans le corpus des exemples de titres appartenant à chaque catégorie, avec des marqueurs variés afin de couvrir une majorité des cas susceptibles d'apparaître dans les données à annoter. A partir des marqueurs trouvés dans les exemples, nous complétons une liste de marqueurs potentiels dans les titres qui pourraient indiquer l'appartenance à une catégorie ou une autre. Les marqueurs peuvent être très précis pour des catégories spécifiques comme « Fuite », où les marqueurs seront de l'ordre de « fuite, écoulement … », alors que les marqueurs seront plus vagues pour des catégories plus globales comme « FonctionnePas » où l'on peut retrouver « ne fonctionne pas, ne [verbe] plus, en panne, hors service … ». Les marqueurs permettent de guider les annotateurs dans leur annotation mais ne constituent pas à eux même des marques sûres de l'appartenance d'un titre à une catégorie, car il est toujours important de se référer à la définition de la catégorie en premier lieu. Pour la catégorie « Fuite » par exemple, le mot « écoulement » peut constituer un marqueur pour cette catégorie s'il marque une fuite, mais « mauvais écoulement de ma machine à café » n'exprime pas une fuite, ce qui signifie que ce titre n'appartiendra pas à cette catégorie, malgré la présence du marqueur.

De plus, un titre peut contenir plusieurs informations et plusieurs types d'expressions qui pourraient le placer dans plusieurs catégories. S'agissant d'une problématique multidimensionnelle, il convient d'indiquer aux annotateurs la bonne façon d'annoter quand de tels cas se présentent. En effet, dans un cas classique dans nos données comme « Réparer ma TV Samsung qui ne fonctionne plus », on peut repérer 3 catégories indiquées par 3 types de marqueurs différents :

- « Réparer » semble indiquer une Demande d'aide à la réparation
- « Ma TV Samsung » correspondrait à Objet
- « ne fonctionne pas » correspond à la catégorie FonctionnePas

Dans un cas comme celui-ci, nous décidons de trancher et d'attribuer à notre titre la catégorie la plus spécifique de notre typologie parmi celles que l'on pourrait lui attribuer. La catégorie d'expression la plus spécifique est celle à partir de laquelle on peut avoir le plus d'informations sur la nature du problème exprimé. De cette façon, dans « Ma voiture ne démarre plus et fait un bruit bizarre », l'information du bruit produit est plus spécifique que l'information du dysfonctionnement de la voiture car elle nous apporte plus d'informations sur la nature du problème et potentiellement sur sa cause. Nous faisons donc figurer dans le guide d'annotation un ordre de spécificité à respecter lors d'un conflit entre différents marqueurs, et les annotateurs devront choisir la catégorie la plus haute dans le tableau suivant parmi celles qui sont applicables au titre.

Catégorie	Etiquette
Fuite	Fuite
Signal qui s'est déclenché	Signal
Dégradation - Usure - Saleté	Dégradation
Objet manquant (dispositif, document, outil)	Absence
Hors spécification	HorsSpec
Dispositif qui ne fonctionne pas	FonctionnePas
Action difficile ou impossible	Impossible
Demande d'aide à la réparation	DemandeAide
Action effectuée par l'appareil	Action
Objet ou composant seul	Objet
Autre type d'expression	Autre

Tableau 8 - Catégories d'expressions du problème dans l'ordre de spécificité et étiquettes correspondantes

Une définition, des exemples, ainsi que des marqueurs et un ordre de spécificité peuvent parfois s'avérer insuffisants lorsque les annotateurs sont devant des cas de titres dont il est difficile de trancher pour une catégorie ou une autre. Pour aider à la désambiguisation et pour renforcer la précision de la définition des catégories, nous choisissons également d'ajouter des contre-exemples à chaque catégorie et de les expliquer afin d'indiquer à un annotateur qui hésiterait entre deux catégories, le choix qu'il devrait faire. Nous mentionnons par exemple que « le tiroir cassette reste ouvert ne se ferme plus » appartient à FonctionnePas et non à Dégradation car le titre ne mentionne pas qu'un composant est cassé, cette dégradation est implicite et seul le dysfonctionnement résultant est exprimé.

Il est important de préciser que la création d'un guide d'annotation n'est pas une action effectuée une seule fois, mais plutôt un processus itératif de création, d'annotation, puis d'évaluation de l'annotation, et enfin de révision du guide, de perfectionnement, et d'ajout d'informations pour

mieux guider les annotateurs (cycle MAMA pour Model, Annotate, Model, Annotate). Tout au long de la création du guide nous annotons des échantillons puis vérifions nos exemples et nos définitions, et vérifions que ceux-ci correspondent aux définitions, et faisons des ajustements si nécessaire. Puis nous nous basons sur de nouvelles données pour affiner les définitions, voire effectuer des changements dans les définitions de catégories. Ce processus nous amène finalement à un guide d'annotation fonctionnel qui peut être utilisé par des annotateurs externes lors d'une campagne d'annotation.

3.2.2. Annotation externe et évaluation

Nous avons mentionné dans la partie précédente la nécessité et l'importance de faire effectuer l'annotation d'un échantillon de données par un tiers afin d'avoir une évaluation représentative du système de classification automatique sur la totalité des données du corpus. En effet, si la personne rédigeant les règles pour le programme est également l'annotateur des données sur lesquelles le programme sera testé, le programme risque de bien fonctionner uniquement sur ces données et l'évaluation d'un tel programme risquera donc de ne pas être représentative d'une catégorisation sur de nouvelles données similaires.

Dans cette démarche, après avoir annoté un échantillon de 100 titres choisis au hasard, nous proposons à un annotateur externe muni de notre guide d'annotation, d'annoter à son tour le même échantillon, afin de s'assurer que l'annotation qui sera effectuée sur de nouvelles données plus tard est représentative de notre typologie et d'une annotation que nous aurions pu effectuer. Il est possible d'expliquer à l'annotateur ce qui est attendu et l'aider à comprendre certains cas délicats, mais lui donner les réponses ou effectuer l'annotation avec lui fausserait les résultats et risquera d'empêcher une bonne annotation lors de la phase d'annotation sur les données pour l'évaluation du système de catégorisation automatique.

A partir de l'annotation interne que nous avons effectué, et de l'annotation externe effectuée par l'annotateur tiers, nous devons nous assurer que les catégories attribuées correspondent assez pour que cet annotateur puisse se consacrer à l'annotation de données que nous ne verrons pas. Ceci passe par le calcul d'un accord inter-annotateur, une mesure qui permet d'évaluer l'annotateur externe par rapport à une référence, c'est-à-dire à notre annotation préalable. Cette mesure constitue une sorte de note de l'annotation externe, et nous permet de décider d'utiliser le guide tel quel si la note est haute, ou de remanier le guide d'annotation pour l'améliorer si la note est plus faible, ce qui correspondrait plus tard à des résultats non représentatifs lors de l'évaluation du système.

Nous utilisons le coefficient Kappa de Cohen (1960) qui est un coefficient permettant de mesurer l'accord entre deux annotateurs seulement. Cette méthode se base sur l'observation de l'accord des réponses données par les deux annotateurs par rapport à un accord attendu si les deux annotateurs catégorisent les titres de façon aléatoire. En effet, même avec une annotation aléatoire, il est possible que l'annotation externe montre un haut accord avec l'annotation interne, ce qui fausserait nos résultats sur une évaluation d'un système. Il convient donc de prendre en compte cet accord aléatoire, que l'on appelle « accord attendu », et de le comparer à l'accord effectivement obtenu entre les réponses des deux annotateurs, que l'on appelle « accord observé ».

Pour calculer le coefficient Kappa de Cohen, nous commençons par représenter les réponses sous forme d'un matrice de confusion, dans laquelle les lignes représentent l'annotation externe, les colonnes représentent l'annotation interne, et leur croisement représente le nombre de titres appartenant à telle catégorie dans l'annotation de référence qui a été associé à telle catégorie dans l'annotation externe. Dans le meilleur cas, nous devrions avoir sur notre tableau une diagonale remplie sans aucun autre nombre autour, cas correspondant à un accord parfait entre les deux annotateurs. Une telle représentation est affichée ci-dessous.

