RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Doctorat de l'Université de Toulouse

préparé à l'Université Toulouse - Jean Jaurès

Approches du dysfonctionnement technique dans les REX d'Ariane 5 : de l'analyse linguistique outillée de son expression vers la modélisation TRIZ du problème

---

Thèse présentée et soutenue, le 16 octobre 2025 par

Mariame MAAROUF

École doctorale
CLESCO - Comportement, Langage, Éducation, Socialisation, Cognition

Spécialité
Sciences du langage

Unité de recherche
CLLE - Unité Cognition, Langues, Langage, Ergonomie

Thèse dirigée par
Ludovic TANGUY

Composition du jury
Mme Amalia TODIRASCU, Présidente, Université de Strasbourg
M. Ahmed SAMET, Rapporteur, INSA Strasbourg
Mme Anne CONDAMINES, Examinatrice, Université Toulouse - Jean Jaurès
Mme Haïfa ZARGAYOUNA, Examinatrice, Université Sorbonne Paris Nord
M. Ludovic TANGUY, Directeur de thèse, Université Toulouse - Jean Jaurès

Membres invités
M. Michal KURELA, CNES
M. Jérôme LAFORCADE, MeetSYS

Table des matières

Liste des tableaux

Table des figures

Remerciements

En premier lieu, je tiens à remercier mon directeur de thèse Ludovic Tanguy pour son accompagnement et son soutien tout au long de ce travail. Ces quatre années ne peuvent pas être qualifiées de long fleuve tranquille, mais tu as toujours su te montrer disponible et impliqué. Tu m'as aussi appris à penser plus loin, à approfondir et à affiner mes réflexions tout en étant toujours ouvert à la discussion et au débat.

Je souhaite aussi remercier les autres membres partenaires de cette thèse : Daniel Galarreta et Michal Kurela du côté du CNES et Jérôme Laforcade de MeetSYS qui ont tous fait preuve de curiosité, de patience et d'ouverture d'esprit face au chemin qu'a pris ce travail. Ils ont aussi su se montrer disponible dès que le besoin sans faisait sentir, et ce même quant il fallait m'expliquer une énième fois ce que pouvait bien vouloir dire des mots aussi ésotériques que « EAP » ou bien « vépole ».

Je remercie les rapporteurs de mon jury de thèse, Amalia Todirascu et Ahmed Samet, pour avoir pris le temps de lire cette thèse et pour leurs retours sur mon manuscrit. Je remercie aussi Haïfa Zargayouna et Anne Condamines d'avoir accepté d'être les examinatrices de cette thèse, mais aussi pour leur accompagnement en tant que membres du CSI (auquel figurait aussi Daniel Galarreta). J'ajoute une mention particulière à Michal Kurela et Jérôme Laforcade qui ont aussi participé à ce jury en tant que membre invité. Les échanges lors de cette soutenance ont été enrichissants.

Je souhaite également remercier les trois institutions partenaires de cette : le laboratoire CLLE, le CNES et MeetSYS. Ils ont apporté un soutien financier, matériel, mais aussi scientifique, et cette thèse n'aurait pas existé sans eux et leur volonté de mener à ben ce projet.

L'expérience de doctorat n'a pas été sans heurts, mais elle a aussi été nettement adoucie, enrichie et ensoleillée grâce aux belles rencontres que j'ai pu faire parmi les doctorants avec qui j'ai partagé cette expérience.

Je tiens donc à remercier Claire pour tout le soutien et pour toutes les fois où on a ri, pleuré et râlé ensemble. Je souhaite remercier Clamença, ma partenaire de squash, mais surtout une amie avec qui j'ai partagé plein d'escapades (mention spéciale à nos sorties spa). J'adresse un remerciement tout particulier à mes deux cobureaux, Killyam et et Lockman qui m'ont supporté pendant quatre ans malgré tout le bruit et le bazar que j'ai mis dans ce bureau.

J'attends toujours qu'on se fasse nos week-ends à la plaine ceci dit. Je souhaite remercier aussi Victoria, Silvia, Florian et Alba qui m'ont aidé à me sentir accueillie dans cette ville que je découvrais à l'occasion de cette thèse et tous les autres doctorants, post-docs et amis avec qui j'ai partagé tant de bons moments : Kévin, Daniele, Valentin, Maxime, Malvina, Dimitri, Brivael et Élodie.

J'adresse mes remerciements à ma famille. Tout d'abord à mes parents qui n'ont jamais douté que j'arriverai à la fin de cette thèse. Aussi à mes frères, Moncef et Sami, pour le soutien et pour les blagues. Et mention spéciale à la bande des cousins : à quand un nouveau week-end à Toulouse ?

Je souhaite aussi remercier mes amis qui m'ont soutenu pendant ces quatre années, parfois de près et parfois de loin. J'adresse ma gratitude à Rémy pour son soutien durant toute cette fin de thèse. Je remercie Océane pour m'avoir accueillie à Toulouse, au CLLE, pour les conseils toujours avisés et l'aide qu'elle m'a prodigué pendant toute cette thèse et encore aujourd'hui, mais aussi pour tous les fou-rires qu'on a partagés. Je remercie aussi Héloïse qui a toujours su me remonter le moral et me booster quand il fallait. Et enfin, je souhaite remercier Barbara. Je ne saurai même pas dire pourquoi parce qu'à ce stade, qu'est-ce qu'on n'a pas partagé ?

Résumé

Les REX (Retours d'EXpérience) sont des documents textuels dont la visée est de rapporter un problème, ou un dysfonctionnement, et qui jouent un rôle important dans la maîtrise des risques au sein d'une organisation. Plusieurs travaux de TAL (Traitement Automatique des Langues) ont donc vu le jour afin de capitaliser les connaissances qu'ils abritent. Par ailleurs, des méthodes de résolution de problèmes techniques ont été développées, comme la méthode TRIZ, et présentent un intérêt non négligeable pour les dysfonctionnements qui peuvent être rapportés dans les REX. De ce fait, un partenariat s'est créé entre le CNES qui cherche à exploiter ses REX liés aux lanceurs spatiaux, et la société MeetSYS, spécialisée dans l'utilisation de la méthode TRIZ pour la capitalisation du savoir expert. Cette thèse s'est vue comme l'opportunité d'explorer l'utilisation du TAL et de la linguistique de corpus pour l'extraction fine d'information dans les REX d'Ariane 5 en vue de modéliser un dysfonctionnement technique sous forme de *vépole* (formalisme propre à la méthode TRIZ). Cela signifie être capable de partir d'un texte brut, spécialisé et bruité vers un formalisme conçu indépendamment des données en question. À cette fin, une démarche en plusieurs étapes a été mise en place en vue de se rapprocher autant que possible de ce formalisme. L'un des piliers sur lequel s'appuie cette démarche est la sémantique des cadres, et la ressource *FrameNet* qui en découle, qui nous permet d'identifier et de qualifier les éléments textuels qui constituent le problème.

Nous explorons dans cette thèse plusieurs approches de TAL et de linguistique de corpus dans l'étude des REX, soit des textes spécialisés et bruités, afin d'identifier les structures sémantiques qui composent l'expression d'un dysfonctionnement technique. Nous mêlons ainsi diverses techniques comme le *Topic Modeling*, *word2vec* et de l'analyse lexicale outillée pour de l'exploration de corpus, du *fine-tuning* de modèles neuronaux pour de l'étiquetage automatique, l'utilisation de LLMs pour de la normalisation et de l'annotation automatique, mais aussi de l'analyse syntaxique et de la reconnaissance de patrons pour l'analyse fine des structures langagières.

Dans un premier temps, une première analyse du corpus nous a permis de dégager une typologie d'expressions d'un dysfonctionnement technique en neuf classes. Elle est basée sur la détection de marqueurs lexicaux au sein de la description de l'anomalie qui a été repérée et décrite. À partir de cette typologie, nous avons pu effectuer une annotation

des marqueurs lexicaux spécifiques au sein du corpus. Celle-ci nous a permis d'explorer l'utilisation d'annotateurs non experts du domaine sur des données spécialisées et, par la suite, d'entraîner un modèle neuronal à base de *transformers* pour l'étiquetage automatique des rapports d'anomalies. Nous avons aussi mené une étude afin de normaliser automatiquement ces rapports pour en supprimer le bruit, avant de tester l'impact de cette normalisation sur l'entraînement du modèle. Cette étude n'ayant pas montré d'améliorations sur la tâche d'étiquetage automatique nous entraîne à interroger la pertinence de la normalisation des données bruitées, et notamment en fonction de la tâche visée.

Par la suite, nous avons pu focaliser notre étude sur deux catégories de la typologie qui sont la *Fuite d'un liquide ou d'un gaz* et la *Présence d'un obstacle*. Pour la première, nous avons mis en place une approche impliquant plusieurs méthodes complémentaires de linguistique de corpus afin de faire émerger un *frame de la fuite dans un environnement technique*. Nous avons ainsi pu identifier les différents éléments qui composent l'expression de la fuite. Pour la catégorie *Présence d'un obstacle*, nous avons utilisé des LLMs génératifs pour l'annotation automatique de ces textes. Par ce biais, nous avons pu explorer les capacités et les limites d'un LLM à effectuer une annotation de type *Frame Semantic Role Labeling*, mais aussi à traiter un texte spécialisé et bruité.

Abstract

Anomaly reports are textual documents whose purpose is to report on a problem or malfunction, play an important role in risk management within an organisation. A number of NLP (Natural Language Processing) projects have therefore been developed to capitalise on the knowledge they contain. In addition, methods for solving technical problems have been developed, such as the TRIZ method, which are of considerable interest for the type of malfunction that can be reported in incident or anomaly reports. As a result, a partnership has been formed between CNES, which is seeking to exploit these reports, and MeetSYS, a company specialised in using the TRIZ method to capitalise on expert knowledge. This thesis was seen as an opportunity to explore the use of NLP and corpus linguistics for the fine-grained extraction of information from the Ariane 5 anomaly reports in order of modeling a technical malfunction in the form of a *vepole* (a formalism specific to the TRIZ method).This means being able to move from a raw, specialised and noisy text to a formalism designed independently of the data in question. To this end, a multi-stage approach has been put in place to get as close as possible to this formalism. One of the pillars on which this approach is based is frame semantics, and the resulting *FrameNet* resource, which allows us to identify and qualify the textual elements that constitute the problem.

In this thesis, we explore several NLP and corpus linguistic approaches in the study of the reports, i.e. specialised and noisy texts, in order to identify the semantic structures that make up the expression of a technical malfunction. We thus combine various techniques such as *Topic Modeling*, *word2vec* and tool-based lexical analysis for corpus exploration, *fine-tuning* of neural models for automatic labelling, the use of LLMs for normalisation and automatic annotation, as well as syntactic analysis and pattern recognition for fine-grained analysis of language structures.

An initial analysis of the corpus enabled us to identify a nine-class typology of expressions of technical malfunction. It is based on the detection of lexical markers within the description of the anomaly that has been identified and described. Using this typology, we were able to annotate the specific lexical markers within the corpus. This enabled us to explore the use of non-domain expert annotators on specialised data and, subsequently, to train a neural model based on *transformers* for the automatic labelling of anomaly reports. We also conducted a study to automatically normalise these reports to remove noise, before

testing the impact of this normalisation on the training of the model. Since this study did not show any improvement in the automatic labelling task, we are in this occasion questioning the relevance of normalising noisy data, particularly in relation to the target task.

These different approaches have enabled us to identify a nine-class typology for the expression of a technical malfunction. It is based on the detection of lexical markers within the description of the anomaly that has occurred. Using this typology, we were able to annotate markers within the corpus. This enabled us to explore the use of non-expert annotators on specialised data and, subsequently, to fine-tune a model for the automatic labelling of anomaly reports. We also conducted a study to automatically normalise these reports to remove noise, before testing the impact of this normalisation on model fine-tuning. As this study did not show any significant improvement in the automatic labelling task, it pushed us to question the relevance of normalising noisy data, particularly as a function of the task in question.

We then focused our study on two categories of the typology : *Leakage of a liquid or a gas* and *Presence of an obstacle*. For the former, we implemented an approach involving several complementary corpus linguistic methods in order to bring out *frame of the leak in a technical environment*. We were thus able to identify the different elements that make up the expression of leakage. For the category *Presence of an obstacle*, we explored the use of generative LLMs for the automatic labeling of these texts. In this way, we were able to explore the capabilities and limitations of an LLM for performing *Frame Semantic Role Labeling* type annotation, as well as processing specialised and noisy text.

Introduction

Contexte et problématiques de recherche

La thèse que nous présentons ici vient en réponse à la fois à des problématiques de recherche, mais aussi à un besoin industriel. Elle se présente comme l'opportunité d'explorer l'interconnexion entre trois domaines distincts que sont : la maîtrise du risque dans un environnement industriel, la résolution de problèmes techniques et la linguistique de corpus et le TAL (Traitement Automatique des Langues). Cette thèse est elle-même issue d'une volonté d'approfondir cette relation par trois acteurs que sont le CNES, MeetSYS et CLLE.

Le CNES (Centre National d'Études Spatiales) est engagé depuis plusieurs années dans des actions de capitalisation de connaissances. Dans le cadre de ce projet, elles concernent les Retours d'Expériences (REX) rédigés dans le cadre d'exploitation d'Ariane 5, incluant les campagnes de lancement et les opérations de maintenance. MeetSYS est une entreprise qui intervient au sein d'organisations industrielles afin de mettre à profit les connaissances expertes. Ils sont spécialisés dans la résolution de problèmes techniques et les stratégies d'innovation par le biais de la méthode TRIZ (Théorie de Résolution des Problèmes Inventifs), que nous présentons au chapitre 2. Ils identifient, au sein de l'organisation en question, les modèles théoriques utilisés par les experts et organisent ces connaissances en réseau. Le laboratoire CLLE propose une expertise en linguistique de corpus et en TAL, mais aussi dans les travaux sur des corpus spécialisés.