Annotation externe \ Annotation interne	Fuite	Signal	Dégradation	Absence	HorsSpec	FonctionnePas	Impossible	DemandeAide	Action	Objet	Autre	Total Résultat
Fuite	2											**2**
Signal		8										**8**
Dégradation			3									**3**
Absence				5								**5**
HorsSpec					24	5	1					**30**
FonctionnePas						20						**20**
Impossible												**0**
DemandeAide								19				**19**
Action									1			**1**
Objet										11		**11**
Autre											1	**1**
Total Résultat	**2**	**8**	**3**	**5**	**24**	**25**	**1**	**19**	**1**	**11**	**1**	**100**

Tableau 9 - Matrice de confusion des annotations interne et externe et accord observé

D'après nos observations de cette représentation, l'accord inter-annotateur semble être presque parfait, sauf pour FonctionnePas pour lequel 5 titres ont été catégorisés dans HorsSpec par l'annotateur externe, et Impossible pour lequel le seul représentant de la catégorie a été également catégorisé dans HorsSpec. A partir de cette matrice, nous obtenons l'accord observé en calculant la somme des titres sur lesquels il y a un accord, que l'on divise par le nombre de total de titres, ce qui donne dans notre cas :

$$A_o = (5+1+1+3+19+20+2+24+11+8)/100 = 0,94$$

Notre accord observé est de 0,94 sur une échelle de 0 à 1, ce qui correspond à un accord quasi-parfait. Nous devons néanmoins le comparer avec l'accord attendu, que nous calculons pour chaque cellule en multipliant le nombre total de titres dans telle catégorie en colonne par le nombre total de titres catégorisés dans telle catégorie par l'annotateur externe en colonne, puis en divisant le tout par le nombre total de titres. Cette mesure globale de l'accord attendu, correspondant à une situation d'annotation aléatoire par les deux annotateurs, a pour valeur :

$$A_a = 0,1806$$

Il s'agit d'un accord attendu faible, qui correspond à une probabilité faible d'accord dans des conditions aléatoires. Nous pouvons à présent appliquer la formule du calcul du coefficient Kappa de Cohen afin d'obtenir une mesure unique de l'accord inter-annotateur. La formule est la suivante :

$$\kappa = (A_o - A_a)/(1 - A_a) = (0,94 - 0,1806)/(1 - 0,1806) = 0,7594/0,8194 = 0,93$$

Nous obtenons un fort coefficient Kappa qui traduit un fort accord inter-annotateur, car un accord supérieur à 0,90 sur une échelle de 0 à 1 peut être considéré comme quasi-parfait et nous permet de justifier de l'utilisation en l'état de notre guide d'annotation et de notre annotateur externe pour l'annotation d'un nouvel échantillon qui sera utilisé lors de l'évaluation de notre système automatique dont nous allons détailler, dans la partie suivante, la méthodologie de création, le fonctionnement, ainsi que les résultats obtenus.

Dans le tableau 9 ci-dessus, nous avons pu observer un accord quasi-parfait sur les différentes catégories d'expression du problème, sauf dans le cas de la catégorie HorsSpec, dans laquelle sont classés les 6 titres sur lesquels un accord n'a pas été observé. Ces écarts toujours présents entre notre annotation et celle de l'annotateur externe sont riches d'enseignements en ce qui concerne la précision de la définition des catégories dans le guide d'annotation, et permet de montrer la différence de traitement entre des cas qui n'ont pas été représentés spécifiquement dans les exemples présents dans le guide. Les 5 cas de désaccord sont les suivants :
1) Plaques de cuisson qui ne marchent plus après avoir pris l'eau
2) Réparer un sèche linge dont l'eau n'arrive plus dans le bac de récupération
3) Smart tv toshiba ne connecte plus a aux apli premieum
4) Ma machine ne prend pas l'eau
5) Réparer FOUR ENCASTRABLE AEG MODELE B 8972-5 qui ne chauffe plus (plus aucune commande ne fonctionn
6) Réparer hdmi atlas hd 200s indetectable sur tv

Les 5 premiers titres correspondent à FonctionnePas dans notre annotation mais ont été classifiés comme HorsSpec dans l'annotation externe. Le dernier titre, en revanche, correspond à la catégorie Impossible dans notre annotation, mais a été assigné à la catégorie HorsSpec dans l'annotation externe. On constate donc une surreprésentation de la catégorie HorsSpec dans l'annotation externe, qui peut être expliquée de différentes façons. Pour les titres correspondant à FonctionnePas, malgré la présence de marqueurs montrant une fonctionnalité attendue qui est rendue dysfonctionnelle, la catégorie HorsSpec est plus spécifique que la catégorie FonctionnePas dans notre typologie. De ce fait, la définition de la catégorie HorsSpec peut amener à penser, dans certains cas, que le dysfonctionnement décrit par la précision de la fonctionnalité touchée (exemple : « ne prend pas l'eau ») fait correspondre le titre à la catégorie HorsSpec, étant plus spécifique.

Dans le cas du dernier titre, correspondant pour nous à la catégorie Impossible, la classification est effectivement plus délicate et sujette à interprétation. Nous avons choisi la catégorie Impossible car le marqueur représentatif de la catégorie, « indetectable », fait part d'une impossibilité de la machine d'activer une fonctionnalité, la détection dans ce cas. Pourtant, « indetectable » peut être également compris comme un adjectif représentant un état hors de la spécification de la machine, qui amène donc un annotateur externe à l'assigner à la catégorie correspondante. L'absence d'exemple spécifique à ce type d'adjectif explique ce désaccord, et la présence d'un exemple similaire dans le guide d'annotation aurait pu permettre un accord sur ce titre et d'autres présentant un adjectif relevant de l'impossibilité.

Dans cette troisième partie, nous avons adapté la typologie de base à nos données, en créant une typologie contenant 11 catégories distinctes, détaillées dans un guide d'annotation. Le guide d'annotation a été ensuite utilisé pour l'évaluation de l'annotation d'un annotateur externe, qui a donné de bons résultats dans l'ensemble, nous permettant de garantir une certaine fiabilité de l'annotation qui sera utilisée comme base pour l'évaluation de notre système de classification automatique.

4. Classification automatique des titres

Dans cette quatrième partie, nous créons un système de classification automatique des titres suivant la typologie décrite dans la partie précédente. Nous détaillons donc les aspects techniques et linguistiques du processus de construction d'un tel système, et expliquons également le passage des catégories de notre typologie à un système de règles utilisables par notre programme pour déterminer de façon automatique la catégorie à laquelle devrait appartenir un titre donné.

4.1. Préparation des données

4.1.1. Récupération des données à utiliser et création de l'échantillon test

Le corpus CoCoRep constitué des données que nous étudions tout au long de ce mémoire, a été créé au format XML suivant la norme TEI, norme permettant d'encoder les informations dans des corpus de façon à pouvoir les utiliser et les échanger facilement (Romary & Hudrisier, 2002). De ce fait, le corpus CoCoRep contient toutes les informations dont nous avons besoin pour notre étude, cependant les informations pour chaque post se trouve dans un fichier séparé, et un nombre considérable d'informations supplémentaires est également disponible sur le corpus et sur chaque post, ce qui entrave la visibilité des informations dont nous avons effectivement besoin pour cette étude. Pour l'analyse de titres que nous nous apprêtons à effectuer, il semble plus facile et plus clair de disposer uniquement des données dont nous avons besoin de façon centralisée, c'est-à-dire dans un même fichier, et dans un format dans lesquelles les données sont plus directement accessibles et plus faciles à analyser.

Dans un premier temps, nous choisissons d'extraire des fichiers XML sous la norme TEI uniquement les informations susceptibles d'être intéressantes, donc dans le cas présent, le titre de chaque post, ainsi que des métadonnées associées à chaque post, comme son identifiant unique, l'auteur du post sous la forme de son nom d'utilisateur, la catégorie et la sous-catégorie de produit dans laquelle le post a été publié, la marque, le modèle, ainsi que l'année de l'appareil dont il est question, et enfin le lien du post sur le site original, ce qui nous permet toujours d'avoir un accès direct à la source des données récoltées. En utilisant uniquement les données susceptibles de nous servir dans le cadre de notre étude, nous créons effectivement une sorte de nouveau sous-corpus à partir de notre corpus général CoCoRep, reflétant la visée de la recherche (Poudat et al., 2020).

A partir des données extraites, nous créons un nouveau fichier unique contenant les informations nécessaires sur la totalité des posts du corpus entier CoCoRep, dans un format facilement manipulable à l'aide d'un tableur et de bibliothèques Python, qui est le format TSV. Le format TSV (acronyme de Tabulation Separated Values) est un format texte dans lequel chaque ligne correspond aux informations d'un post, et chaque information sur le post est séparée par une tabulation. Cela a pour effet de créer un tableau facilement manipulable et exploitable par différents logiciels dont nous aurons besoin lors de nos analyses. Le tableau au format TSV est également plus rapide à charger dans ces programmes en comparaison au corpus au format XML TEI, car pour celui-ci, il faudrait que le programme cherche les données dans chaque fichier XML séparément au chargement, pour ensuite mettre à disposition les informations importantes.

A partir du fichier TSV, long de 61269 lignes correspondant chacune à un post et à son titre, nous avons créé deux échantillons de données. Le premier échantillon créé est un échantillon de 100 titres utilisés pour la création du programme et le calcul de l'accord inter-annotateur. En effet, cet échantillon visible tout au long de l'étude nous permet d'avoir un fenêtre sur les données composant notre corpus et les catégories d'expression du problème que celui-ci contient. Il est donc utilisé pour l'observation et la rédaction des règles nécessaires au système de classification automatique dont nous expliquons la mise en place dans cette partie. Il s'agit également de l'échantillon utilisé pour

l'évaluation de la pertinence des catégories et le calcul de l'accord inter-annotateur effectué dans la partie précédente. De plus, nous utilisons cet échantillon pour évaluer le modèle de lemmatisation que nous utiliserons pour analyser et catégoriser les titres. Nous proposons ci-dessous un aperçu de cet échantillon accompagné des catégories attribuées par notre annotation, ainsi que par l'annotateur externe.

id	auteur	...	titre	Annotation interne	Annotation externe
CR-1-32626	gégé	...	Réparer lave-vaisselle indesit dpg 36 aix qui ne tourne plus	FonctionnePas	FonctionnePas
CR-125-34193	Gugu	...	Comment réparer ma console de vélo SVP?	DemandeAide	DemandeAide
CR-16-55122	Kolawolé	...	Télévision écran incurvé cassé	Dégradation	Dégradation
CR-16-15731	Breathless	...	Téléviseur LG 32LH4000 ne s'allume plus	FonctionnePas	FonctionnePas
CR-14-44946	denis	...	Tous les voyant clignotent sur sèche linge Ariston TCLG31X	Signal	Signal
CR-14-32615	Pat	...	Sèche-linge Bosch wtc84101ff se coupe apres quelques minutes	HorsSpec	HorsSpec
CR-66-64424	Hervé	...	Thermomix TM21	Objet	Objet
CR-2-47571	Golgoth51	...	Réparer cafetière Dolce Gusto Drop eau chaude reste froide	HorsSpec	HorsSpec
CR-16-53658	Tania	...	Mise en veille automatique de ma tv sans que je ne lui demande	HorsSpec	HorsSpec
CR-16-26510	gaellez	...	TV Toshiba	Objet	Objet

Tableau 10 - Extrait de l'échantillon de 100 titres au format TSV

Le deuxième échantillon que nous créons est un échantillon test, également appelé « test-set », constitué cette fois-ci de 200 titres, avec des données différentes mais similaires à notre échantillon de 100 titres. Les données de ce test-set ne sont pas destinées à être visibles pour la conception du système de catégorisation automatique, mais plutôt à être annotées par notre annotateur externe, afin de constituer une annotation de référence (un *gold*) sur la base duquel notre classifieur pourra être évalué. Puisque les exemples de notre test-set sont similaires à nos données de l'échantillon de 100 titres mais tout de même différentes, le système ne sera pas « entraîné » sur celles-ci. Néanmoins, un système de catégorisation performant doit tout de même être capable de catégoriser ces données de façon assez efficace et représentative de notre typologie.

4.1.2. Méthode de parsing et évaluation

Afin d'analyser les titres de chaque post pour pouvoir les placer dans une catégorie, nous nous penchons vers une approche lexicale et syntaxique. De cette façon, les règles que le système de classification utilisera pour décider d'attribuer une catégorie à chaque titre seront des règles basées sur le repérage de lemmes ainsi que le repérage de patrons syntaxiques récurrents dans les différentes catégories d'expression du problème de notre typologie.

Pour le repérage lexical, il faut donc que le classifieur soit capable de repérer les mots séparément. Nous utilisons donc un outil de tokenisation qui permet de transformer un titre en une

succession de tokens qui auront ensuite chacun des informations supplémentaires permettant l'analyse. Chaque token, dans une forme particulière, se voit attribuer un lemme à des fins de normalisation, dans un processus appelé « lemmatisation » (Chanier et al., 2014). La lemmatisation consiste en l'attribution d'un lemme, c'est-à-dire d'une forme canonique d'un mot variable, ce qui est d'autant plus important dans un contexte de CMR où la variabilité de l'orthographe des mots est considérablement élevée.

Le repérage de patrons syntaxiques, d'un autre côté, ne peut se faire sans l'ajout d'informations syntaxiques et morphosyntaxiques aux tokens de la phrase. Cet ajout d'informations se fait grâce à un analyseur morphosyntaxique qui pourra indiquer, pour chaque token, sa relation avec d'autres tokens ainsi que la nature de la relation, et une étiquette indiquant la nature ou la « partie du discours » correspondant au token. Cette analyse préalable nous permet ensuite d'utiliser des règles syntaxiques et morphosyntaxiques afin de classifier les titres.

Les outils utilisés pour ces analyses préalables font partie d'une boîte à outils informatique appelée Stanza (Qi et al., 2020) développée par l'université de Stanford. Stanza est une bibliothèque Python de modèles d'analyse de langage qui peut être utilisée pour convertir un texte brut en liste de tokens accompagnés d'informations supplémentaires comme son lemme, sa nature, ou sa fonction syntaxique dans la phrase. Des modèles d'analyse existent dans plus de 70 langues du monde, et de plus en plus de modèles performants continuent à voir le jour. Nous utilisons donc l'analyseur de Stanza qui fait appel entre autres à des systèmes de tokenisation, lemmatisation, ou encore d'analyse et annotation morphosyntaxique.

Le document obtenu à partir de ces analyses est un document au format CoNLL-U, similaire au format TSV, pour lequel chaque token de la phrase correspond à une ligne, et chaque colonne nous donne une information différente sur ce token. Les informations de nature, de fonction, ou encore de dépendances syntaxiques, sont proposées dans un framework international appelé « Universal Dependencies », ou UD (Nivre et al., 2020). Universal Dependencies est un projet international visant à atteindre une annotation linguistique consistante et similaire à travers toutes les langues prises en compte, et à proposer une typologie consistante de parties du discours et d'informations syntaxiques permettant un traitement automatique du langage robuste et similaire peu importe la langue.

Malgré la puissance des modèles de langage proposés par Stanza, il reste important de vérifier que les méthodes proposées sont applicables à nos données, car les données textuelles de CMR présentent certaines spécificités qui rendraient une analyse automatique délicate si les modèles n'ont pas été entraînés sur ce type de données. Afin de vérifier que ces outils soient utilisables, nous observons le résultat du processus de lemmatisation sur l'échantillon de 100 titres que nous avons annoté en catégorie. Parmi ces 100 titres, nous comptons au total 812 tokens, pour lesquels 793 (97,66%) ont été lemmatisés correctement par le modèle de lemmatisation du français de Stanza, ce qui correspond à un score assez satisfaisant pour justifier de l'utilisation de cette méthode d'analyse préalable sur nos données.

Parmi les erreurs que peut faire le lemmatiseur de Stanza, il peut y avoir par exemple des erreurs de lemmatisation dues à des fautes d'orthographe dans le texte original qui sont lemmatisées telles-quelles. Nous pouvons prendre l'exemple de « dépane », variante orthographique de « dépanne », qui est lemmatisée en « dépaner » et reconnu comme verbe, alors que le verbe normalisé qui devrait être associé est « dépanner ». Ces variations orthographiques amènent d'autres erreurs, comme « Lave Vaisselle » reconnu comme deux noms propres, qui gardent donc leur majuscule à la lemmatisation, au lieu d'être reconnus comme un nom commun. Nous aborderons ces spécificités et les difficultés qu'elles amènent dans la sous-partie suivante.

4.1.3. Difficultés rencontrées face aux données

Nous avons plusieurs fois au cours de ce mémoire évoqué la spécificité des données de CMR et leurs différences comparées à des données langagières plus typiques. La majorité des systèmes d'analyse automatique de la langue se basent sur une version normalisée de la langue, autrement dit une version standard de la langue qui fonctionne moins bien avec des données qui diffèrent de cette norme sur différents points, ce qui est également le cas sur des données orales avec la reconnaissance vocale sur des accents autres que l'accent standard, ce qui peut notamment poser des problèmes éthiques.

La différence avec des données typiques qui complique le plus le traitement des données textuelles se trouve dans la variation orthographique. Dans les données provenant de la CMR, les fautes d'orthographe sont nombreuses et dans une analyse textuelle, ces fautes seront analysées telles quelles. La présence de fautes d'orthographe ne met pas à mal uniquement le modèle d'analyse automatique, mais également la compréhension humaine, ce qui peut jouer un rôle également lors d'une annotation manuelle comme nous le faisons avec nos données. En effet, si un analyseur automatique est capable de comprendre et de normaliser l'orthographe d'un mot, mais que celui-ci est incompréhensible pour un annotateur humain, l'annotation perdra en qualité et l'évaluation du système ne sera pas représentative de la réalité de l'efficacité de celui-ci. Il faut donc prendre en compte les fautes d'orthographe et autres variations orthographiques dans l'analyse des données, sans pour autant altérer les données qui pourraient finir par ne plus refléter les réalités linguistiques initiales. Le traitement automatique doit donc être assez robuste pour permettre une prise en compte de formes altérées des mots (Chanier et al., 2014).

Un autre degré de variabilité à prendre en compte se retrouve sous la forme d'abréviations. Les données de CMR contiennent souvent des abréviations qui constituent un gain de temps à l'écriture pour l'auteur d'un message ou d'un post, et ne pose pas de problème lorsque cette abréviation est partagée et comprise par les interlocuteurs ou autres utilisateurs susceptibles de voir le message. Cependant, comme pour les fautes d'orthographe, la variation amenée par la présence d'abréviations est forte, et complique d'autant plus l'analyse. Un système assez robuste serait donc également capable de prendre en compte les formes abrégées des mots, et au mieux serait capable de normaliser les abréviations en leur associant le lemme normalisé « complet » correspondant.

Sur le plan syntaxique, l'ordre des mots et le caractère télégraphique des expressions dans la CMR représente également une spécificité de ces données qui amènent une difficulté supplémentaire lors du traitement du texte. Les schémas syntaxiques ne sont parfois pas naturels si enlevés de leur contexte, et ceux-ci sont également très variables et difficilement prédictibles. Malgré des ordres de mots inchangés, le cas des titres est un exemple assez poussé de cet argument. Les informations contenues dans les titres sont souvent dans un ordre difficilement prédictible et les mots grammaticaux sont souvent omis par principe d'économie, puisque ceux-ci sont vides de sens. Ceci peut rendre le sens d'une phrase ou d'un titre moins clair, comme dans l'exemple « Machine laver plus eau », pour lequel on imagine que la machine à laver n'a plus d'eau, mais pour lequel il pourrait également être acceptable de comprendre que l'utilisateur voudrait laver sa machine avec plus d'eau, ou qu'il y a un surplus d'eau dans sa machine à laver, ce qui amène une interprétation contraire à la première.