Cette étude vise donc à explorer comment et dans quelle mesure la linguistique et le TAL peuvent permettre de faire le lien entre des REX rédigés dans un environnement industriel et à risque, et une formalisation d'un problème technique telle que définie dans la méthode TRIZ.

Les REX sur lesquels nous avons travaillé sont des rapports d'anomalies, nommés Fiches Anomalies (FA). Ils visent à rendre compte d'écarts à l'attendu, quel que soit leur apparent degré de gravité. Le REX est aussi plus globalement une pratique systémique et systématique de documentation des anomalies tel qu'il en est l'usage dans les organisations industrielles des domaines à risques. En revanche, les documents auxquels nous avons accès, et qui constituent notre corpus d'étude, sont des versions minimalistes de ces FA. Les rapports avec

lesquels nous travaillons prennent la forme de phrases descriptives courtes, avec parfois un diagnostic absent et des données peu renseignées. Nous pouvons prendre un des rapports du corpus en exemple :

> LED DE BON FONCTIONNEMENT ETEINTESSUR LES CHARGEURS, DU BABAR, Nr 6, 9, 10, 13, 17 ET 18, LES LED "MARCHE", "CHARGE", "25VR16", EN FACE AVANT DE LA CARTE SEQ (DES CHARGEURS), NE SONT PAS ALLUMES ALORS QUE LA CHARGE FINALE EST LANCEE.SUR LES AUTRES CHARGEURS , Nr1, 2, 5 ET 14, TOUT EST NOMINAL.

Nous avons choisi ici de sélectionner uniquement le champ du rapport qui concerne la description de l'anomalie, qui est celui sur lequel nous avons concentré notre étude. Nous pouvons noter deux éléments qui caractérisent le texte que sont le vocabulaire spécialisé et le bruit textuel (texte en majuscule, absence d'accents...). Ce sont des caractéristiques récurrentes dans ce type de texte et, par conséquent, dans nos travaux, qu'il aura fallu prendre en compte dans l'établissement de notre méthode de traitement.

Parmi les études de linguistique et de TAL pour les textes spécialisés, deux ressources essentielles sont communément mises à profit : les ressources langagières métiers et le recours à des experts. Toutefois, les ressources langagières supplémentaires concernant ces données n'existent pas. En effet, les FA avec lesquelles nous travaillons couvrent des sous-domaines variés, leur point commun étant principalement leur zone géographique : la base de lancement guyanaise du lanceur Ariane 5. Par ailleurs, le corpus couvre une période de plus de 10 ans (de 1998 à 2011), et certaines conventions, acronymes et autres codes ont évolué au fil du temps. De ce fait, il n'existe pas de ressources couvrant toute l'étendue de ces données. Par rapport aux experts, leur implication dans la thèse a permis d'éclaircir de nombreux phénomènes langagiers que nous avons rencontrés. En revanche, leur disponibilité a aussi été limitée et une participation d'ampleur telle que la mise en place d'une campagne d'annotation a été impossible. Ces deux éléments ont donc aussi participé à l'élaboration de la méthode que nous avons mise en place, qui contourne l'absence de ressources langagières par une approche basée reconnaissance de patrons langagiers, et fait usage d'une annotation experte de façon limitée.

L'objectif industriel de cette thèse nous permet d'aborder un certain nombre de thématiques de recherche. En effet, le traitement automatique de données textuelles techniques est un domaine de recherche encore actif. Les outils actuels sont toujours peu adaptés à ce type de textes et génèrent des problématiques de recherche, malgré l'évolution des systèmes de TAL et les performances des modèles récents comme les *transformers* et les LLM (Large Language Models), que nous avons eu tous deux l'occasion de mettre en pratique.

Les caractéristiques des données et du contexte de recherche permettent aussi de soulever des problématiques de recherche liées au travail sur des données de domaine spécialisé. En effet, nous explorons dans cette thèse, l'étude de données en langage spécialisé en l'absence de ressources langagières conjoint à des difficultés d'accès à du temps d'experts peu. Nous faisons donc une proposition méthodologique autour de la question de l'étude de textes techniques et bruités en contournant ces écueils.

Démarche d'étude

L'objet de cette thèse est le passage d'un texte brut, spécialisé et bruité vers une formalisation qui implique un certain degré d'abstraction. Cette abstraction permet d'obtenir en définitive des solutions génériques au problème énoncé, grâce à la généralisation des composants du dysfonctionnement technique. Cela nécessite à la fois de réussir à identifier et à qualifier les éléments précis du texte, mais aussi de pourvoir les généraliser. Pour réduire la distance qui sépare ces deux pôles, nous avons adopté une approche basée sur la linguistique de corpus, tout en incorporant des techniques de TAL actuelles. Cette démarche vise donc de passer d'un texte brut à un texte analysé, vers une représentation basée sur la sémantique des cadres. Cette même représentation nous sert de point d'accès vers une modélisation du contenu des énoncés afin de formaliser les données avec la méthode TRIZ.

La question qui sous-tend l'entièreté de cette thèse est de savoir comment aborder des données aussi particulières de par leur nature de REX, qui implique un format et un objectif de communication spécifique, leur domaine, visible dans le lexique de spécialité omniprésent, et le bruit qu'elles contiennent, impactant les performances d'outils et donc d'une chaîne de traitement. Pour cela, nous avons choisi de focaliser notre étude sur la forme du texte sous plusieurs angles, et de questionner les outils traditionnels afin de les employer à des fins d'exploration de corpus, plutôt que de comme producteurs de résultats à exploiter.

Notre démarche a donc impliqué plusieurs étapes afin d'aborder ces données. Tout d'abord, une première étude du corpus nous a permis de faire émerger une typologie d'expression des problèmes techniques. Nous avons utilisé conjointement plusieurs techniques de fouille de texte, comme le *Topic Modeling ou Word2Vec*, afin de repérer un ensemble de régularités dans l'expression des problèmes techniques et d'y associer des indices et des marqueurs lexicaux. Ces mêmes indices sont à la base de cette typologie.

Une fois cette typologie validée, nous avons pu la mettre en œuvre dans l'annotation d'un *dataset*. Cela nous a permis de tester la pertinence de faire travailler des annotateurs non-experts sur des données complexes. Cela a été rendu faisable en raison d'une annotation

centrée sur la forme plutôt que sur l'interprétation des données et du repérage d'indices qui ne consistent pas en du vocabulaire technique. En effet, la construction d'une typologie sur des indices lexicaux centrés sur l'expression du problème en lui-même impliquent des indices faisant partie du vocabulaire courant. Ainsi une annotation non-experte devient possible dans un contexte d'accès limité à un expert. Nous avons ensuite pu mettre à profit ces données annotées pour entraîner un modèle neuronal sur une tâche d'étiquetage automatique. Cette tâche d'étiquetage a par la suite pu être reproduite après une opération de normalisation du corpus. En effet, nous avons mis en place une normalisation des données en utilisant un LLM (*Large Language Model*) génératif afin de corriger les phénomènes de bruit qui parsèment le corpus. Nous avons ainsi pu étudier, d'une part, la performance de ces modèles sur une telle tâche et des données très bruitées, mais aussi la pertinence d'une normalisation aussi lourde pour la tâche d'étiquetage des indices lexicaux.

Enfin, nous profitons de ces données étiquetées, obtenues en quantité suffisante pour adopter une approche inductive, pour nous concentrer sur deux types de la typologie qui sont *Fuite d'un liquide ou d'un gaz* et *Présence d'un obstacle*. L'étude de ces deux classes est l'occasion d'effectuer une analyse fine du contenu textuel des données. Pour cela, nous nous appuyons sur la théorie de la sémantique des cadres et la ressource *FrameNet* qui l'implémente. Cette théorie permet d'avoir une première représentation du contenu sémantique des données. Nous aurons pourtant aussi l'occasion de voir les limites des *frames* sélectionnés dans *FrameNet* face au domaine dans lequel s'inscrit le corpus et l'objectif de modélisation que nous nous sommes donné.

Ainsi, la catégorie *Fuite d'un liquide ou d'un gaz* fait l'objet d'une étude en soi, qui s'appuie sur plusieurs approches de linguistique de corpus et d'un entretien avec un expert afin de faire ressortir le *frame de la fuite dans un environnement technique*. Nous analysons donc, à partir des *patterns* récurrents que nous retrouvons dans ce sous-corpus, quels éléments textuels constituent la situation de fuite en contexte. Cela nous permet aussi d'identifier et de repérer les éléments nécessaires à la constitution du *vépole*, formalisme issu de TRIZ que nous visons pour représenter le dysfonctionnement technique.

Pour la catégorie *Présence d'un obstacle*, nous nous appuyons cette fois sur le *frame* de *FrameNet* (légèrement modifié). Celui-ci nous sert de point d'appui pour la mise en place d'une annotation automatique de type *Frame Semantic Role Labeling* avec un LLM génératif. Par ce biais, nous évaluons un outil récent et moderne dans l'opération de cette tâche complexe, qui l'est d'autant plus sur des données spécialisées et bruitées. Nous pourrons aussi voir les difficultés que pose cette catégorie pour une modélisation et en quoi les REX, tels qu'ils sont rédigés à ce jour, comportent des limites dans l'élaboration d'une chaîne de traitement automatique vers TRIZ.

Publication et communications

Les recherches effectuées pendant cette thèse ont mené à une publication :

— Maarouf, M. et Tanguy, L. (2025). Automatic normalization of noisy technical reports with an LLM : What effects on a downstream task ? In Bak, J., Goot, R. v. d., Jang, H., Buaphet, W., Ramponi, A., Xu, W. et Ritter, A., éditeurs : *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pages 38–44, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Elles ont aussi donné lieu aux communications suivantes :

— Maarouf M., Tanguy L. (2025) Fine grain semantic analysis by LLMs : real-world technical reports vs simple synthetic sentences, *LLM Fails Workshop*, Mannheim, Allemagne

— Maarouf M., Tanguy L., Kurela M. (2023) Analyses fines de textes techniques bruités : stratégies de contournement, *Journée d'étude Bruit de fond ou valeur ajoutée ? Gérer le bruit lors des traitements informatiques des corpus linguistiques*, Grenoble, France

— Maarouf M. (2023), Traitement automatique de rapports d'incidents : application au domaine spatial, *Printemps des Jeunes Chercheurs*, Toulouse, France

Plan de la thèse

Le manuscrit de thèse tel que nous l'avons conçu est constitué de quatre parties distinctes.

Dans la première partie de la thèse, nous faisons l'état de l'art des trois composantes essentielles de notre travail. Dans le premier chapitre, nous nous focalisons sur le REX et sa place dans la culture de sécurité. Nous envisageons aussi le REX sous l'angle de la linguistique et du TAL. Dans le chapitre 2, nous présentons la méthode de résolution de problème TRIZ et comment nous l'appréhendons dans notre travail. Enfin, dans le chapitre 3, nous discutons de la sémantique des cadres, théorie linguistique qui nous sert de point d'appui dans l'analyse fine de la représentation sémantique des rapports d'anomalies.

La seconde partie de cette thèse vise à présenter les données avec lesquelles nous travaillons. Dans le chapitre 4, nous présentons le corpus et son contenu langagier. Dans le chapitre 5, nous explorons plus avant ce contenu et notamment la description du problème à travers plusieurs méthodes de linguistique de corpus. Cela nous a permis de construire une typologie d'expression du problème technique et nous détaillons à cette occasion sa construction et son contenu.

La troisième partie présente un premier travail de traitement automatique du corpus. Le chapitre 6 met en avant le processus d'annotation et d'étiquetage du corpus en fonction de la typologie d'expression d'un problème technique à partir duquel nous pouvons effectuer un

apprentissage automatique pour une classification du corpus. Au chapitre 7, nous reproduisons ces traitements mais après un travail de normalisation des données. Nous présentons par la même occasion une méthode de normalisation des données par l'utilisation d'un LLM.

La quatrième et dernière partie de ce manuscrit est constituée d'une analyse fine du contenu textuel de deux catégories de la typologie. Le chapitre 8 se concentre sur la catégorie *Fuite d'un liquide ou d'un gaz*. Nous y proposons une approche méthodologique pour l'étude d'énoncés techniques à travers le repérage de patrons. Dans le chapitre 9, nous nous focalisons sur la catégorie *Présence d'un obstacle* et étudions la mise en place d'une annotation automatique de type *Frame Semantic Role Labeling* par le biais de LLM. Pour ces deux catégories, nous étudions aussi dans quelle mesure leur modélisation, telle que conçue par la méthode TRIZ, est réalisable et quelles en sont les limites.

Première partie

Du texte aux connaissances dans un contexte industriel et technique

Chapitre 1

Le REX : un objet multiple

Sommaire

1.1 Introduction : La nécessité du REX

Imaginons que je vous informe que la porte du local 803 est ouverte. *A priori*, votre réponse serait de l'ordre du « Et alors ? ». En revanche, si je vous dis que la porte du local 803 est restée ouverte et que l'opératrice Dupont est rentrée tête la première dans cette fameuse porte, vous le considéreriez peut-être différemment. Et si maintenant, j'ajoute que Dupont était en charge de lancer le test de vérification de température (test très important dans le cadre de cette histoire), et que personne n'a noté son absence, vous commencerez peut-être à vous inquiéter. Et oui, vous avez raison, le lanceur a explosé. Pourquoi ? Car Dupont n'a pas procédé au test de vérification de température, ou parce que Dupont s'est cognée la tête, ou bien parce que la porte est restée ouverte, ou bien tout cela à la fois.