Enfin, la façon d'écrire, au-delà de l'orthographe des mots, amène également une grande variabilité qui rend l'analyse plus complexe. Le parseur peut être induit en erreur par des lettres majuscules, par la présence ou non de traits d'union, ou encore par la ponctuation, qui peuvent être utilisés d'une façon différente du standard linguistique. De cette façon, « lave-vaisselle » écrit « Lave Vaisselle » pourra être considéré comme deux noms propres plutôt qu'un nom commun. En plus d'une spécificité de la CMR, la présence de majuscule au début de chaque mot est notamment courante dans des titres, comme des titres de film, de chansons, ou de livres, et les utilisateurs peuvent appliquer les standards du titre à ce qui diffère tout de même de ces autres exemples.

4.2. Systématisation des catégories

4.2.1. Repérage des caractéristiques formelles

Pour créer notre système de catégorisation automatique, nous devons attribuer à chaque catégorie un nombre de règles qui permettra de catégoriser les titres. Ce système de règles se base sur le repérage de marqueurs et de caractéristiques formelles, lexicales ou syntaxiques, à la manière de l'annotation manuelle, dans le texte que nous donnons en entrée à notre catégoriseur. Nous listons ici les différentes caractéristiques retenues et les différents patrons observés, puis les organiserons dans la prochaine sous-partie en les regroupant respectivement sous la catégorie qu'ils nous permettent de détecter.

Tout d'abord, les titres contiennent des noms communs qui peuvent représenter une variété de concepts. Compte tenu de la qualité concise et impersonnelle attendue d'un titre, il arrive souvent dans nos données que chaque titre contienne un nom commun donnant une information différente sur la nature du problème. Il sera donc important de déterminer le contexte dans lequel les noms communs apparaissent pour déterminer à quoi ceux-ci correspondent. Les noms communs peuvent correspondre au type d'appareil ou à la catégorie d'appareil impacté par le problème ou le dysfonctionnement, ce qui ne nous donne pas d'information supplémentaire sur la nature du problème, mais peut constituer un marqueur indiquant l'objet (et si celui-ci est seul, il constitue un marqueur de la catégorie Objet). Un nom propre peut également apparaître dans les titres pour marquer une information similaire, qui peut être la marque ou le nom de l'objet ou de l'appareil dysfonctionnel. Ces deux informations peuvent coexister dans un titre, par exemple : « Ma voiture Twingo est en panne », ou séparément dans « Ma Twingo est en panne » ou « Ma voiture est en panne ».

Les noms communs peuvent également correspondre directement au type de problème, auquel cas il conviendra de séparer les lemmes dans leurs catégories respectives afin d'obtenir une information sur le type de problème. De cette façon, le nom commun « fuite » nous indiquera une appartenance à la catégorie Fuite, là où le nom « panne » nous indiquera une appartenance à la catégorie FonctionnePas. Autrement, les noms communs peuvent aussi nous donner une information sur l'action effectuée par l'appareil ou l'action envisagée ou entreprise par l'auteur du post, auquel cas ce nom traduire une action et pourra prendre la forme d'un nom déverbal (exemple : « Réparation » ou « Mise à jour »).

Les titres peuvent contenir des verbes qui, selon leur signification, mais également leur forme, pourront nous informer sur la catégorie correspondante et le type d'action exprimée. Un verbe à la troisième personne ou à une forme participe présent ou passé correspondra à une action ou au résultat d'une action faite ou non par l'appareil, ou à un état résultant d'une action. Des formes comme « cassé », « tournant », « siffle » nous donnent donc des indications sur l'état ou l'action de l'appareil plutôt que celle de l'utilisateur. En revanche, un verbe à l'infinitif, comme certains noms déverbaux vus plus haut, correspondra plutôt à une action de l'utilisateur auteur du post. Celui-ci demande souvent implicitement, à travers un simple verbe à l'infinitif, comment effectuer l'action décrite par le verbe. Il arrive également que la demande soit explicite, auquel cas les adverbes interrogatifs seront également présents et rattachés au verbe infinitif pour demander comment effectuer une réparation.

Les adjectifs peuvent indiquer l'état d'un appareil défectueux, et peuvent être liés syntaxiquement à un nom correspondant à l'appareil ou l'objet en question. Une difficulté concernant ces adjectifs est la complexité de l'identification de la fonction sémantique de l'adjectif, c'est-à-dire de savoir si un adjectif lié à un nom correspondant à un objet nous donne une information sur le problème que rencontre cet objet, ou bien simplement une information supplémentaire sur l'objet. La différence est donc assez complexe à déterminer entre des cas comme « Radio réveil gris » et « Radio réveil bruyant » sur ce point.

Les adverbes négatifs sont très intéressants à observer car ils nous donnent presque toujours une information sur une absence d'objet, de fonctionnalité, ou de comportement attendu d'un appareil. Un verbe à la troisième personne pourra donc être entouré d'adverbes négatifs comme « ne » et « pas » ou « ne » et « plus » pour parler d'un comportement attendu d'un appareil mais qui ne montre plus ce comportement, alors qu'un adverbe négatif lié syntaxiquement à un nom représentera plutôt une expression d'absence (avec par exemple « Pas de son » ou « Plus de piles »).

Enfin, nos données contiennent également des formules et des marques de politesse ou d'adresse aux interlocuteurs typiques des CMR, malgré le caractère assez impersonnel attendu dans un titre de post sur un forum. Si les aides à la rédaction d'un post indiquent que le titre doit contenir les informations les plus importantes en donnant des exemples canoniques comme « Comment réparer ma baignoire qui fuit ? », les utilisateurs intègrent également des formules de politesse ou des interjections, ou encore des mots à valeur vocative. On peut citer par exemple « Bonjour », « S'il vous plaît » ou encore « Merci d'avance » dans le titre, là où l'on s'attendrait plutôt à observer ce genre d'occurrences dans les corps de post.

4.2.2. Attribution de règles formelles aux catégories d'expression

Dans cette sous-partie, nous regroupons les caractéristiques formelles observées dans les titres sous formes de règles associées à chaque catégorie d'expression du problème afin d'expliquer le fonctionnement du programme de catégorisation automatique que nous avons mis en place et apporter des clarifications sur sa catégorisation et les erreurs qui peuvent être causées par le choix des règles. Le coeur du programme se base sur un fonctionnement simple sous forme de choix conditionnel ordonné : si un titre contient un marqueur ou répond à une règle caractéristique de la première catégorie dans notre tableau 1, le classifieur attribue à ce titre l'étiquette correspondant à la première catégorie (dans notre cas la catégorie Fuite). Si toutefois le titre ne correspond à aucune de ces règles, le classifieur teste les marqueurs et les règles de la deuxième catégorie, et ainsi de suite dans l'ordre de spécificité des catégories.

Nous rédigeons donc un ensemble de règles à respecter pour chaque catégorie, que ce soit des marqueurs lexicaux ou des règles syntaxiques, et expliquons pour chaque règle, les raisons pour lesquelles nous avons choisi de l'intégrer au programme. Si une omission notable d'un marqueur ou d'une règle existe, nous la justifions également.

a) Fuite (Fuite) :

Un titre est considéré comme faisant partie de cette catégorie s'il contient au moins une occurrence du lemme « fuite » ou « fuir ». Malgré la mention du lemme « écoulement » dans notre guide d'annotation, celui-ci est plus souvent associé à un comportement ou une fonctionnalité normale d'appareils électroménagers, comme l'écoulement d'eau d'un lave-linge ou l'écoulement d'une machine à café. Pour cette raison, nous préférons ne pas inclure ce lemme dans la liste des marqueurs pour cette catégorie.

b) Signal qui s'est déclenché (Signal) :

Un titre est considéré comme faisant partie de cette catégorie s'il contient au moins une occurrence des lemmes suivants : « voyant », « bip », « témoin », « alarme », « signal » ou « clignotement ». Les lemmes « code » et « erreur » sont également pris en compte uniquement dans le cas où ils sont suivis d'un nombre ou d'un code d'erreur, ce qui évite que les titres contenant ces lemmes sans la signification de signal associé, comme « code pin » ou « erreur de manoeuvre », soient pris en compte. Les verbes « afficher » et « clignoter » sont également pris en compte s'ils sont caractéristiques d'une action effectuée par l'appareil, donc à la troisième personne ou en forme participe passé ou présent.

c) Dégradation - Saleté - Usure (Dégradation) :

Un titre est considéré comme faisant partie de cette catégorie s'il contient au moins une occurrence des lemmes suivants : « casse », « casser », « briser », « fissurer », « griller », « abimer », « déchirer », « tâcher », « tâche », « corrosion », « rouille », « rouiller », « usure », « user », « sale », « saleté ».
Ces mots sont tous liés au vocabulaire de la dégradation, de la saleté, ou de l'usure, ce qui correspond à notre typologie. Cette liste est renforcée avec des variations orthographiques qui pourraient ne pas être prises en compte ou reconnues par le lemmatiseur, comme l'absence d'accent (« salete », « use » etc…).

d) Objet manquant - Dispositif, document, outil (Absence) :

Un titre est considéré comme faisant partie de cette catégorie s'il contient au moins une occurrence d'un nom commun non sujet associé syntaxiquement à un de ces marqueurs, qui marquent l'absence : « pas », « plus », « perte », « absence » et « aucun ». « Plus » est notamment compliqué à analyser car les occurrences de « plus » pour parler d'un surplus d'un objet ou d'une substance sont également considérés comme faisant partie de cette catégorie par le programme.

e) Hors spécification (HorsSpec) :

Un titre est considéré comme faisant partie de cette catégorie s'il contient au moins une occurrence des lemmes suivants : « problème », « erreur », « défaut », « soucis », « anormal », « bruit », « mauvais », « mais ». Cette liste a été enrichie avec des exemples de variations orthographiques (« pb », « probleme », « problems », « defaut » …). Un nom lié à un adjectif peuvent constituer un marqueur de cette catégorie si le nom n'est ni objet, ni sujet d'un autre composant syntaxique, et que le nom ou l'adjectif n'appartiennent pas aux marqueurs d'autres catégories, comme FonctionnePas, Impossible, ou DemandeAide. Un verbe quelconque à la troisième personne ou à la forme participe peut également constituer un marqueur si celui-ci est dans une forme positive. Enfin, un adverbe donnant une information sur le problème constitue également un marqueur. De cette façon, « Mon lave-linge tourne mal » correspondra à cette catégorie, alors que « Mon lave-linge ne tourne pas » correspondra à la catégorie suivante.

f) Dispositif qui ne fonctionne pas (FonctionnePas) :

Un titre est considéré comme faisant partie de cette catégorie s'il contient au moins une occurrence des lemmes ou suites de mots suivants : « hs », « hors service », « panne », « bloquer », « blocage », « bug », « dysfonctionnement », « dysfonctionner », « faible », « hors d'usage », « court circuit ». Un verbe à la troisième personne ou à la forme participe uniquement dans une forme négative, ainsi qu'un adjectif à la forme négative, peuvent également constituer un marqueur pour cette catégorie. La seule exception ici sera le verbe « pouvoir », car « ne pas pouvoir » constitue un marqueur typique de la catégorie suivante.

g) Action difficile ou impossible (Impossible) :

Un titre est considéré comme faisant partie de cette catégorie s'il contient au moins une occurrence des lemmes suivants : « impossible », « difficile », « dur », « echec ». Le verbe « pouvoir » dans une forme négative correspond également à un marqueur de cette catégorie. De plus, un adjectif notant l'impossibilité, donc précédé du préfixe « in- » et suivi du suffixe « -able », constitue également un marqueur de cette catégorie.

h) Demande d'aide à la réparation (DemandeAide) :

Un titre est considéré comme faisant partie de cette catégorie s'il contient une marque d'interrogation, comme un adverbe interrogatif (« comment », « où » etc …), une formulation interrogative comme « est-ce que », ou un point d'interrogation. Le verbe « aider » à toutes ses formes, ou n'importe quel autre verbe à l'infinitif constitue également un marqueur pour cette catégorie. Enfin, un nom déverbal en racine syntaxique du titre sera également un marqueur pour cette catégorie. Nous donnons deux exemples de noms déverbaux typiques de cette catégorie, « demande » et « recherche », mais définissons également les noms déverbaux comme tout autre nom commun se terminant par « -tion » (exemple : « Réparation ») ou « -ure » (exemple : « Ouverture »).

i) Action (Action) :

Un titre appartient à cette catégorie si sa racine syntaxique est un verbe quelconque et que le titre n'a pas été catégorisé par une des autres catégories.

j) Objet (Objet) :

De façon similaire, un titre appartient à cette catégorie si sa racine syntaxique est un nom commun, propre, ou non catégorisé et que le titre n'a pas été catégorisé par une des autres catégories. Les mots non catégorisés (l'étiquette UD correspondante étant X) correspondent souvent à des noms propres ou des noms communs décrivant un objet mais ayant été mal orthographiés au point où l'analyseur morphosyntaxique ne peut pas trancher sur sa nature.

k) Autre type d'expression (Autre) :

La catégorie « Autre type d'expression » est différente des autres en ce qu'elle est qualifiée par son absence de marqueurs. Au contraire, un titre ne contenant aucun marqueur permettant de le classer dans une autre catégorie sera catégorisé comme Autre. Les autres règles étant assez larges dans la définition de leurs marqueurs, et également par rapport à nos données, nous ne nous attendons pas à un grand nombre de titres contenus dans cette catégorie.

Dans cette quatrième partie, nous avons pu passer de la définition des catégories de notre typologie à une représentation formelle des caractéristiques de celles-ci, à l'aide de marqueurs repérables dans les titres et d'une hiérarchisation de ces marqueurs en spécificité du type d'expression du problème. Le système de classification automatique créé à partir de ces règles représente le fruit du travail d'étude que nous avons effectué, et doit être évalué afin de déterminer l'efficacité, mais également la pertinence de l'utilisation d'un tel système de règles, et la pertinence des règles elles-mêmes, sur de telles données.

5. Résultats de la classification

Dans cette cinquième et dernière partie, nous évaluons le système de classification automatique dont les réflexions et le processus de création sont décrits dans la partie précédente. Nous évaluons le système en comparant ses résultats à l'annotation externe d'un échantillon aléatoire de titres du corpus, et observons les forces et les faiblesses du système afin d'en tirer des conclusions sur son fonctionnement. De plus, nous effectuons une classification automatique du corpus entier, grâce au système créé, afin d'obtenir un aperçu statistique des types d'expressions du problème contenus dans un corpus de CMR contenant des demandes d'aide à la réparation.

5.1. Résultats obtenus sur l'échantillon test et évaluation

Dans cette partie, nous détaillons les résultats obtenus à la sortie du classifieur automatique sur notre échantillon test et proposons des mesures d'évaluation ainsi que des explications sur les résultats. L'échantillon test qui a été annoté automatiquement correspond à l'échantillon de 200 titres annotés au préalable manuellement par notre annotateur externe, contenant des données dont nous n'avions pas conscience lors de la rédaction des règles, ce qui garantit une fiabilité supérieure des résultats de l'évaluation du programme. Après l'annotation automatique de cet échantillon test, nous obtenons les résultats que nous affichons sous forme de matrice dans le tableau 5 ci-dessous. Dans ce tableau, les lignes correspondent aux catégories données aux titres par l'annotateur externe, qui constituent donc les catégories de référence, et les colonnes correspondent aux catégories attribuées automatiquement par notre programme de classification automatique. Parmi les catégories attribuées automatiquement, plus la couleur d'une cellule est intense, plus la proportion d'annotation est forte par rapport au total des titres annotés avec cette catégorie. La couleur verte correspond aux bonnes réponses, tandis que les cases en rouge correspondent aux mauvaises classification du système automatique.

Annotation manuelle \ Classifieur	Fuite	Signal	Dégradation	Absence	HorsSpec	FonctionnePas	Impossible	DemandeAide	Objet	Autre	Total Résultat
Fuite	1										**1**
Signal		11			2	1					**14**
Dégradation			7		1			2			**10**
Absence				1					2		**3**
HorsSpec		1		3	34	3		4	1		**46**
FonctionnePas		1		2	5	42		2			**52**
Impossible							1				**1**
DemandeAide					5			35	1		**41**
Objet					5				25		**30**
Autre									2		**2**
Total Résultat	**1**	**13**	**7**	**6**	**52**	**46**	**1**	**43**	**31**	**0**	**200**

Tableau 11 - Matrice des résultats du classifieur comparés à l'annotation de l'échantillon de 200 titres

Nous pouvons déjà voir des résultats qui semblent assez bons, sauf pour deux catégories, qui sont Absence et Autre, pour lesquelles la majorité ou la totalité des titres qui ont été placés dans ces catégories n'y correspondaient pas. Néanmoins, afin d'obtenir une évaluation numérique, nous calculons les mesures de précision et de rappel de notre système de classification automatique à partir des résultats ci-dessus.

La précision et le rappel sont deux mesures statistiques qui permettent d'évaluer un système automatique de classification ou de recherche d'information, en comparant les résultats obtenus par le système aux informations de référence, c'est à dire aux « bonnes réponses » auxquelles devraient correspondre totalement les résultats d'un classifieur parfait. La mesure de précision correspond à la proportion d'individus qui se voient attribués par le système la bonne catégorie parmi tous les individus qui se voient attribuer cette catégorie, alors que la mesure de rappel correspond à la proportion d'individus qui se voient attribués par le système la bonne catégorie parmi tous les individus correspondant effectivement à cette catégorie dans notre annotation de référence. Sachant qu'un bon score sur les deux mesures est important pour qualifier de bon un système automatique, il est important d'obtenir une bonne moyenne des deux mesures. La f1-mesure (ou f1-score) correspond à la moyenne harmonique de la précision et du rappel afin d'obtenir, par catégorie, une mesure globale de l'efficacité du système de classification automatique sur cette catégorie. Nous calculons également une moyenne de ces f1-mesures afin d'obtenir une mesure globale de l'efficacité de notre système de classification automatique sur toutes les catégories. Le tableau contenant ces mesures se trouve ci-dessous. L'échelle rouge-vert permet de mieux visualiser les mesures en distinguant les bonnes mesures des moins bonnes.

Catégorie	Précision	Rappel	F1-mesure
Fuite	1	1	1
Signal	0,85	0,79	0,81
Dégradation	1	0,7	0,82
Absence	0,17	0,33	0,22
HorsSpec	0,65	0,74	0,69
FonctionnePas	0,91	0,81	0,86
Impossible	1	1	1
DemandeAide	0,81	0,85	0,83
Objet	0,81	0,83	0,82
Autre	0	0	0
TOTAL	**0,72**	**0,71**	**0,71**

Tableau 12 - Mesures d'évaluation du classifieur en fonction de chaque catégorie

Ces mesures correspondent effectivement à nos observations préalables sur la matrice de résultats. On peut voir que la moyenne des f1-mesures se situe à environ 0,71, ce qui est assez bien dans l'ensemble, et reflète une f1-mesure très bonne voire parfaite sur la majorité des catégories, ainsi qu'une f1-mesure moins bonne voire nulle sur d'autres catégories. Les catégories dont la f1-mesure s'élève à 1, la valeur la plus haute, correspondent à des catégories assez faciles à repérer dans la nature de leurs marqueurs. En revanche, plus les marqueurs se complexifient et les catégories deviennent vagues, plus la f1-mesure baisse. Il est important de noter l'absence de la catégorie « Action » dans ce tableau, car l'échantillon test ne contenait aucun titre appartenant à cette catégorie. Effectivement, malgré la présence de cette catégorie dans la typologie, correspondant à une action neutre effectuée par l'appareil, cette catégorie existe en réalité assez peu dans nos données, ce qui peut être expliqué par une des spécificités de nos données qui est que les utilisateurs rédigeant un post le font en général à cause d'un problème qui est survenu, ce qui signifie qu'une action effectuée par la machine, et apparaissant dans les données, a beaucoup plus de chances d'appartenir à une autre catégorie comme HorsSpec, qu'à une catégorie plus neutre comme Action.