À travers cet exemple imagé, nous avons souhaité montrer comment une simple anomalie peut avoir des conséquences imprévues, et en l'occurrence bien plus graves, que supposées. Dans les REX (Retour d'Expérience) dont nous faisons l'étude dans cette thèse, soit les FA (les Fiches Anomalies) du lanceur spatial Ariane 5, les problèmes évoqués sont avant tout de l'ordre d'anomalies, exposées de façon succinctes, plutôt que des rapports d'accidents. Ils consistent en des rapports rédigés sur le vif, de la part d'opérateurs s'adressant à d'autres experts, dont les anomalies rapportées visent à être résolues et non reproduites. Ainsi, nous avons pu trouver dans notre corpus la description du problème ayant donné lieu à l'exemple illustré ci-dessus : « PORTE DU LOCAL 803 OUVERTE PENDANT L'OPERATION. ». Ce procédé qui consiste à rapporter tout écart à l'attendu n'est pas exceptionnel. Il découle d'une culture de la sécurité et du risque qui a identifié un besoin, celui d'une approche systématique du rapport d'anomalies afin d'empêcher les incidents, les accidents, voire les catastrophes. Ces rapports prennent forme par le biais de la démarche du REX.

Dans ce chapitre, nous étudions les REX sous trois aspects pertinents à la thèse. Le premier présente le REX selon son contexte industriel et ses liens avec la maîtrise des risques. Il permet de placer le contexte de recueil et d'utilisation des REX, d'un point de vue pratique, mais aussi tel qu'il est théorisé du point de vue de la recherche académique. Le second aspect se focalise sur la composante linguistique du REX en tant qu'objet textuel et relevant d'un langage de spécialité. Enfin, nous présentons le REX tel qu'il est considéré en TAL, des difficultés qu'il pose aux possibilités qu'il ouvre.

1.2 Les REX : contexte et sécurité

Le REX est une démarche privilégiée de la maîtrise des risques dans les organisations (Casse et Caroly, 2014). Il fait partie intégrante de la culture de sécurité à des fins de gestion des risques. Afin d'illustrer ce propos, nous présentons dans un premier temps ce qu'est la culture de sécurité. Dans un second temps, nous revenons sur les différentes définitions du REX et sur son usage. Enfin, nous présentons une des composantes essentielle d'un dispositif de REX, la recherche de la cause d'un problème technique.

1.2.1 La culture de sécurité

La culture de sécurité est présente dans toutes les organisations ayant affaire à la gestion des risques, et plus spécifiquement, celles qui traitent avec les risques importants pouvant entraîner des accidents majeurs, graves ou mortels (Institut pour une Culture de Sécurité Industrielle (ICSI), 2017) (nous revenons sur la distinction entre ces différents niveaux de risques en partie 1.2.3). Elle a été étudiée dans différents domaines de recherche (la sociologie

du travail, l'ergonomie, la psychologie…) mais demeure un concept dont la définition, la portée et le statut restent peu clairs (Hale, 2000). Dans cette section, nous tenterons de circonscrire le concept de « culture de sécurité » tel qu'il se manifeste dans le contexte industriel dans lequel s'inscrit notre étude. Nous évoquerons son lien avec les REX et la gestion des risques.

La **culture de sécurité** possède plusieurs définitions, modèles et même dénominations : « culture en relation à la sécurité », « culture de sûreté », « culture de sécurité » (« *safety culture* »), climat de sécurité (« *safety climate* »). Nous différencions les trois premières appellations de la dernière (« climat de sécurité ») en nous fondant la distinction effectuée par Fucks (2013), elle-même inspirée des travaux de (Hale, 2000), Guldenmund (2000) et (Wiegmann *et al.*, 2002) :

> « La culture de sécurité renvoie à une formation complexe, relativement stable, ancrée dans des manières de faire et de penser et nécessite des méthodes qualitatives d'analyse ; le climat de sécurité renvoie à une photographie, prise à un moment donné de la vie d'une organisation, des diverses perceptions que les personnels ont au sujet de la sécurité ».

Dans le cas présent, c'est à la culture de sécurité telle que décrite ci-avant que nous nous intéressons. Nous différencions par ailleurs la « sécurité » de la « sûreté » industrielle en nous basant sur les définitions de l'IMDR (Institut pour la Maîtrise des Risques). La première concerne « l'ensemble des dispositions techniques et des mesures prises pour prévenir des risques découlant d'événements dommageables aux travailleurs (sécurité professionnelle), au public, aux biens (matériels, immatériels) et à l'environnement naturel et protéger ces derniers en cas de survenance d'incidents et d'accidents ». Elle se différencie de la seconde qui traite des malveillances intentionnelles : « l'ensemble des dispositions techniques et des mesures prises pour prévenir et se protéger d'actes de malveillance intentionnelle contre des personnes et/ou des actifs (corporels, incorporels, financiers) de l'entreprise »[1].

Choudhry *et al.* (2007) listent un ensemble de définitions de la culture de sécurité. Parmi ces définitions, les trois suivantes nous semblent couvrir les différents aspects de ce concept tel que nous l'entendons :

- Hale (2000) considère la culture de sécurité comme « *the attitudes, beliefs and perceptions shared by natural groups as defining norms and values, which determine how they act and react in relation to risks and risk control systems* » ;
- Guldenmund (2000) en fait une définition plus large « *those aspects of the organizational culture which will impact on attitudes and behavior related to increasing or*

1. https://www.imdr.eu/818_p_57385/maitrise-des-risques.html

decreasing risk » ;

— Pour Cooper (2000) la culture de sécurité est « *the product of multiple goal-directed interactions between people (psychological), jobs (behavioral) and the organization (situational); while safety culture is 'that observable degree of effort by which all organizational members directs their attention and actions toward improving safety on a daily basis* »

Selon ces définitions, la culture de sécurité est un concept centré sur la relation entre les membres de l'organisation, la sécurité et le risque. La première évoque les « normes et valeurs », la seconde les « attitudes et comportement » et la dernière « l'attention et les actions ». Si Hale et Guldemund restent plus généraux, Cooper spécifie quant à lui cette notion sur la sécurité uniquement (et non le risque) dans l'idée d'**améliorer** la sécurité au sein de l'organisation. Le concept d'organisation revient aussi avec le concept de « culture organisationnelle » dans les deux dernières définitions ci-dessus. La culture de sécurité y est décrite comme une composante de la culture organisationnelle. Cette dernière lui préexiste puisque la culture de sécurité naît suite à l'accident de la centrale nucléaire de Tchernobyl[2]. En effet, l'apparition du terme « *safety culture* » est constatée dans le *INSAG's Summary Report on the Post-Accident Review Meeting on the Chernobyl Accident* (alors traduit par « culture de sûreté », appellation encore utilisée aujourd'hui mais principalement dans le domaine nucléaire). Elle est donc une sous-composante de la culture organisationnelle, mais focalisée sur le lien avec la sécurité.

La **maîtrise des risques** est définie selon l'IMDR (Institut pour la Maîtrise des Risques) et Francis *et al.* (2020) comme :

> "l'aptitude d'un organisme (entreprise, collectivité…) placé dans un contexte d'efficacité, à optimiser la performance de ses activités sans subir ou faire subir à ses parties prenantes et à son environnement des risques de dommages technologiques, environnementaux, économiques ou humains. Pour ce faire, l'organisme utilise un processus de prévention et de protection pour rendre tous les risques pris individuellement acceptables".

Elle souhaite répondre à la problématique de connaître les risques (et les risques qu'on accepte de prendre), les mesurer et les traiter. Elle concerne ainsi les différentes méthodes et outils pour répondre à ces besoins (Tea, 2009). Par ailleurs, le concept de la culture de sécurité comprend aussi bien un outil de compréhension des pratiques, que le système que peuvent implémenter et améliorer les entreprises (Fucks, 2013). Dans le cas du retour d'expérience, c'est ce deuxième aspect qui est sollicité. Le REX devient une forme de mise en pratique d'une

2. La catastrophe de Tchernobyl a lieu lors d'un essai technique. Plusieurs facteurs (humains et techniques) ont simultanément entraîné une forte perturbation du réacteur nucléaire.

culture de sécurité à des fins de maîtrise des risques. En effet, la culture de sécurité, par ces objectifs d'étude, devient utile et peut aider à une maîtrise des risques (Chevreau et Wybo, 2007).

Pour le domaine spatial, certaines lois viennent définir ces différentes notions. La **sécurité** est définie dans la loi n° 2008-518 [3] comme l'« ensemble des dispositions destinées à maîtriser les risques dans le but d'assurer la protection des personnes, des biens et la protection de la santé publique et de l'environnement ». Dans un arrêté du CNES disponible publiquement [4], nous retrouvons aussi les définitions de :

— la **sûreté** : « mesures relatives à la protection des personnes et des installations prévues par la législation et la réglementation applicables et dont la mise en œuvre est placée sous la coordination du président du Centre national d'études spatiales, au titre de l'article 14-11 du décret relatif au CNES précité. R331-14 du Code de la recherche » ;
— la **sauvegarde** : « ensemble des dispositions destinées à maîtriser les risques techniques liés à la préparation et à la réalisation des lancements afin d'assurer la sécurité des personnes et des biens et la protection de la santé publique et de l'environnement, au sol et en vol ». Ce dernier prend deux dimensions qui sont la *sauvegarde au sol* et la *sauvegarde en vol*.

Ces trois éléments, *sûreté*, *sauvegarde en vol* et *sauvegarde au sol* se retrouvent dans trois services distincts au CNES : un service « Sûreté Protection » qui concerne la malveillance ou le sabotage, un service « Sauvegarde Vol » et un service « Sauvegarde-Environnement » qui gèrent les risques non-intentionnels liés aux activité des opérations spatiales. Les FA que nous étudions dans cette thèse se concentrent sur le dernier point, soit les risques en lien avec les opération spatiales au sol quelque soit leur niveau de gravité.

1.2.2 Définir le REX

Suite à différents accidents industriels majeurs, la nécessité de se pencher spécifiquement sur la démarche de retour d'expérience se fait sentir. En France, le Groupe d'Intérêt Scientifique (GIS) « Risques » du CNRS dressa un bilan en 1998 sur la mise en pratique de procédés de retour d'expérience au sein des industries (Tea, 2009). Ces premières études font émerger une définition générale du REX selon laquelle la démarche du retour d'expérience consiste en la collecte d'information sur les événements « répétitifs, traitables "en interne" et appréhendés sous l'angle des facteurs techniques » (Gilbert, 2001).

3. https://www.legifrance.gouv.fr/loda/id/JORFTEXT000024095828
4. https://cnes.fr/sites/default/files/2024-07/REI-NG-VF-15072024-pour-consultation.pdf

D'un point de vue pratique, le dispositif de REX est constitué de cinq étapes décrites par Hadj-Mabrouk et Hamdaoui (2008) :

1. la collecte de données : la collecte de toute anomalie rencontrée par les opérateurs pour en recueillir un maximum ;
2. l'analyse des données : l'établissement des différentes causes possibles à l'anomalie, et pas uniquement les causes premières apparentes ;
3. le stockage des données : l'archivage des informations (données et analyse) récoltées ;
4. l'exploitation des données : apprendre du processus par les différents résultats issus des traitements des rapports et extraire les événements qui sont prédictifs ;
5. les recommandations : identifier les mesures à mettre en place pour empêcher la reproduction des événements, tirant ainsi profit de l'expérience en améliorant la sécurité.

Ces étapes, illustrées en Figure 1.1, font partie d'une démarche globale du REX, que les organisations mettent en place par le biais d'un ensemble d'outils et de procédures. Elles impliquent différents acteurs du REX que ce soit les opérateurs qui rencontrent l'anomalie, les experts en charge d'identifier la cause de cette anomalie, les responsables techniques impliqués en fonction de la gravité de la situation. Il est à noter qu'un même acteur peut prendre plusieurs rôles. Ils ont cependant la particularité de posséder un langage technique commun, question que nous aborderons en section 1.3 de ce chapitre.

Le REX est donc plus que le simple constat d'une anomalie, c'est un processus composé de plusieurs étapes précises et définies qui est au cœur de la gestion des risques.