Les résultats obtenus sur certaines catégories sont assez mauvais. La catégorie Autre en est l'exemple le plus flagrant, avec une précision et un rappel de 0 pour une f1-mesure nulle également. La catégorie Autre souffre de problèmes similaires à la catégorie Action, en ce que les titres appartenant effectivement à cette catégorie sont très peu nombreux. En effet, la majorité des titres contiennent au moins une information qui permet de donner des indications sur la nature du problème, ce qui place une grande majorité de titres dans d'autres catégories. Cependant dans notre échantillon test, deux titres ont été catégorisés comme Autre par l'annotateur externe. Ceux-ci ne comportent effectivement pas d'information sur le problème, mais le fait que le mot racine est un nom commun a contribué à un placement automatique dans la catégorie Objet.

Le bas score de la catégorie Absence est dû à plusieurs facteurs, dont des négations qui ne sont pas reconnues comme telles par le programme, qui analyse « PAS » dans « PAS DE SON » comme un nom propre, ou « plus » dans « plus d eau » comme un surplus d'eau et non comme un adverbe négatif. A l'inverse, certains titres contiennent « plus » en tant que surplus sont reconnus comme des adverbes négatifs, comme dans « s'arrête au bout d'1 heure et même plus ». D'autres problèmes de rappel sont dûs aux règles qui ne prennent pas en compte d'autres marqueurs comme « manque » dans « manque pression d eau lave vaisselle ».

De plus, l'efficacité sur la catégorie HorsSpec est assez complexe à obtenir, car les marqueurs sont très variés et souvent confondus avec des marqueurs d'autres catégories moins spécifiques, ce qui explique que notre système attribuera à certains titres cette catégorie alors qu'il n'y correspond pas et inversement. Pour l'exemple « Combien coûte la réparation de mon écran », la catégorie correspondant devrait être DemandeAide, mais le verbe à la 3e personne du singulier « coûte » détecté par le système amène le titre à être classé dans HorsSpec. Pour « Commandes digitales d'un four SAUTER sfp650B inopérantes », il s'agit d'un problème de marqueurs, car la liste de marqueurs donnés au système pour le repérage de la catégorie « FonctionnePas » ne contient pas tous les marqueurs possibles et donc omet le marqueur « inopérantes ». Cependant cette catégorie obtient tout de même d'assez bons scores, même si un nombre très grand de titres classés automatiquement en HorsSpec est à modérer et à ne pas considérer comme représentatif de la réalité du corpus étudié dans les mêmes proportions.

Dans l'ensemble, le classifieur donne quand même des résultats assez bons à partir des règles données, ce qui témoigne d'un certain lien entre les catégories observées et les marqueurs proposés au système de classification automatique. Ce classifieur peut donc, tout en prenant un peu de recul par rapport aux résultats, être utilisé pour quantifier et catégoriser les titres du corpus entier.

5.2. Catégorisation du corpus entier

Afin d'avoir un aperçu des catégories d'expression du problème dans le corpus entier, nous proposons en entrée du système de classification automatique créé le fichier TSV contenant les titres du corpus entier. Cela nous permet d'avoir des informations sur le corpus comme le nombre de titres par type d'expression en fonction de la catégorie d'appareil dans laquelle le post contenant le titre a été publié. La table ci-dessous organise ces informations sous forme de matrice pour laquelle les lignes correspondent aux catégories d'appareil, et les colonnes correspondent aux types d'expression du problème comme définis dans notre typologie. Plus une case est bleue, plus la proportion de ce type d'expression est élevée dans cette catégorie d'appareil par rapport au nombre total de posts dans cette même catégorie d'appareil.

Type d'appareil \ Résultat classifieur	Fuite	Signal	Dégradation	Absence	HorsSpec	FonctionnePas	Impossible	DemandeAide	Action	Objet	Autre	**Total Résultat**
Electroménager	725	3205	331	837	10606	8538	105	5738	173	4370	83	**34711**
Audio-vidéo	5	502	279	771	3645	1964	52	2053	69	1728	27	**11095**
Electronique, informatique	7	251	402	153	2153	1378	54	1980	48	976	22	**7424**
Réparations diverses	2	60	83	44	813	459	10	781	7	299	9	**2567**
Jardinage, bricolage	75	10	38	102	898	368	16	555	16	358	4	**2440**
Plomberie-Chauffage	38	76	7	51	527	244	7	233	9	145	4	**1341**
Auto-moto	7	25	17	24	311	92	5	219	4	163	5	**872**
Mobilier, Maison	0	22	34	21	258	99	4	211	3	120	1	**773**
Vêtements, linge, bijoux	0	0	4	1	18	1	0	18	0	4	0	**46**
Total Résultat	**859**	**4151**	**1195**	**2004**	**19229**	**13143**	**253**	**11788**	**329**	**8163**	**155**	**61269**

Tableau 13 - Nombre de titres dans chaque catégorie d'expression en fonction du type

d'appareil dans le corpus entier

Parmi ces résultats, il est intéressant d'observer une tendance très claire de surreprésentation des catégories HorsSpec et FonctionnePas, mais également de la catégorie DemandeAide qui est propre à nos données. De plus, certains types d'expression du problème apparaissent beaucoup plus ou beaucoup moins pour certaines catégories d'appareils. En effet, certains types de problèmes sont spécifiques à certaines catégories d'appareils, ce que nous pouvons voir dans les valeurs élevées de Signal et Fuite dans la catégorie Electroménager. En revanche, FonctionnePas est presque absent de la catégorie Vêtements, linge, bijoux, puisqu'il ne s'agit pas d'appareils et que ces objets ne font en général pas d'action d'eux même, et l'expression d'un dysfonctionnement s'applique donc moins bien à cette catégorie. Au contraire, le type Dégradation est assez représenté dans les catégories Electronique, informatique, et Vêtements, linge, bijoux, ce qui nous donne une idée des types de problèmes associés à ces catégories.

Il est important de modérer ces résultats en se référant aux mesures de précision et de rappel pour chaque catégorie, qui nous indiquent que les valeurs de certaines catégories, comme Absence, Autre, ou dans une moindre mesure HorsSpec, ne sont pas forcément représentatives des types de titre effectivement contenus de chaque catégorie d'appareil. Cependant, nous obtenons des informations qui peuvent se révéler importantes pour des analyses futures, sur le type de problème et la façon dont ceux-ci sont exprimés dans une majorité de catégories de façon assez fiable, selon les résultats de notre échantillon test.

Dans cette dernière partie consacrée à l'évaluation et à l'utilisation de notre système de classification automatique, nous avons pu observer son comportement face à nos données et discuter de ses qualités comme de ses défauts. Malgré un système manquant encore d'une certaine fiabilité, les résultats encourageants dans un bon nombre de catégories nous a confortés dans l'idée qu'une telle approche est pertinente mais peut encore être améliorée afin d'obtenir des résultats beaucoup plus représentatifs de la réalité des données. L'observation quantitative du corpus entier annoté automatiquement par notre système montre des résultats attendus quant au type de données, mais montre également les limites du système, par exemple dans sa surreprésentation de certaines catégories.

Conclusion

Dans ce projet de recherche, nous avons tenté de créer un système de catégorisation automatique d'expressions de problèmes techniques sur un forum d'aide à la réparation. Cette étude se base sur le travail parallèle de Mariame MAAROUF concernant la catégorisation automatique d'expressions de problèmes techniques dans les fiches d'anomalie du domaine spatial, afin d'automatiser la modélisation des problèmes et la recherche de solutions dans le cadre de la méthodologie TRIZ.

Nous avons commencé par établir un état présentant les caractéristiques et les applications des fiches d'anomalie dans le cadre du processus de Retour d'Expérience, et avons également présenté la typologie de Mariame MAAROUF sur laquelle se base notre typologie. Les spécificités de notre type de données, étant des données issues d'internet et entrant dans le cadre de la communication médiée par les réseaux, ou CMR, sont également abordés sous différents points de vue dans cet état de l'art.

Pour mener à bien notre étude, nous choisissons de créer un corpus contenant des expressions de problèmes qui pourraient être similaires aux données de Mariame MAAROUF, puisque pour des raisons de confidentialité et de sécurité, il est délicat d'utiliser les données originales venant de fiches d'anomalie du domaine spatial. La création du corpus, ses problématiques à différents niveaux, que ce soit, linguistique, mais aussi social, éthique, juridique, mais également technique et méthodologique, ont été abordées. Le corpus créé, baptisé CoCoRep, et composé du premier message de chaque fil de discussion sur le site commentreparer.com, est utilisé ensuite pour l'élaboration du système de classification automatique.

Les données que nous utilisons étant différentes, nous avons adapté la typologie de base pour en créer une nouvelle spécialement adaptée à nos types de données. Bien que la visée de l'étude soit similaire à celle de l'étude parallèle, les spécificités des données amènent de nouvelles problématiques différentes de celles relatives au travail sur des données appartenant à une agence spatiale.

La typologie créée a ensuite donné lieu à un guide d'annotation permettant de faire annoter les données par un annotateur externe. D'un autre côté, nous avons créé un système de classification automatique à base de règles et de repérage de marqueurs que nous avons évalué sur les données catégorisées par notre annotateur externe.

La différence entre les données de base et les données de CMR ont pu poser quelques difficultés, sur le plan technique comme sur un plan plutôt méthodologique. Sur le plan technique, la lemmatisation et l'annotation morphosyntaxique automatique des outils Stanza est imparfaite sur ce type de données, car les modèles sont entraînés sur des données ne suivant pas les mêmes codes et caractéristiques. L'utilisation de règles plutôt que d'un système d'apprentissage automatique à partir d'un certain nombre de données aurait pu également permettre une meilleure efficacité de notre système de catégorisation automatique.

Bien sûr, le nombre de titres annotés, plutôt faible, constituerait également un obstacle pour un apprentissage automatique fiable, mais aussi pour la formulation de règles formelles plus robustes pour la catégorisation automatique, car il pourrait être notamment possible de s'aider d'un processus automatisé pour extraire des exemples plusieurs règles utilisables par le classifieur. De plus, la définition des catégories de notre typologie reste floue sur certains points et bénéficierait notamment d'une prise en compte plus complète d'exemples et de marqueurs supplémentaires pour obtenir un accord parfait entre annotateurs.

Pour améliorer un tel système de classification automatique, il pourrait être intéressant de pouvoir identifier de façon automatique quelle partie du texte correspond à quelle information, et partir de cette identification pour construire nos règles. Les règles que nous avons formulé se basent sur des indices syntaxiques et sémantiques mais ne sont pas suffisantes pour repérer précisément ce

que comprennent les titres sur un plan sémantique. Il pourrait être également intéressant de prendre en compte d'autres champs en compte dans l'analyse, comme la marque ou le modèle de l'appareil, afin d'améliorer la reconnaissance des informations du texte et donc la classification.

Malgré ces limites et améliorations possibles, le travail nous a notamment permis d'apprendre à faire la passerelle entre la définition de la typologie et l'automatisation de la classification à partir de celle-ci, tâche assez complexe et demandant plus de recherche afin de perfectionner la technique et la méthodologie suivie.

Le corpus que nous avons construit, ou des données similaires, pourraient être étudiés plus en profondeur et sous différents angles, en prenant en compte d'autres champs de texte, pour une étude plus complète de l'expression d'un problème. Il pourrait être également intéressant de prendre en compte les autres messages du fil de discussion et l'interaction entre les différents locuteurs autour d'un problème. Une étude portée sur d'autres particularités des CMR comme la multimodalité est également possible avec le type de données que nous avons récolté. En effet, sur le forum, il est possible d'accompagner son expression du problème d'un document ou d'une image, ce qui ajoute des indices à la compréhension du problème et constitue une autre facette de l'expression du problème. Nous pensons que l'étude linguistique de l'expression des problèmes peut nous aider à comprendre et à prévenir ces problèmes de façon plus consistante, et ceci à une échelle de plus en plus grande à l'ère d'internet et de la communication médiée par les réseaux.

Bibliographie

Amato, S., & Boutin, E. (2013). Rites d'interaction et forums de discussion en ligne. Les Cahiers Du Numérique, 9(3), 135–159. https://doi.org/10.3166/LCN.9.3-4.135-159

Ansoff, H. I. (1975). Managing Strategic Surprise by Response to Weak Signals. *California Management Review*, *18*(2), 21‑33. https://doi.org/10.2307/41164635

Blatter, C. & Raynal, C. (2015). Méthodes d'analyse textuelle pour l'interprétation des REX humains, organisationnels et techniques. *Congrès Lambda Mu 19 de Maîtrise des risques et sûreté de fonctionnement*. Dijon, France. https://doi.org/10.4267/2042/56070

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C. R., Hriba, L., Longhi, J., & Seddah, D. (2014). The CoMeRe corpus for French : structuring and annotating heterogeneous CMC genres. Journal for Language Technology and Computational Linguistics, 29(2), 1–30. https://doi.org/10.21248/jlcl.29.2014.187

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 37‑46. https://doi.org/10.1177/001316446002000104

Condamines, A. & Vergely, P. (2001). Expression du dysfonctionnement dans un corpus dialogique de la Navigation Aérienne : Mise au jour de régularités. *Terminologie et intelligence artificielle. Rencontres*, 22‑32.

Gaillard, I. (2005). Facteurs socio-culturels de réussite du REX industriel par l'analyse bibliographique. *Cahiers de la Sécurité Industrielle*, *2008*(01). Fondation pour une Culture de Sécurité Industrielle, Toulouse, France.

Galand, L., Kurela, M., Clavijo, H.R. (2018). Techniques de TAL pour la Recherche des « Signaux Faibles » et Catégorisation des Risques dans le REX SDF des Lanceurs Spatiaux. *Congrès Lambda Mu 21 de Maîtrise des risques et transformation numérique : opportunités et menaces*. Reims, France.

Ghliss, Y., & André, F. (2017). Après la collecte, l'anonymisation : enjeux éthiques et juridiques dans la constitution du corpus 88MILSMS. HAL (Le Centre Pour La Communication Scientifique Directe). https://hal.archives-ouvertes.fr/hal-01722169/document

Heiden, S., Magué, J. & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, *2*(3), 1021‑1032.

Ilevbare, I. M., Probert, D. & Phaal, R. (2013). A review of TRIZ, and its benefits and challenges in practice. *Technovation*, *33*(2‑3), 30‑37. https://doi.org/10.1016/j.technovation.2012.11.003

Krotov, V., Johnson, L., & Silva, L. (2020). Legality and Ethics of Web Scraping. Communications of the Association for Information Systems, 47, 539‑563. https://doi.org/10.17705/1cais.04724

Mondada, L. (1999). Formes de séquentialité dans les courriels et les forums de discussion. Une approche conversationnelle de l'interaction sur Internet. Alsic, 2(1). https://doi.org/10.4000/alsic.1571

Nejad, M. B., Golshan, M., & Naeimi, A. (2021). The effect of synchronous and asynchronous computer-mediated communication (CMC) on learners' pronunciation achievement. Cogent psychology, 8(1). https://doi.org/10.1080/23311908.2021.1872908

Nivre, J., De Marneffe, M., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F. M., & Zeman, D. (2020). Universal Dependencies v2 : An Evergrowing Multilingual Treebank Collection. *arXiv (Cornell University)* (p. 4034‑4043). Cornell University. https://arxiv.org/pdf/2004.10643

Poudat, C., Wigham, C. R. & Liégeois, L. (2020). Les corpus de la communication médiée par les réseaux : une introduction. *Corpus*, *20*. https://doi.org/10.4000/corpus.4720

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza : A Python Natural Language Processing Toolkit for Many Human Languages. https://doi.org/10.18653/v1/2020.acl-demos.14

Romary, L. & Hudrisier, H. (2002). TEI – Text encoding initiative. *Études et Documents Berbères*, *N° 19–20*(1), 319‑322. https://doi.org/10.3917/edb.019.0319

Romiszowski, A. & Mason, R. (2004). Computer-mediated Communication. *Handbook of Research on Educational Communications and Technology*, 2. Routledge.

Tanguy, L., Tulechki, N., Urieli, A., Hermann, E. & Raynal, C. (2016). Natural language processing for aviation safety reports : From classification to interactive analysis. *Computers in Industry*, *78*, 80‑95. https://doi.org/10.1016/j.compind.2015.09.005

Annexes

Annexe 1 : Guide d'annotation pour la catégorisation des titres du corpus CoCoRep

Dans ce guide d'annotation, nous présentons les consignes de catégorisation des titres du corpus CoCoRep, contenant des posts sur le forum d'aide à la réparation commentreparer.com. Dû à la grande ressemblance de ces posts avec les fiches d'anomalies que nous pouvons trouver dans le contexte de Retours d'EXpérience (REX) de domaines techniques, nous nous basons sur certaines catégories définies dans le guide d'annotation de Mariame MAAROUF, que nous détaillerons ici et adapterons à notre contexte de recherche. Cependant, puisque les posts contenus dans ce corpus sont rédigés dans un but différent et à destination d'un public différent que dans un contexte de REX industriel, de nouvelles catégories spécifiques à nos données apparaissent, lesquelles sont absentes des catégories proposées dans le guide d'annotation mentionné ci-dessus. Pour cette raison, nous proposons également des catégories supplémentaires adaptées à un tel contexte de communication médiée par les réseaux dans ce guide.

Pour la catégorisation, il conviendra de se baser uniquement sur le texte contenu dans le titre du post, et ne pas inférer de contexte à partir d'autres informations comme les métadonnées où le jugement de l'annotateur. Chaque titre se verra attribuer une étiquette correspondant à une catégorie. La liste des étiquettes est présentée dans le tableau ci-dessous. Dans le cas où un titre peut correspondre à plusieurs catégories, il conviendra d'utiliser l'étiquette la plus haute dans le tableau ci-dessous parmi les catégories correspondant au titre (ex: Si un titre peut correspondre à FonctionnePas, mais également à Fuite, l'étiquette à utiliser sera Fuite car celle-ci est plus haute dans le tableau).