La littérature actuelle fait état de nombreuses définitions du REX, qui viennent notamment de la pluralité des pratiques observées par les entreprises. Wybo *et al.* (2003) et Tea (2009) en listent plusieurs qui montrent que le REX peut être considéré selon différents angles. Dans notre cas d'étude, nous retenons quatre définitions, qui nous semblent correspondre au cas d'étude des Fiches Anomalies de notre corpus d'étude :

1. Abramovici (1999), tel que cité par Tea (2009), décrit la démarche du REX comme « l'ensemble des moyens mis en place afin de conserver formellement les connaissances issues de l'analyse du fonctionnement réel du système et permettre leur exploitation ». Une définition plutôt généraliste et qui met en avant l'objectif de construction de connaissances, à partir de ce recueil de données.
2. Amalberti et Barriquault (1999), toujours cités par Tea (2009), apportent une nuance capitale en précisant que le REX « sert à récupérer et à exploiter une information sur

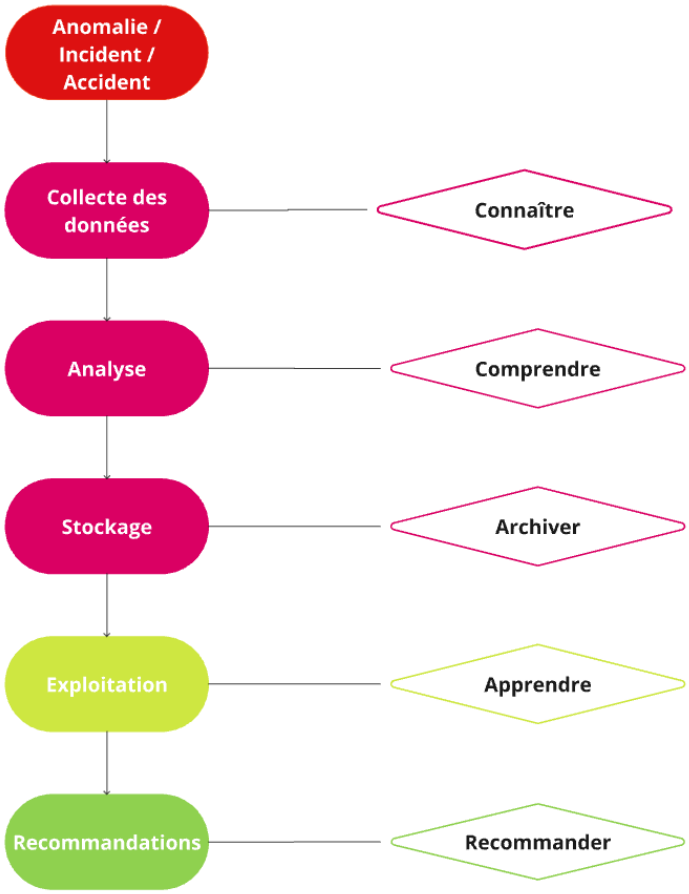


FIGURE 1.1 – Schéma du processus du REX d'après Hadj-Mabrouk et Hamdaoui (2008)

les dysfonctionnements des situations de travail des acteurs de premières lignes ». Il mettent ici l'accent sur la composante humaine des auteurs des REX, mais aussi sur l'immédiateté qui découle de leur position « en première ligne » dans la rédaction de ces retours.

3. Dal Pont (2003) décrit les REX par leurs objectifs qui sont « de saisir les anomalies, les déviations par rapport au prescrit et à l'attendu, d'analyser les incidents et accidents. C'est une source d'enseignement, d'engagement et d'amélioration ». Cette définition met en avant le fait que ce n'est pas la notion de problème à laquelle on s'intéresse au sein des REX, mais avant tout la notion d'écart. Toute situation anormale par rapport à l'attendu, peut être source de problème, notamment dans un contexte de gestion des risques. Un écart à la norme ou à l'attendu, aussi petit soit-il, est une raison suffisante à la production d'un REX, et à l'analyse qui s'ensuit.

Si les définitions du REX peuvent être multiples, son objectif principal n'en est pas moins identifié : réduire et éviter la reproduction des risques. Pour Wybo *et al.* (2003), un des principes fondamentaux du retour d'expérience est son déclenchement suite à une anomalie. Une démarche de REX vise avant tout à éviter qu'un dysfonctionnement se reproduise, peu en importe l'auteur (Wybo *et al.*, 2001). Le REX, en lui-même, n'est pas un objet d'intérêt, c'est sa participation à la gestion des risques qui le rend important, voire primordial. Selon Gilbert (2001), il est nécessaire pour maintenir la poursuite des activités dans les secteurs à risques. Plus le contexte industriel présente des risques importants, plus les besoins de le gérer, et d'apprendre des expériences passées peut se faire sentir. Ceci conduit à la mise en place d'une démarche de REX qui peut porter sur des écarts qui peuvent paraître de nature anodine. La gravité de la situation ne peut pas être systématiquement estimée au moment du constat de l'anomalie.

Dans le domaine spatial, l'accident de la navette *Challenger*[5] en 1986 remet en question la gestion des risques jusque-là pratiquée. En effet, en plus du problème technique qui a provoqué l'accident, le rapport de Vaughan (1990) appuie sur la nécessité de repenser la démarche du retour d'expérience telle qu'elle est pratiquée qui s'est montrée insuffisante pour limiter les risques. Cet accident a mis en avant le besoin d'un système de gestion des risques fondé sur une pratique systématique et organisationnelle, dans un contexte de sécurité important.

Ainsi, le processus de REX est un outil privilégié dans des organisations avec une démarche de gestion des risques. Malgré des définitions multiples et qui continuent d'évoluer, c'est un procédé défini et organisé dont la visée est une amélioration de la sécurité qui passe par un apprentissage des expériences passées.

1.2.3 L'analyse de causalité

Les REX ont différents objectifs que Tea (2009) nous liste de la façon suivante :
— « déterminer des mesures correctives » ;
— « vérifier le bon fonctionnement » (avec par exemple un processus de veille) ;
— « déterminer des mesures préventives » ;
— « identifier des points de faiblesses » (avec par exemple la recherche de signaux faibles).

5. La navette *Challenger* a explosé 73 secondes après son décollage, causant la mort des sept astronautes à son bord. La défaillance est causée par une fuite d'un des joints non adapté aux températures basses (soit 1° Celsius lors du lancement.)

Ces objectifs visent tous, de façon plus ou moins directe, à éviter la reproduction d'accidents. Le rapport de ces anomalies est donc suivi d'un processus d'analyse (comme nous l'avons vu chez Hadj-Mabrouk et Hamdaoui (2008)) afin d'en identifier la cause.

Dans les rapports d'anomalie dont nous faisons l'étude, la visée est de trouver les causes organisationnelles et structurelles de l'anomalie, plutôt que de chercher un responsable humain au problème. Selon Daniellou *et al.* (2010), la cause d'un accident n'est plus considérée comme « l'erreur d'un opérateur » dans la communauté scientifique : la focalisation sur l'erreur humaine est inefficace. L'erreur humaine a cependant été un point focal dans l'attribution d'une cause à un accident d'après Reason (1997) qui met en faute le *« blame cycle »*. Ce cycle empêche la recherche d'une autre cause puisqu'elle enclenche une boucle d'attribution du *blâme* à un responsable. Pourtant, afin d'obtenir un rapport à l'erreur prompte à une culture de sécurité efficace, il est nécessaire de rechercher les causes structurelles à l'apparition de l'erreur. Selon Reason (1997), pour briser le « cycle du blâme », il faut mettre en place ces cinq composantes :

- *reporting culture* : une atmosphère de liberté à rapporter les erreurs et les risques ;
- *just culture* : encourager le rapport d'informations concernant la sécurité dans un climat de confiance ;
- *flexible culture* : un traitement des erreurs par des experts compétents avant d'être considérés par la hiérarchie ;
- *learning culture* : une disposition à apprendre des erreurs et des accidents de la part des organisations.

Catino (2008) met en contraste ce procédé de blâme d'un individu, l'*individual blame logic*, avec une approche de la cause organisationnelle, l'*organizational function logic*. Cette dernière vise à trouver les causes structurelles et organisationnelle à un accident. Elle s'inscrit dans l'idée développée par Perrow (2011) selon laquelle la survenue d'une anomalie est normale dans un contexte très technique, comme celui de la maintenance des fusées Ariane. Par ailleurs, dans le rapport de l'accident de la navette *Challenger* produit par Vaughan (1990), elle appuie la notion selon laquelle, malgré le fait que les accidents résultent d'erreurs humaines, comme tout système impliquant des personnes, ils découlent néanmoins de problèmes organisationnels. Il est donc nécessaire de chercher les raisons systémiques de l'accident plutôt que les causes individuelles, de s'interroger sur quels événements ont conduit à l'erreur (Leveson, 2004). La recherche de causes structurelles et organisationnelles apparaît notamment lors de cas d'accidents majeurs. La visée est d'identifier les causes structurelles à l'arrivée de l'événement, au-delà des causes matérielles et de l'erreur humaine, apporter un recul afin d'observer le système dans son ensemble, de prendre en compte sa dynamique (Lim *et al.*, 2002).

Les rapports avec lesquels nous travaillons ne font cependant pas état d'accidents majeurs mais plutôt d'anomalies (pannes, usure...). De nombreuses typologies existent en fonction des domaines, des entreprises et des « stratégies » de classification[6]. L'échelle INES[7] (International Nuclear and radiological Event Scale) en Figure 1.2 présente la catégorisation de ce type d'événement dans le domaine nucléaire. Nous y retrouvons cependant les différentes catégories mentionnées jusqu'à présent, à savoir les anomalies (tout en bas de l'échelle), les incidents catégorisés en deux niveaux, et les accidents catégorisés en 4 niveaux jusqu'à « accidents majeurs ». Dans le domaine de l'aviation civile en France, nous retrouvons 6 catégories différentes en fonction de la gravité qui sont : Accident / Incident grave / Incident majeur / Incident significatif / Aucune incidence immédiate sur la sécurité / Non déterminé. L'accident implique des conséquences de type blessures, décès ou destruction de l'aéronef alors que l'incident concerne les cas où un effet (plus ou moins important) est constaté sur la sécurité des vols. Le classement d'un événement peut toutefois évoluer dans le classement au fur et à mesure de son investigation (Johnson, 2003). Dans le cas d'une anomalie, la gravité peut être plus ou moins importante. Pour certaines anomalies, un expert est à même d'en déterminer la gravité à la lecture des rapports à partir des éléments mentionnés. Pour d'autres, il peut être nécessaire d'effectuer un diagnostic et d'envisager les conséquences éventuelles du problème pour en déterminer la gravité. Cette notion de gravité est essentielle lorsqu'il est question des REX. Des travaux de classification automatique des FA en fonction de la gravité par des méthodes de TAL ont d'ailleurs été explorés par Kurela *et al.* (2020), travaux que nous présentons plus en détail dans la section 1.4.

Des méthodes pratiques existent afin de rechercher la cause de ces problèmes. Hesslow (1988) distingue deux aspects du diagnostic d'une anomalie :

- *the « connection problem »* : pouvoir repérer un lien de cause à effet entre deux événements ;
- *the « selection problem »* : sélectionner, parmi les différentes causes possibles, la plus pertinente pour le problème.

Ces deux aspects découlent du fait que (1) plusieurs événements concourent à l'occurrence d'un accident, (2) une chaîne de cause à effet peut-être retracée suite à un accident, (3) une cause peut être conceptualisée de différentes manières. La question de la sélection des causes a donné lieu à un processus appelé « analyse des causes premières » (*« root cause analysis »*). Cette pratique inclut plusieurs techniques, outils et méthodes dédiées à la sélection de la cause *première* du problème. C'est cet événement causal qu'il est nécessaire de régler pour

6. Par exemple, Johnson (2003) développe plusieurs stratégies pour classifier les accidents/incidents.
7. https://afcn.fgov.be/fr/situations-durgence/echelle-ines

empêcher une reproduction de l'accident. Si on reprend les étapes du REX telles que définies dans Hadj-Mabrouk (2004), le procédé d'analyse des causes premières correspond à l'étape d'analyse des données (étape 2) et se voit mis à profit lors de l'étape 4 d'exploitation des données pendant laquelle on détermine quel élément causal, événement ou particularités, doit être prévenu pour empêcher la reproduction de l'anomalie. Il permet de repérer quel élément doit être modifié afin d'éviter une reproduction de l'accident.

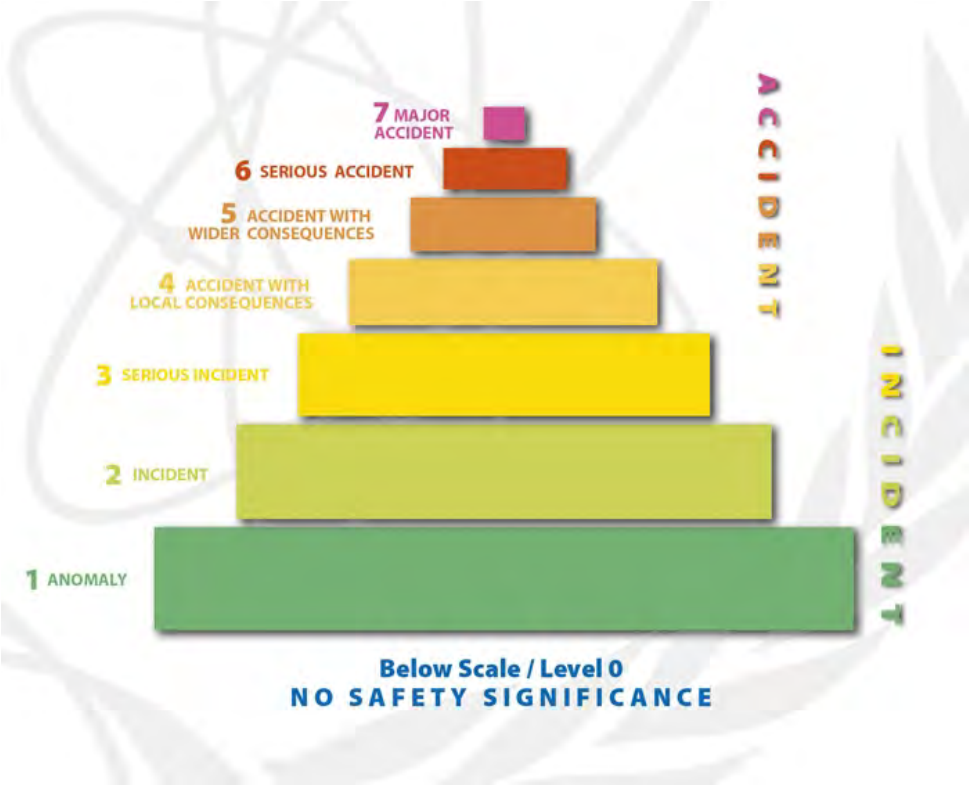


FIGURE 1.2 – Échelle INES

Dans cette section, nous avons exploré l'univers du REX et notamment celui de la sécurité dans des organisations liées à de hauts risques techniques et industriels. Cet univers est fortement lié à la culture de sécurité, qui a marqué un changement dans la gestion des risques telle qu'elle pouvait être considérée préalablement. Ce changement s'est aussi constaté dans la façon dont est envisagée la cause d'un problème, s'éloignant de l'idée d'une cause individuelle pour aller vers des raisons liées au contexte structurel de l'organisation. Le REX est un procédé majeur dans cette recherche de la cause, et l'objectif d'empêcher la reproduction d'un accident. Il est une étape désormais incontournable d'une organisation telle que le CNES afin d'obtenir une maîtrise des risques efficace. La première manifestation de ce processus dans les FA est un rapport rédigé sur le vif, par un expert, et à destination

d'un expert, dans un contexte professionnel. De ce fait, son contenu reflète une langue de travail, une communication en domaine de spécialité. Elle implique donc l'usage d'un langage spécialisé, que nous explorons dans la section suivante. Nous reviendrons sur le procédé des FA tel qu'il est implémenté au CNES et chez Arianeespace dans le chapitre 4.

1.3 Les REX : un objet linguistique

Dans cette section, nous souhaitons introduire différentes notions linguistiques autour du REX. À partir des différents aspects du REX et des quelques exemples de FA que nous avons vus dans la section 1.2, nous pouvons déjà commencer à circonscrire le type de texte qui constitue notre corpus. En effet, nous avons pu voir que la situation de communication implique un certain nombre d'éléments. Ces documents prennent forme dans un contexte professionnel, et sont des communications d'expert à expert. Ils impliquent donc des dénominations spécifiques pour les éléments auxquels ils font référence et une phraséologie spécialisée. Ces REX font aussi état d'un besoin de précision et de concision dans l'expression des anomalies.

Afin d'aborder le REX comme un objet linguistique, nous définissons d'abord les notions de « langage de spécialité » et ces différentes dénominations. Nous présentons ensuite le rôle du contexte, notion essentielle dans ce type de texte. Enfin, nous discuterons des caractéristiques essentielles et récurrentes des langues de spécialités que nous retrouvons dans le corpus.

1.3.1 Langue, langage et sous-langage ?

L'étude de textes dits "de spécialité" a donné lieu à différentes dénominations. Ces dénominations ne sont pas anodines car elles véhiculent une certaine considération de l'objet d'étude.

Harris (1990) utilise l'appellation de « sous-langage » (*sublanguage* en anglais) pour faire référence aux sous-ensembles fermés de phrases qui n'utilisent qu'une partie des règles d'une langue. Il considère que ces sous-langages possèdent des règles communes avec la langue source (notamment syntaxiques), et des règles présentes dans la langue source mais qui ne seront jamais appliquées dans les sous-langages. Ces situations seraient particulièrement présentes dans les « langues de la science » (en. « *language of various science* ») et donc proche de la notion de langue de spécialité. Lerat (1995) rejette le nom de sous-langage (plus propre aux dialectes). Il lui préfère le nom de « langue de spécialité » ou « langue spécialisée » qu'il décrit comme « renvoy[ant] au système linguistique pour l'expression et aux professions pour le savoir ». Chez Boutet (2001b), le terme de « langue » ne peut être

utilisé, « car nous n'avons jamais affaire à des "langues" qui seraient distinctes du français commun », et le terme de « langage » lui est préféré. Pour notre part, nous sommes en accord avec la distinction qu'opère Boutet et nous utilisons dans cette thèse la dénomination « langage » de spécialité de façon préférentielle, et parfois « langue » par abus de langage. Il existe d'autres dénominations (« langages opératifs » chez Falzon, « langues spéciales » chez Saussure, « langues techniques » chez Dauzet...). Les appellations varient mais certains traits semblent communs à la définition d'un langage de spécialités :

1. sa parenté avec la langue générale (Condamines, 1997) ;
2. la communication comme fonction principale (Kocourek, 1991) ;
3. le caractère spécialisé (*ibid*).

Cette parenté avec la langue générale implique que, si un locuteur maîtrise la langue générale et est familier avec le domaine en question, il est capable de comprendre la langue de spécialité en question (Charnock, 1999). Les langages de spécialités sont aussi définis comme moins étendus par rapport à la langue générale (Falzon, 1989) « quel que soit l'aspect considéré (lexique, syntaxe, sémantique, pragmatique), ils sont restreints et déformés par rapport au langage naturel ». Il les décrit comme issus de l'adaptation de la langue naturelle afin de mieux convenir aux connaissances du domaine (Falzon, 1987). Boutet (2001a) va jusqu'à définir les pratiques langagières comme dépendantes des conditions matérielles de travail. Dans le corpus des FA, nous verrons dans quelle mesure (cf. chapitre 4) le fait de rédiger ces rapports dans l'immédiateté et à destination d'opérateurs experts qui partagent une connaissance commune entraîne une brièveté des rapports, ainsi qu'un recours aux acronymes et codes de façon abondante.

La notion de sous-langage vue chez Harris a donné lieu à des rapprochements avec d'autres notions. En effet, comme dit chez Somers (1998), le terme de « sous-langage » a pu être associé à celui de « registre » (Biber, 1988) ou celui de « genre textuel ». Bhatia (1993) définit le genre textuel de la façon suivante : « a recognizable communicative event characterized by a set of communicative purpose(s) identified and mutually understood by the members of the professional or academic community in which it regularly occurs ». Cette définition remplit les conditions d'émergence d'un sous-langage telle que définis par Kittredge (2005) : « a community of speakers (i.e. 'experts') shares some specialized knowledge about a restricted semantic domain » et « the experts communicate about the restricted domain in a recurrent situation, or set of highly similar situations ».

La question se pose alors pour les REX avec lesquels nous travaillons. Ceux-ci relèvent sans équivoque des langages de spécialités. Nous y retrouvons les caractéristiques de tout langage de spécialité, que nous explorons plus en détail dans la partie 1.3.3. Selon ces défini-

tions, et de même que dans les travaux de Warnier (2018) portant sur la rédaction d'exigence techniques dans le domaine spatial, les Fiches Anomalies semblent remplir les critères d'appartenance à un même genre textuel. Nous y retrouvons la régularité du procédé, ici conséquence de la culture de sécurité et de l'incitation à rapporter toute anomalie. La communauté de laquelle ces rapports sont issus et celle à laquelle ils s'adressent est la même, ils sont membres du même corps professionnel et travaillent sur une même installation.

Dans cette partie, nous avons pu voir différentes dénominations d'un langage de spécialité et ainsi que sa définition et comment, de par leur nature, les FA répondent à ces caractéristiques. Les différents langages de spécialités présentent des variations internes qu'il faut étudier, ils ne répondent pas à des codes rigides, et nous aurons l'occasion de le faire dans le chapitre 8 de cette thèse. Nous pouvons aussi faire l'hypothèse que les différents items de notre corpus relèvent du même genre textuel. Dans la section suivante, nous aborderons plus avant les particularités extra-linguistiques des langages de spécialité à travers la notion de contexte.

1.3.2 L'importance du contexte

Vergely (2004) cite deux notions extralinguistiques inhérentes aux langues de spécialités et intrinsèquement liées qui sont : le domaine et l'expert.

La première notion extralinguistique, le **domaine**, est définit comme tel : « Les communications opératives et plus largement toute communication de travail s'interprètent de façon particulière en fonction du domaine dans lequel elles s'insèrent. C'est le domaine qui actualise le sens (voire parfois la signification) et l'interprétation des communications » (Vergely, 2004) . De même, Lerat (1995) nous dit que : « Le sens, en pareil cas, n'est pas seulement dépendant de la connaissance de la langue, mais ici la langue ne fait que mettre de l'ordre dans la connaissance des choses ». Boutet (2001a) qualifie le domaine de travail comme étant à la fois le contexte, mais aussi un élément au centre de l'activité langagière sur le lieu de travail : « De telles pratiques langagières ne peuvent que difficilement être traitées comme des corpus autonomes, tant est décisif leur ancrage dans les conditions matérielles de leur énonciation. Leur interprétation, comme leur analyse, ne peuvent se passer d'une prise en compte du cadre matériel de leur production, envisagé non seulement comme une situation sociale ou un contexte, mais surtout comme un composant central de cette activité de langage ». Ainsi, le contexte de travail donne sens à l'énoncé langagier.

Vergely (2004) considère la définition traditionnelle d'un **expert** comme une personne possédant des connaissances approfondies dans un domaine spécifique, c'est un spécialiste qui peut aussi faire office de référence dans son domaine, un dépositaire de connaissance. Elle défend néanmoins une prise en compte de l'expert plus nuancée. Un secteur de travail implique lui-même une pluralité de domaines, de spécialités. Un expert ne saurait être versé dans toutes les spécialités mises en œuvre, il est lui-même spécialisé. De ce fait, un échange entre experts ne signifie pas une compréhension totale systématique. Cette définition vient nuancer celle de Falzon (1986), cité par Vergely, : « il est communément admis que les experts ont des connaissances partagées relatives à une tâche, ce qui leur permet d'engager des communications verbales compréhensibles et efficaces uniquement par eux ».

Dans notre étude, les REX ont pour destinataires des opérateurs qui auront pour tâche de diagnostiquer et de résoudre le problème rapporté. Ces destinataires peuvent être un collègue, un supérieur, ou parfois eux-mêmes. Ils donnent donc lieu à un document le plus complet et précis possible. Notre objet d'étude concerne cependant une description minimale du problème. Cette description doit être suffisamment claire pour qu'un opérateur soit à même de comprendre l'étendue du problème, où il a lieu et sa gravité, tout en restant brève et concise. Ce minimalisme est aussi permis par l'utilisation de termes spécialisés, d'acronyme, de code. Ils sont importants pour la clarté, la précision de la description du problème. Lors d'un entretien avec un expert, nous avons rencontré la FA : "YSS022 ET YSS513 FUITE AU TEST HE > 10-7Ncm3/s". Les composants "YSS..." ne lui étaient pas familiers. Cependant, le reste des informations présentes lui ont permis de savoir quel type de problème est évoqué : une fuite (facilement compréhensible en ce cas avec l'indice lexical « FUITE »), ce qui fuit (du gaz) et la précision du débit. Ces informations lui ont suffi pour en déduire le type de composant qui fuit. La question se pose alors du degré d'expertise qui est requis pour comprendre un REX écrit en langage spécialisé. Nous pouvons supposer différents niveaux de compréhension. Un non-expert est aussi capable de comprendre qu'il est question d'une fuite et qu'un débit est mentionné. En revanche, le type de composant qui fuit, le fluide qui fuit et même le type de test en question (« TEST HE ») ne sont pas évidents.

Dans cette continuité, se pose la question de la place du linguiste dans l'analyse de textes spécialisés, textes difficilement compréhensibles et accessibles pour un non-expert (voire parfois même pour un expert). Condamines (1997) décrit le rôle du linguiste travaillant sur des textes spécialisés :

> (1) « Remarquons tout d'abord que la démarche du linguiste analysant les textes spécialisés a ceci de particulier qu'il n'a pas la compétence linguistique de la langue qu'il étudie, ou du moins pas toute la compétence. Il n'est pas à la fois juge et parti. »

> (2) « Tout l'art du linguiste étudiant des corpus spécialisés consiste à s'appuyer sur ce qu'il connaît d'un fonctionnement "standard", attendu de la langue tout en sachant que ce fonctionnement va, dans certains cas, ne pas être suivi. »

Le linguiste, ayant une expertise de la langue, est donc à même de travailler sur des corpus spécialisés. Cependant, le manque de connaissances expertes du domaine est un frein pour une compréhension complète des textes étudiés. De plus, comme dans de nombreuses études sur les langages spécialisés, obtenir la disponibilité des experts est parfois difficile. De ce fait, dans cette thèse, nous avons pu explorer la problématique suivante : jusqu'où une approche basée sur des caractéristiques linguistiques et sur des motifs répétitifs peut être menée dans l'étude de textes spécialisés ? À quel grain de l'analyse le savoir expert est-il nécessaire ?

La méthodologie que nous avons mise en place nous a permis de catégoriser des REX à partir d'une approche basée sur une analyse du corpus et de ses régularités (cf. chapitres 4 et 5). Les résultats de cette analyse et la typologie qui en découle ont par la suite été validés (et mis en applications) par un expert (voir chapitre 6). En revanche, une annotation plus fine, à la *FrameNet*, sans un expert s'est avérée en partie impossible. Nous explorons donc, dans les chapitres 8 et 9, l'analyse fine de ces textes avec un recours limité à un expert en vue d'obtenir une annotation similaire à celle de *FrameNet*.

Le domaine est une des caractéristiques élémentaires du corpus spécialisé, de même que l'expert en tant que destinataire des énoncés. Toutefois, dans notre travail, l'expert prendra aussi le rôle de « point d'accès », de « médiateur » entre le linguiste et la connaissance véhiculée par le texte. Il permettra de dépasser (ou de contourner) les obstacles à la compréhension et, en finalité, d'assurer le traitement de ces données.