Catégorie	Etiquette
Fuite	Fuite
Signal qui s'est déclenché	Signal
Dégradation - Usure - Saleté	Dégradation
Objet manquant (dispositif, document, outil)	Absence
Hors spécification	HorsSpec
Dispositif qui ne fonctionne pas	FonctionnePas
Action difficile ou impossible	Impossible
Demande d'aide à la réparation	DemandeAide
Action effectuée par l'appareil	Action
Objet ou composant seul	Objet
Autre type d'expression	Autre

Tableau 1 : Catégories d'expressions du problème et étiquettes correspondantes

1. Fuite (Fuite) :

La catégorie Fuite correspond à la présence d'une fuite, ou d'un écoulement d'une substance. Ne correspondent à cette catégorie que les titres pour lesquels il est question d'une fuite, et non du résultat de la fuite sous forme de tâche ou de dégradation. Les marqueurs typiques de cette catégorie sont : « fuite », « écoulement » …

Exemples :

- Fuite sur pompe filtration leroy somer LS71P
- Problème WC suspendu : écoulement d'eau permanent

Contre-exemple : Réparer pompe immergée de puit (appartient à DemandeAide car nous n'avons pas d'information explicite sur une fuite)

2. Signal qui s'est déclenché (Signal) :

La catégorie Signal correspond au déclenchement d'un signal visuel ou sonore, à l'affichage d'un code d'erreur, à l'allumage d'un témoin. Le problème qui engendre le déclenchement de ce signal n'est pas décrit. Les marqueurs typiques de cette catégorie sont : « signal », « alarme », « témoin », « code », « erreur » …

Exemples :

- ERREUR E33 sur seche linge CURTISS msc 70 ebdi
- Arrêt en cours de cycle code erreur A5 lave-vaisselle Ariston
- Lave-vaisselle far LV1614 bipe au démarrage programme « normal » clignote sans cesse et rien ne

Contre-exemple : Comment réparer ma télé qui siffle et s'éteint tout seul? (appartient à HorsSpec car le sifflement ne correspond pas à un signal marquant une erreur, mais plutôt une action imprévue de l'appareil)

3. Dégradation - Usure - Saleté (Dégradation) :

Correspond à une constatation de dégradation, de saleté, ou d'usure. Entrent dans cette catégorie les titres dans lesquels il y a mention d'un composant cassé, déchiré, abîmé, mais également la présence de tâches, de corrosion, d'insectes etc. Les marqueurs typiques sont : « cassé », « déchiré », « abîmé », « corrosion », « tâche » …

Exemples :

- Poignée compartiment congelateur s'est cassé
- Technics SC-EH750 qui à les têtes de lecture sale et révision tuner radio
- Rupture du fil supérieur sur la brodeuse 955

Contre-exemple : le tiroir cassette reste ouvert ne se ferme plus (appartient à FonctionnePas car on ne mentionne pas qu'un composant est cassé, mais plutôt le dysfonctionnement résultant)

4. Objet manquant - dispositif, document, outil (Absence) :

Un objet qui devrait exister à un endroit où être fourni est manquant ou a été déplacé et est introuvable, ou une fonctionnalité n'est plus présente. Les marqueurs sont : « sans », « absent », « pas de … » …

Exemples :

- Pas de son
- Jbl charge 3 : plus de bass
- Perte de signal hdmi

Contre-exemple : Télevison continental edison le son ne fonctionne plus (appartient à FonctionnePas car on ne qualifie pas le son par son absence, mais par son dysfonctionnement)

5. Hors spécification (HorsSpec) :

Evocation d'un état indésirable en comparaison avec un état attendu, le type d'écart n'est pas spécifié mais la situation est décrite comme incorrecte et l'écart résultant n'est pas une situation normale dans les fonctionnalités de l'appareil. Les marqueurs peuvent être « mauvais », « mal », « au lieu de » etc., mais il peut également y avoir la description d'une action ou d'un état indésirable ou inattendu sans ces marqueurs, ou un marqueur général de type « problème » lorsque le problème n'est pas spécifié.

Exemples :

- Mauvaise connectique de mon pc Acer aspire 7735zg
- Problème régime moteur aspirateur Miele
- Aspirateur Siemens VS 07 qui s'arrête souvent en cours de travail

Contre-exemple : Platine vynile Technics : moteur faible (appartient à FonctionnePas car on ne parle pas d'un moteur qui fait quelque chose d'inattendu, mais d'un moteur qui fait moins bien une action attendue)

6. Dispositif qui ne fonctionne pas (FonctionnePas) :

Un dispositif ou un objet qui ne fonctionne pas parce qu'il est en panne ou défectueux. La raison du problème n'est pas explicitée, il n'y a pas d'information sur la façon dont l'appareil dysfonctionne, juste qu'il ne fonctionne pas. Certains marqueurs peuvent être : « HS », « en panne », « ne fonctionne pas », « bloqué »…

Exemples :

- Mon pavé souris tactile ne fonctionne plus
- Sèche linge Whirlpool minuterie bloqué
- Réparer Beko DC713 qui ne chauffe plus

Contre-exemple : Mon four électrique Ariston se met en route tout seul. (appartient à HorsSpec car il n'est pas question d'une action ou d'une fonctionnalité qui ne se fait pas, mais plutôt de l'apparition d'un comportement inattendu)

7. Action difficile ou impossible (Impossible) :

Correspond à une action impossible ou difficile à effectuer. La raison du problème n'est pas explicitée, le titre ne présente qu'un simple constat d'impossibilité d'effectuer la tâche. Les marqueurs peuvent être : « impossibilité », « difficulté », « dur », « ne peux pas », « echec » …

Exemples :

- Impossible de MAJ windows Update
- Je ne peux pas appeller à partir de l'appareil - mais on peut m'appeller
- Impossible de choisir un programme après la mise sous tension avec le

tableau tactile

Contre-exemple : Pas de fonction detartrage broyeur krups intuition EA870 (appartient à Absence car l'attention est portée sur l'absence de fonction plutôt que l'impossibilité d'effectuer l'action associée)

8. Demande d'aide à la réparation (DemandeAide) :

La catégorie « DemandeAide » correspond aux titres commençant par « réparer » ou un terme équivalent (« démonter », « changer », « réparation » etc.) suivi (et, de façon beaucoup moins fréquente, précédé) du nom de l'appareil défectueux, de sa marque, ou du composant défectueux. Si le problème est décrit, la catégorie correspondante au problème sera attribuée au titre. Des marques interrogatives peuvent être présentes (exemple : « comment … »).

Exemples :

- Réparer lisseur babyliss Modele ST95E
- Demontage Aspirateur Bosch Relaxx's pour nettoyage Capteur
- Comment accéder au moteur sur un rowenta X-TREM POWER.

Contre-exemple : Comment réparer ma télé qui siffle et s'éteint tout seul? (appartient à HorsSpec car le problème est décrit, et ce problème appartient à la catégorie HorsSpec qui se trouve plus haut dans la hiérarchie des catégories)

9. Action (Action) :

La catégorie « Action » correspond à un titre ne mentionnant qu'une action effectuée par la machine ou l'appareil défectueux, l'action étant neutre et ne supposant pas un dysfonctionnement ou un blocage.

Exemples :

- Mise à jour android
- Mise en veille automatique de ma tv

Contre-exemple : Mise en veille automatique de ma tv sans que je ne lui demande (appartient à HorsSpec car l'action est décrite comme inattendue avec « sans que je ne lui demande »)

10. Objet (Objet) :

La catégorie « Objet » correspond aux titres ne mentionnant que le nom de la machine, la marque, le nom du composant défectueux, ou une combinaison de ces trois indices.

Exemples :

- TV Toshiba
- Bouton platine vinyle
- Batterie

Contre-exemple : Réparer televiseur pilips (appartient à DemandeAide car en plus de l'objet, il y a un verbe indiquant une demande d'aide)

11. Autre type d'expression (Autre) :

La catégorie Autre type d'expression ne s'applique qu'aux titres qui ne trouvent leur place dans aucune catégorie. Il s'agit d'une expression qui ne transcrit pas un problème, ni une action, ni un objet, mais apparaît tout de même en titre d'un post.

Nous utilisons donc cette catégorie selon sa définition comme indiquée ci-dessus, mais également comme catégorie « par défaut » dans laquelle seront classés les titres qui ne présentent aucune marque existante dans une autre catégorie. Les titres accrocheurs présentant un seul mot comme « Urgent » ou « Vite ! », ou les marques de politesse seules comme « Bonjour » ou « S'il vous plaît » correspondent à cette catégorie.

Exemples :

- Bonjour
- Quelle panne..?
- URGENT !

Contre-exemple : Bonjour j'ai un problème ma télé est en panne (appartient à FonctionnePas, malgré la présence de marques de politesse car le problème est décrit et il correspond à une panne)

UNIVERSITÉ TOULOUSE Jean Jaurès

SCIENCES DU LANGAGE

Déclaration sur l'honneur de non-plagiat

(à joindre au mémoire à la fin du document)

Je soussigné.e,

Nom, Prénom : CASSAM SULLIMAN Shaad
Régulièrement inscrit.e à l'Université de Toulouse II Jean Jaurès
N° étudiant : 21700514

Année universitaire : 2022-2023

certifie que le document joint à la présente déclaration est un travail original, que je n'ai ni recopié ni utilisé des idées ou des formulations tirées d'un ouvrage, article ou mémoire, en version imprimée ou électronique, sans mentionner précisément leur origine et que les citations intégrales sont signalées entre guillemets.

Fait à : Toulouse

Le : 16 juin 2023

Signature :