1.3.3 Quelles caractéristiques ?

La caractéristique saillante principale des langages de spécialités est son lexique, produisant ainsi de nombreuses études en terminologie pour tel ou autre domaine. Cependant, comme dit par Lerat (1995) : « Une langue spécialisée ne se réduit pas à une terminologie : elle utilise des dénominations spécialisées (les termes), y compris des symboles non linguistiques, dans des énoncés mobilisant des ressources ordinaires d'une langue donnée ». Si un langage de spécialité fait appel au système linguistique de la langue générale, il est donc cohérent qu'il puisse aussi être étudié sur d'autres plans que la terminologie. En revanche, Lerat et Falzon (1989), cités par Vergely (2004), s'accordent pour dire qu'il est question d'un système plus « simple », plus « restreint » et donc « déformé » par rapport à la langue générale. Dans cette partie, nous aborderons les aspects lexico-sémantiques des langages spécialisés et ses différences avec la langue générale.

Chaque domaine technique et scientifique possède son vocabulaire particulier, son lexique (Guilbert, 1973). Guilbert et Falzon observent un caractère monosémique à ce lexique :

> « Il semblerait par contre que, dans chaque vocabulaire particulier, la référence à un objet ou un concept puisse être obtenue par l'intermédiaire d'un seul terme. Ce serait même, à première vue, la condition pour que soient introduits de l'ordre et de la cohérence dans la connaissance de la réalité et dans la communication de l'expérience » (Guilbert, 1973).

Les besoins de clarté, d'empêcher les confusions, dans les domaines techniques et d'autant plus dans les domaines à risques, sont vecteurs de cette monosémie. Cependant, plusieurs auteurs viennent nuancer ces propos, que ce soit en parlant de polysémie à proprement parlé (Mortureux, 1995) ou d'effets de sens (Gentilhomme-Koutyrine, 1994), comme exposé dans Condamines et Rebeyrolle (1997) pour qui la polysémie est avant tout corrélée à la notion de point de vue. Ces points de vue engendrent une *polyacception* des termes en contexte spécialisé, puisqu'ils peuvent différer de leur association « à un choix individuel ou à un choix collectif à relier à des compétences socioprofessionnelles communs à un groupe identifiable de locuteurs ». Selon (L'Homme, 2023), dans les travaux de terminologie qui impliquent cette polysémie, la majorité cherchent à l'éviter afin de supprimer l'ambiguïté dans les communications en contexte professionnel. Cependant, elle met en avant l'importance et la récurrence de ce phénomène dans les textes spécialisés et la pertinence de sa prise en compte lors de la création de ressources terminologique pour les domaines spécialisés.

Dans Kocourek (1991), l'auteur cite aussi les unités lexicales « non linguistiques », ou brachygraphiques, comme caractéristiques des textes de spécialités. Elles désignent les unités avec des formes graphiques plus concises, de l'ordre de l'abréviation ou l'idéographie. Dans notre corpus, elles sont présentes en masse en tant abréviations, qu'unités de mesures ou encore symboles. L'exemple vu dans la partie précédente comportait deux unités lexicales brachygraphiques malgré la brièveté de l'énoncé : "YSS022" et "YSS023". Dans le chapitre 4, nous aurons l'occasion d'observer plus avant les différentes marques de la spécialité que nous retrouvons dans les FA.

Le dernier point sur lequel nous souhaitons appuyer est celui selon lequel les termes spécialisés sont des dénominations de connaissances (Lerat, 1995). Traditionnellement en terminologie, un terme désigne un « concept » ou une « notion », et un concept désigne une « unité de connaissance » (Roche, 2022). Le concept, en fonction des disciplines dans lequel il est évoqué, peut être lui-même défini différemment ou rapproché d'autres notions.

Selon Condamines (2005), en ingénierie des connaissances, le concept prend toute sa place puisqu'il est au centre de la formalisation des connaissances. Un graphe de connaissance va vouloir relier des concepts entre eux par des relations « sémantiques ». Ainsi, de nombreux projets ont eu pour objectif la création de « bases de connaissances terminologiques » reliant ainsi la terminologie et l'intelligence artificielle (Bourigault *et al.*, 2004). En effet, des projets en TAL ont vu le jour à des fins de créations d'outils pour l'extraction de concepts pour alimenter des bases de connaissances. Pour modéliser les informations contenues dans un corpus, trois questions se posent : le type de modélisation visé, le corpus à étudier et les méthodes à utiliser pour arriver à la modélisation (Condamines, 2005). Dans le cadre de cette thèse, nous cherchons à modéliser les éléments du problème que le REX rapporte. Nous travaillons donc avec un corpus déjà décidé car fourni par le CNES. Concernant le type de modélisation, celle-ci est aussi préétablie puisqu'elle s'appuie sur la méthode de résolution de problème TRIZ, que nous présentons dans le chapitre 2. Elle s'éloigne cependant des graphes de connaissances tels qu'envisagés habituellement en ingénierie des connaissances puisqu'elle cherche à modéliser l'anomalie, le dysfonctionnement, et non un réseau étendu de concepts du domaine. Ce travail de thèse se concentre donc sur la troisième question posée par Condamines (2005), c'est-à-dire les méthodes à employer pour la dite modélisation. Nous proposons ainsi l'exploration de diverses approches dans l'étude des REX du domaine spatial, à des fins de modélisation du dysfonctionnement qu'ils expriment.

Le dispositif de REX est donc producteur d'une vaste quantité de données textuelles de façon systématique. Il est pleinement intégré dans les organisations et participe à une culture de sécurité et une maîtrise des risques efficace. Un rapport nécessite à la fois une résolution de l'anomalie, mais aussi une exploitation de l'information pour en empêcher la reproduction. Elle nécessite d'être conservée, mais aussi exploitée sur un temps long. En réponse à ce besoin d'exploitation, un traitement automatique est envisageable. En revanche, les caractéristiques de ces textes de spécialités signifient aussi le besoin de traitements adaptés à ces spécificités. Dans la section suivante, nous explorons donc la question de l'utilisation du TAL pour effectuer un traitement automatique des REX.

1.4 Analyser un REX : le rôle du TAL

La masse de données textuelles produites par un dispositif de REX pose un défi dans une perspective d'exploitation sur le long terme, en-dehors d'une résolution immédiate. Ainsi, le besoin de traiter ces données textuelles de façon automatique s'est fait ressentir. Le traitement de ces REX entraîne lui-même différents besoins. Tout d'abord, il nécessite l'accès

à des données propriétaires. Une organisation peut posséder plusieurs raisons pour ne ne pas rendre accessible ces données : ne pas révéler des informations, voire des failles, à la concurrence, un contenu confidentiel protégé... Il est aussi nécessaire de posséder un savoir expert afin d'accéder à la connaissance véhiculée par les données. Enfin, nous pouvons aussi citer le besoin d'adapter les méthodes et les outils standards du TAL afin de prendre en compte les caractéristiques de spécialités de ces textes. Dans cette section, nous discutons de ces aspects en deux temps. Premièrement, nous explorons l'apport du TAL pour les REX. Puis, nous discutons des différents travaux autour du TAL et des textes techniques, des défis qu'ils posent et des méthodes et des ressources qu'ils développent.

1.4.1 Le TAL pour les REX : pour quoi faire ?

Plusieurs travaux se sont penchés sur les REX en tant qu'objet d'étude. Parmi ces études, nous notons la présence de nombreuses études en rapport avec le domaine aéronautique. Ce domaine bénéficie de plusieurs avantages non négligeables qui expliquent cette tendance. Tout d'abord, la culture de la sécurité dans l'aviation a été encouragée à un niveau institutionnel par le biais de plusieurs agences comme l'OACI (Organisation de l'Aviation Civile Internationale) qui a pour vocation d'aider à la coopération et au partage du ciel pour les pays des Nations Unies, ou bien l'EASA (European Union Safety Agency) spécialisée pour l'Europe et qui a pour but de veiller à la sécurité. Ces agences ont toutes deux un aspect international car le domaine de l'aviation implique par nature de nombreux acteurs de par le monde, entraînant la nécessité de ressources partagées telles que des bases de données, des nomenclatures, des ontologies et aussi des outils. Parmi ces derniers, nous retrouvons le programme ASRS (Aviation Safety Reporting System)[8] qui récolte, traite et analyse des REX soumis volontairement par des pilotes, des contrôleurs aériens et autres acteurs du transport aérien. Ces REX sont ensuite mis à disposition (de façon anonyme) et ont pu servir de corpus dans des études en linguistique ou en TAL comme dans Tulechki (2015).

L'étude de Andreani *et al.* (2013) identifie les tâches dans lesquelles le TAL est un atout pour le traitement des REX :

1. l'aide à la rédaction ;
2. la catégorisation automatique des documents ;
3. la vérification de cohérence intra-document et inter-documents ;
4. la recherche d'information dans une base de données ;
5. le calcul de similarité entre REX ;

8. https ://asrs.arc.nasa.gov/

6. la fouille de données textuelles ;
7. le regroupement de REX ;
8. l'extraction d'informations sémantiques spécifiques ;
9. la constitution de ressources langagières.

Ces deux derniers points (8 et 9), sont présentés comme des aspects transversaux, qui peuvent impliquer les différentes tâches susmentionnées. Dans cette thèse, nous nous situons dans la tâche numéro 8, c'est-à-dire l'extraction d'informations sémantiques spécifiques. Les auteurs la définissent comme :

> « Typiquement, cette tâche consiste à repérer des entités et des relations entre ces entités, leur nature variant selon les domaines et les objectifs : analyse de rapports médicaux pour assister le diagnostic (les entités extraites sont alors les symptômes, les diagnostics, les traitements, etc.), analyse de dépêches financières pour assister le suivi des opérations boursières (les entités sont alors des organismes, des opérations financières, des montants, etc.). »

Ils identifient trois composants généralement utilisés pour l'extraction d'information précise qui sont les systèmes d'étiquetages (segmentation, analyse morphosyntaxique et analyse syntaxique), les ressources lexicales dédiées (liste d'entités nommées, lexiques...) et des règles de reconnaissances des entités et des relations (grammaires locales, ciblées sur le repérage d'éléments d'information spécifiques). Parmi ces trois éléments, nous aurons dans cette thèse l'occasion d'explorer particulièrement le troisième : les règles de reconnaissances des entités. Nous réalisons ce travail, non par le biais de grammaires locales, mais plutôt en nous appuyant sur la sémantique des cadres et la ressource *FrameNet*, que nous présentons dans le chapitre 3 et dont nous verrons la mise en application (et les limites) dans le chapitre 8 et 9. Les deux premiers éléments qui sont les systèmes d'étiquetage et les ressources lexicales dédiées ne seront en revanche pas exploités pour cette thèse. Dans le chapitre 4, nous verrons en effet que les systèmes d'étiquetages usuels montrent des résultats peu satisfaisant sur un corpus aussi technique et bruité, phénomène récurrent dans les textes techniques sur lequel nous revenons dans la section suivante. Quant aux ressources dédiées, nous n'avons pas eu accès à des lexiques, ontologies... spécifiques aux REX dont nous faisons l'étude. Cette thèse se voit donc aussi comme un moyen d'explorer les différentes méthodes pour exploiter ce corpus, à des fins d'extraction des informations sémantiques en l'absence de ces ressources.

L'étude de Tanguy *et al.* (2016) a mis en application certaines de ces tâches (2, 6, 5, 9) sur le corpus des REX de l'ASRS et d'ECCAIRS (European Coordination Centre for Accident and Incident Reporting Systems). Le premier comprend des écrits en anglais alors que le second inclut de l'anglais et du français. Pour la tâche de classification automatique de documents

(2), les auteurs ont utilisé une méthode d'apprentissage automatique supervisée[9], le SVM (Support Vector Machine) (Fan *et al.*, 2008). Une approche non supervisée a aussi été utilisée afin d'effectuer du *topic modeling* sur les données, dans l'objectif de repérer des thématiques sous-jacentes dans le texte grâce à cette méthode de fouille de textes (6). Ces deux tâches ont obtenu des résultats satisfaisants mais les auteurs ont tout de même noté la redondance de la tâche de fouille de texte étant donné la complétude des métadonnées déjà existantes et définies dans l'organisation ayant mis à disposition le corpus d'étude. Ces deux tâches démontrent néanmoins l'efficacité de l'utilisation de techniques de TAL sur des REX. La troisième tâche évaluée lors de cette étude consiste en l'utilisation d'un outil afin d'effectuer un rapprochement de documents par similarité textuelle (5). L'évaluation qualitative a notamment montré que les utilisateurs experts ont fait un usage de l'outil différent de celui attendu par les auteurs. Ce détournement » de l'outil montre la nécessité de l'intégration d'un expert lors du travail avec des textes spécialisés. À cet effet, les auteurs ont effectué une troisième tâche d'*active learning*, c'est-à-dire un système permettant aux utilisateurs, lors de l'utilisation de l'outil, de catégoriser les éléments. Ils deviennent ainsi des annotateurs et produisent des ressources supplémentaires (9).

Toujours dans le domaine de l'aviation, nous pouvons noter les travaux de Persing et Ng (2009) qui se concentrent sur une tâche de classification de documents selon leur cause, ou l'identification de pourquoi un événement a eu lieu qui, comme nous en avons discuté dans la section 1.2.3, est un pan important de la démarche de REX. Dans cet article, les facteurs identifiés comme contributeurs à la difficulté de la tâche sont :

— la quantité de données annotées disponibles ;
— la multiplicité des classes ;
— la catégorisation *multi-label* (c'est-à-dire la possibilité de plusieurs classes par document) ;
— une distribution des documents par classe inégale ;
— l'appartenance des documents à un même domaine (ici des recommandations de films) impliquant une similarité dans le vocabulaire.

Ces difficultés sont par ailleurs parties prenantes de notre étude, excepté la catégorisation *multi-label*, même si nous retrouvons des rapports d'anomalies qui dépendent de plusieurs classes à la fois. L'apport de ces travaux a été de proposer un algorithme pour pallier le problème du manque de données annotées. Cet algorithme vise à augmenter les données utilisées pour l'entraînement. Pour ce faire, il considère un ensemble de documents non annotés. Parmi ces documents, il identifie les mots similaires à ceux d'une classe donnée. Si ces mots

9. L'apprentissage supervisé signifie que les documents sont annotés manuellement. Le modèle apprend les caractéristiques de cette annotation afin de le reproduire sur des documents qui lui sont inconnus

sont suffisamment caractéristiques d'une classe, le document sera automatiquement annoté comme appartenant à la classe en question. Cette méthode a montré une augmentation significative des résultats de la classification automatique effectuée par la suite avec, à nouveau, un *SVM*. Plusieurs méthodes pour contourner les facteurs de difficultés au traitement des REX susmentionnés ont déjà été développées. Dans notre cas, nous choisissons de faire usage des données telles quelles, sans augmentation artificielle. Dans le chapitre 5, nous présentons les différentes classes, issues de la typologie que nous avons développée. Dans le chapitre 6, nous verrons comment la répartition inégale des documents dans les différentes classe, et le parti pris de ne pas effectuer d'augmentation artificielle a pu impacter les résultats.

Les REX d'Ariane 5 ont déjà fait l'objet de travaux en TAL. En effet, dans Galand *et al.* (2018) une première analyse de ces REX avec des techniques de TAL est effectuée, dans l'objectif de circonvenir à l'analyse manuelle habituelle extrêmement chronophage. Ces premiers travaux se sont concentrés sur des REX qui diffèrent de notre corpus, des Fiches de Points Critiques, et ont exploré l'application de trois tâches de TAL : le *topic modeling* afin de classifier les documents sans classes préétablies, la classification de documents vers des classes définies, la recherche de similarités afin de retrouver des problèmes antérieurement rencontrés. Pour ces deux premières tâches, des méthodes par sac-de-mots ont été utilisées, alors que la dernière a fait le cas d'une approche par sac-de-mots et avec des *embeddings* (plongements de mots) avec Word2Vec (Mikolov *et al.*, 2013). Ces tâches ont montré des résultats prometteurs qui prouvent la pertinence de l'utilisation du TAL sur ce corpus. Elles offrent un premier aperçu de ce qu'une approche en surface du texte peut donner.

Une seconde étude a été effectuée sur les FA d'Ariane 5 par Kurela *et al.* (2020). Celle-ci s'est concentrée sur une tâche de classification des FA en fonction de leur degré de gravité. Les auteurs ont testé plusieurs modèles de classification automatique afin de repérer le plus efficace. Ils ont ensuite fait varier la quantité de champs textuels (champ libre ou métadonnées) pris en compte et ont observé une amélioration suite à l'ajout de ces informations supplémentaires. Avant cela, ils ont tout de même effectué un certain nombre d'opérations de prétraitements avec un système par règles afin de nettoyer, au moins en partie, le corpus de travail. Ces travaux se sont avérés concluants et ont permis d'identifier des pistes prometteuses, ainsi que des pistes d'amélioration pour la classification des FA en fonction de leur gravité.

Ces travaux ont été les prémices de ce que nous présentons dans cette thèse. En effet, comme précédemment mentionnés, nous étudions aussi des FA et leur capacité à être traité par des techniques de TAL. Cependant, cette thèse s'est vue comme l'opportunité d'explorer ce corpus avec un grain plus fin, un niveau de proximité avec le texte et sa structure plus

élevée. Nous retrouvons donc les mêmes difficultés de traitement que celles énoncées tout au long de cette partie, mais testerons plusieurs approches, avec et sans nettoyage de données. Par ailleurs, nous emploierons des techniques neuronales mais aussi de linguistique de corpus afin d'explorer plus en détail les énoncés auxquels nous sommes confrontés (cf. chapitres 6, 8 et 9).

1.4.2 TAL et textes techniques

Les LLM (Large Language Models) sont aujourd'hui omniprésents dans la sphère du TAL, depuis la sortie de BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al.*, 2019). Ces modèles neuronaux, contextuels, et pré-entraînées, ont l'avantage d'avoir été entraînés sur des corpus conséquents et ont acquis la capacité de généraliser des représentations linguistiques utiles à de nombreuses tâches de traitement automatique des langues. Ils obtiennent des résultats performants sur une grande variété de tâches de TAL (Jurafsky et Martin, 2024). En revanche, ces modèles ont été entraînés sur des données variées et non contrôlées issues du Web, et ont vite montré leurs limites dans leurs performances lors de tâches effectuées sur des corpus spécialisés. En effet, le peu de données accessibles pour les domaines de spécialités rend impossible l'entraînement d'un LLM uniquement sur des données spécialisées (El Boukkouri *et al.*, 2019). Ainsi, des modèles ayant été *fine-tunés* (c'est-à-dire qu'un modèle déjà existant a été partiellement réentraîné sur le corpus d'étude) ont vu le jour. On peut citer, parmi d'autres, BioBERT (Lee *et al.*, 2019) et SciBERT (Beltagy *et al.*, 2019), respectivement adaptés au domaine biomédical et aux articles scientifiques. Par ailleurs, il a été démontré qu'un modèle entraîné spécifiquement pour un domaine n'est pas nécessairement plus performant qu'un modèle *fine-tuné* (El Boukkouri, 2020). Dans cette partie, nous nous proposons de revenir sur quelques travaux récents ayant été effectués, soit dans le domaine de la maintenance, soit dans le domaine du spatial, et ce, afin de présenter les problématiques rencontrées dans le traitement de données similaires (dans une certaine mesure) aux nôtres. Nous pourrons voir qu'une des problématiques de ces domaines est le manque de ressources langagières, ressources pourtant jugées nécessaires pour le traitement de ce type de données comme démontré par Dima *et al.* (2021) et le manque d'outils pour les traiter.

Si nous prenons le temps de nous arrêter sur le traitement de textes de maintenance, c'est parce que nous avons pu observer une similarité entre ces textes et les données de notre corpus. Dans l'article de Dima *et al.* (2021), le corpus utilisé est composé de *maintenance work orders*. Ce sont des « bons de commande » d'opérations de maintenance qui doivent être (ou ont été) exécutés au sein d'une organisation. Dans la Table 1.1, exemple 1, nous pouvons

Numéro	Référence	Exemple
1	Dima *et al.* (2021)	RESET FAN FAILS AND START EQUIOPMENT
2	Bikaun *et al.* (2024b)	air freight for pump TBC
3	FA : description du problème	VENTILATION COIFFE : VANNE GVC 8130 BLOQUEE.

TABLE 1.1 – Exemples de données de maintenance et d'une FA

observer un texte tout en majuscules (ce qui n'est pas systématique dans leur corpus) et une faute d'orthographe au mot "EQUIOPMENT" ("equipment"). Ce type de phénomène se retrouve aussi dans les FA (troisième ligne) où on a cette fois un texte systématiquement en majuscules et des fautes typographiques récurrentes. Les deux textes sont aussi très courts (pour les FA, ces textes peuvent être un peu plus longs mais restent relativement brefs avec une moyenne de 19 mots pour la description de l'anomalie dans notre corpus).Dans cet article Dima *et al.* (2021) nous informent aussi sur l'importance de l'utilisation de ressources externes telles que des taxonomies, des *thesaurus*, des ontologies. Ils recommandent aussi l'utilisation de *fortuitus data* telles que défini par Plank (2016), c'est-à-dire, des métadonnées ou des données supplémentaires pouvant venir, par exemple, d'autres tâches ou expérimentations sur ces données. Nous retrouvons aussi la mise en avant de la présence indispensable d'un expert lors du traitement de données techniques. Ces éléments sont exposés dans la pipeline définie par Dima *et al.* (2021) et Brundage *et al.* (2021) dans la Figure 1.3. Elle montre bien le besoin conjoint de deux pôles, avec d'un côté l'expert et les ressources du domaine, et de l'autre, le linguiste/taliste, le texte, et les méthodes et les outils de traitement.

Le projet *MaintIE* Bikaun *et al.* (2024b) a aussi pour corpus des textes de maintenance. Ils travaillent avec des *maintenance short texts*, textes dérivés des *maintenance work orders*, qui récapitulent les éléments clés de ces derniers. Nous pouvons observer dans l'exemple 2 de la Table 1.1 que le texte est, lui aussi, bref et contient des acronymes, ici "TBC". La question d'un prétraitement afin de normaliser les données s'est donc posée. Le corpus a été nettoyé de deux manières, la première visant à normaliser le lexique à l'aide du corpus *MaintNorm* créé spécifiquement pour la normalisation des *Maintenance Short Text* (Bikaun *et al.*, 2024a). La seconde consiste au masquage des mots sans portée sémantique ou des informations confidentielles. Dans ces travaux, les auteurs ont fait effectuer une annotation fine par des experts. Nous observons donc à nouveau le recours à des ressources langagières dédiées et des experts pour le traitement de données techniques.

L'annotation effectuée est de type *semantic role labeling* (Gildea et Jurafsky, 2002), qui vise donc à annoter les différents éléments du texte selon leur rôle sémantique et répondre

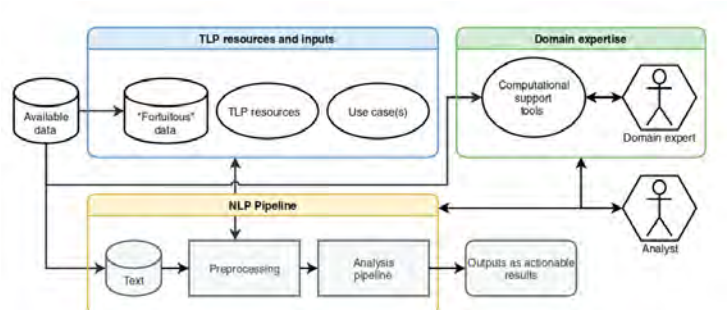


FIGURE 1.3 – Schéma de Dima *et al.* (2021) présentant la pipeline pour le TLP

ainsi à la question de "who is doing what to what, and why ?". Ainsi, un corpus annoté [10] en entités et en relation a été mis à disposition. Nous retrouvons des similarités entre cette approche et celle que nous cherchons à mettre en place dans la thèse. En effet, afin de guider l'analyse fine des rapports d'anomalies, nous nous appuyons sur la sémantique des cadres pour laquelle il est aussi question de rôles sémantiques, et de *frame semantic role labeling* pour l'annotation automatique en cadre sémantique. Nous revenons sur ces notions essentielles plus avant dans le chapitre 3, et proposons une approche de type *frame semantic role labeling* dans le chapitre 9.

Dans l'étude de Usuga Cadavid *et al.* (2022), les travaux effectués sont toujours dans le domaine de la maintenance mais portent sur un corpus français. Les auteurs ont utilisé un modèle neuronal entraîné pour le français, *CamemBERT* (Martin *et al.*, 2020), afin d'effectuer deux tâches de classification. Pour se faire, ils ont fait usage de la méthode de *transfer learning* (Pan et Yang, 2010). Celle-ci consiste a appliquer un modèle entraîné sur des données différentes du corpus d'étude. Elle est fondée sur le principe qu'un modèle appris sur des données est capable de suffisamment généraliser ses « connaissances » sur la langue pour être efficace sur un corpus différent, et implique la possibilité d'obtenir un modèle performant sur des corpus avec très peu de données. Elle présente deux modalités, une approche *feature-based* (dans laquelle le modèle est appliqué tel quel et sert d'extracteur de caractéristiques avant l'application d'un classifieur qui est spécifié sur les données) et du *fine-tuning* (dans laquelle on réentraîne les couches supérieures du modèle pour le spécialiser sur nos données annotées). Dans cet article, les deux modalités ont été appliquées pour les deux tâches. La méthode du *fine-tuning* (que nous utiliserons dans nos travaux) a montré de meilleurs résultats dans les deux cas. Par rapport à ce que nous avons pu voir des langages de spécialités, nous savons que ces dernières peuvent être considérées comme un sous-ensemble de la langue générale, dans lequel on retrouve à la fois un fonctionnement

10. https ://github.com/nlp-tlp/maintie

similaire à cette dernière, mais aussi d'éventuelles règles langagières qui lui sont spécifiques. La méthode du *transfer learning* fait donc sens puisque, malgré l'utilisation d'un corpus qui diffère des données utilisées pour l'entraînement du modèle, le principe de généralisation implique une reconnaissance des systèmes langagiers utilisés dans la langue générale et qui se retrouvent en langue de spécialité, notamment pour l'approche *feature-based*. L'approche par *fine-tuning* permet elle d'entraîner le modèle sur phénomènes langagier spécifiques du corpus d'étude. De ce fait, les systèmes (et phénomènes) langagiers spécifiques à un langage de spécialité, malgré une faible représentation, voire une inexistence dans la langue générale, pourraient être reconnus par le modèle.

Les travaux de Gómez-Pérez *et al.* (2022) font état des mêmes difficultés précédemment évoquées quant au traitement de données de spécialités, et notamment du spatial, c'est-à-dire le manque de données annotées, de modèles pré-entraînés pour un domaine, auxquels ils ajoutent les problèmes d'explicabilité des modèles[11] ainsi que les problématiques liées au coût environnemental du développement et de l'application de ces modèles. Les auteurs proposent aussi, à travers quatre cas d'étude, un *framework* pour l'utilisation du TAL dans le domaine du spatial. Celui-ci vise à proposer une approche symbolique, automatique, ou hybride en fonction du cas d'étude, mais aussi des ressources disponibles dans le domaine.

Dans cette partie, nous avons présenté un aperçu des projets de recherche se focalisant sur deux domaines proches du nôtre : la maintenance et l'espace. Nous avons pu observer que, malgré une variété de travaux, ils restent des domaines avec des difficultés de traitement propres aux langues de spécialité. Ces recherches récentes montrent les limites du TAL sur le traitement des corpus spécialisés, avec notamment un manque de ressources langagières et d'outils adaptés.

1.5 Conclusion

À travers les différentes sections de ce chapitre, nous avons pu explorer le REX sous différents aspects. Tout d'abord, nous avons pu voir comment le REX est une démarche ancrée dans une pratique plus générale de culture de sécurité pour la maîtrise des risques. En effet, c'est une démarche systématique qui s'est mise en place afin de résoudre des problèmes et empêcher leur reproduction, mais aussi afin de faire émerger des dysfonctionnements systémiques et/ou des anomalies avant même de constituer un véritable incident (ou plus). Ainsi, le corpus avec lequel nous travaillons est constitué de Fiches Anomalies et ne fait pas

11. L'explicabilité fait référence à la capacité à décrire et expliquer les mécanismes des modèles d'apprentissage automatique, souvent qualifiés de boîtes noires.

le rapport d'accidents ou d'incidents. Il fait état de situations qui diffèrent de l'attendu et dont le niveau de gravité peut varier, dont la présence de faits rapportés qui peuvent paraître anodins.

En second lieu, nous avons abordé le REX sous l'angle de la linguistique. Cela nous a permis d'examiner la notion de « langage de spécialité » et de quelle façon le REX y correspond. Par ce biais, nous avons mis en avant la notion de contexte avec les deux éléments que sont le domaine et l'expert. Ces derniers montrent une importance indéniable dans la caractérisation de ces textes. Par ailleurs, nous avons mis l'accent sur plusieurs caractéristiques lexico-sémantiques pertinentes à l'étude des REX comme la monosémie/polysémie et les termes spécialisés comme dénomination de connaissances.

Enfin, en dernier lieu, nous avons exploré le REX par rapport au TAL. En effet, le TAL peut venir en réponse à différents besoins de traitement des REX par le biais de diverses tâches (dans notre cas, l'extraction d'information fine). Cela nous a aussi permis, à travers différentes études, de faire ressortir deux composantes récurrentes au traitement de données techniques similaires à notre corpus que sont le recours à un expert et à des ressources langagières dédiées. Notre étude se place dans la lignée de ces dernières mais passe par une utilisation limitée du temps d'experts et l'absence de ressources langagières spécifiques au corpus. De ce fait, cette thèse se voit comme l'exploration de différentes approches de ces REX spécialisés avec une utilisation parcimonieuses d'expertise du domaine en se fondant sur des méthodes de linguistique de corpus et de TAL.

Chapitre 2

Formaliser un problème : la méthode TRIZ

Sommaire

2.1 Introduction : TRIZ comme point d'arrivée

Cette thèse s'inscrit dans un contexte de partenariat entre l'unité de recherche qu'est le laboratoire CLLE, le CNES et MeetSYS[12]. Elle se doit donc de répondre à la fois à des problématiques de recherche, mais aussi à des objectifs industriels. Dans ce contexte, l'objectif principal mis en avant est d'explorer l'adéquation de la méthode de résolution de problèmes qu'est TRIZ (Théorie de Résolution des Problèmes Inventifs) avec un corpus de REX. Cette

12. https://www.meetsys-i2kn.com/

théorie recouvre un champ très large. Elle propose des concepts, des méthodes et des outils, et a fait l'objet de nombreuses évolutions depuis sa création. Elle est aussi considérée comme difficile à implémenter et Bultey *et al.* (2007) cite trois raisons à cela : une traduction depuis le russe qui a souffert d'une mauvaise interprétation (Cavallucci *et al.*, 2002), une application pratique difficile étant donné la multiplicité des outils (Mann, 2002) et l'augmentation exponentielle des découvertes en physique et en ingénierie que doit en conséquence posséder l'utilisateur de la méthode. Les auteurs répondent à ce dernier point par l'utilisation de l'Intelligence Artificielle (IA) pour assister l'utilisateur, et notre recherche s'inscrit dans cette lignée et plus précisément, dans un objectif de résolution de problèmes techniques. L'avantage de TRIZ est que la méthode proposée passe par une abstraction des éléments en jeu. Elle permet ainsi, de contourner certains aspects, comme le lexique spécialisé hautement technique, en proposant une représentation du monde physique indépendant du domaine d'application. La question focale de cette thèse est donc de faire le lien entre notre texte composé de REX, et TRIZ. Cette question en entraîne une autre : comment décrire un problème technique afin de pouvoir utiliser TRIZ ? Pour répondre à cette question, nous avons choisi de faire usage d'un des outils de TRIZ qu'est la modélisation sous forme de *vépole* afin de représenter un problème technique et que nous décrivons dans la section 2.3.1. Mais avant cela, nous faisons un tour d'horizon de cette théorie.

Dans ce chapitre, nous commençons donc par présenter les principes fondateurs de TRIZ et quelques-uns de ses principaux outils afin d'expliciter le contexte théorique et méthodologique dans lequel s'inscrit ce mode de résolution de problèmes. Cette présentation n'a pas visée d'exhaustivité et s'envisage plutôt dans une perspective de vulgarisation. Nous ne prétendons pas nous-même à une expertise poussée en la matière et avons bénéficié de l'appui de MeetSYS pour son utilisation. Dans un second temps, nous nous focalisons sur l'implication de TRIZ dans notre travail en présentant d'abord la modélisation en *vépole* que nous utilisons pour représenter le problème dont il est question dans le texte. Puis, nous présentons la base de connaissances I4KN développée par MeetSYS qui permet de faire usage des informations issues de cette modélisation. Enfin, nous présentons quelques travaux qui ont déjà exploré l'interaction entre le TAL et TRIZ, et les similitudes et les différences qu'ils présentent avec nos propres travaux.

2.2 Une introduction à TRIZ

TRIZ (*Teorija Rezhenija Izobretatelskih Zadach*) est traduit en français par « Théorie de Résolution des Problèmes Inventifs ». Cette théorie a été développée par Genrich Altshuller,

un scientifique et ingénieur russe et ses collègues d'après l'étude de brevets[13] et à travers le repérage de « patrons » et de régularités concernant la résolution de problèmes, la création de nouvelles avancées technologiques et l'innovation technologique (Ilevbare *et al.*, 2013). Cavallucci (1999) établit quatre sources différentes aux travaux d'Althsuller qui sont : l'analyse de brevets, l'analyse de comportement psychologiques des inventeurs, l'analyse des outils et des méthodes existants et l'analyse de la littérature scientifique. Livotov (2008) et Souchkov (2019) identifient la date de la première publication présentant les travaux d'Altshuller à 1956 avec l'article « About Technical Creativity ». Cette méthode marque, selon lui, une rupture avec le système de résolution de problème et d'avancée technologique précédemment utilisé basé sur l'expérience : le « *trial and error* ». Altshuller définit TRIZ comme étant une « science de l'invention »[14].

Cette « nouvelle science » repositionne la conception d'inventions sous l'angle de la résolution de problèmes (Rousselot *et al.*, 2008). Selon Ragab-Zouaoua (2012), on peut retrouver de nombreuses applications de TRIZ dans le champ de la résolution de problème technologique avec 1550 entreprises recensées qui font l'usage (Cavallucci, 1999). En revanche, on en retrouve peu l'utilisation dans les problèmes non technologiques. D'après Ragab-Zouaoua (2012), on peut tout de même trouver quelques études sur son utilisation dans le champ de l'éducation, l'art, l'économie, la politique...

Parmi les propositions de cette théorie, on retrouve un ensemble de principes, notions fondatrices ainsi que de nombreux outils qui prennent la forme de matrices, d'algorithmes, de schémas et de méthodes à mettre en application. Ces derniers se sont développés sur des années, au fur et à mesure des travaux d'Altshuller et de son équipe, et sont issus de plusieurs ouvrages qui ne nous sont en revanche pas tous accessibles. De nombreux travaux ont repris ceux d'Althuller afin de parfaire, développer et étendre ses propositions.

2.2.1 Principes généraux

Selon Moehrle (2005), TRIZ est née avec l'idée d'utiliser les connaissances d'anciens inventeurs afin de soutenir les ingénieurs dans la résolution des problèmes techniques. La théorie propose ainsi un certain nombre de concepts, que nous développerons dans cette partie, mais aussi d'outils pour les mettre en œuvre. Ilevbare *et al.* (2013) citent plusieurs auteurs qui mettent en avant les différents aspects de TRIZ :

— Savransky (2000) met en valeur l'aspect méthodologique comme étant basé sur la connaissance ;

13. Une première version est issue de l'étude de 40 000 brevets. Dans un second temps, 400 00 brevets ont été analysés. Suite à la chute du mur de Berlin, la méthode s'est répandue et plus de 3 millions de brevets ont été étudiés.

14. https://www.youtube.com/watch?v=1uFWSk2W3Go&t=1225s

— Fey et Rivin (2005) reconnaissent les principes dégagés qui décrivent l'évolution des systèmes et des technologies mais appuient l'objectif de création de nouveaux systèmes techniques ;
— Gadd et Goddard (2011) mettent en avant l'ensemble des outils qui rendent compte des différents aspects de la compréhension et de la résolution des problèmes techniques ;
— Livotov (2008) la décrit comme « the most comprehensive, systematically organized invention knowledge and creative thinking methodology known to man ».

Par ces différents aspects, les auteurs cherchent à prouver que TRIZ n'est pas juste une théorie ou un ensemble de principes. Elle est surtout une méthodologie basée sur une analyse systématique de la connaissance des systèmes techniques. Elle permet la résolution de problèmes techniques et la création de nouveaux systèmes techniques de façon inventive. Cavallucci (1999) met aussi en avant l'un des objectifs de TRIZ qui est de lutter contre « l'inertie psychologique ». Cette inertie a pour cause des « habitudes, des compétences trop pointues dans un domaine particulier et [...] le « jargon » du spécialiste » limitant à l'expérience la recherche de solutions à un problème. L'inertie psychologique est présentée comme un « frein à la créativité ».

La notion de « problème inventif » est centrale à TRIZ. Altshuller, d'après Ragab-Zouaoua (2012), définit les « problèmes inventifs » comme des « problèmes sans solution connue ou des problèmes dont la solution, généralement connue et acceptée, génère d'autres problèmes inventifs »[15]. Pour résoudre les problèmes techniques, il propose une approche qui passe par une abstraction du problème, pour aller vers sa résolution, plutôt qu'une pratique « à tâtons » d'essais (et d'échecs) jusqu'à trouver une solution. La Figure 2.1 illustre ce procédé. Nous y retrouvons un problème technique pour lequel il est nécessaire de trouver un équivalent abstrait en ramenant chaque élément du problème à une forme plus élémentaire. À cette forme élémentaire est associé un ensemble de solutions prédéfinies. Ces solutions sont constituées des principes d'inventions, des lois des évolutions techniques et des standards inventifs proposés par TRIZ, Gadd et Goddard (2011) d'après Ilevbare *et al.* (2013), que nous développons ci-dessous. De cet ensemble de solutions, est tirée une solution spécifique pour résoudre le problème. En passant par cette étape d'abstraction, le problème peut donc être envisagé selon un autre angle. Des problèmes qui paraissent différents au premier abord sont rapprochés puisque leur forme élémentaire s'avère similaire et les possibilités de solutions en deviennent plus étendues. (Ilevbare *et al.*, 2013)

15. Nous utilisons aussi dans ce chapitre l'appellation problème technique pour renvoyer aux problèmes inventifs

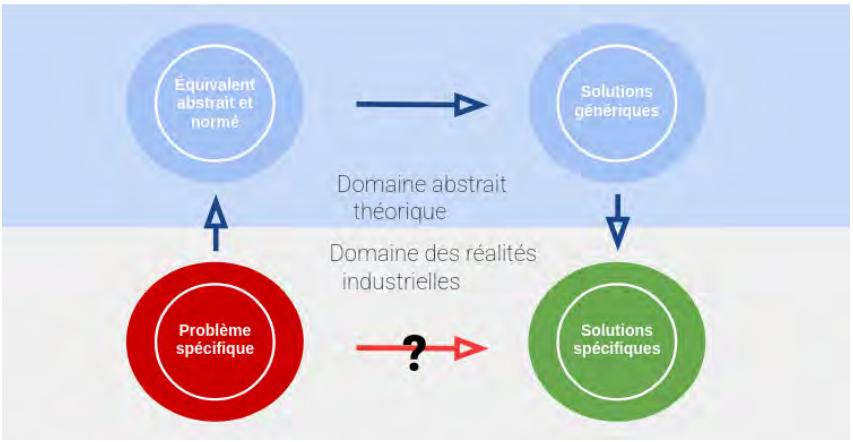


FIGURE 2.1 – Résolution de problème « à la TRIZ » d'après (Laforcade, présentation non publiée, 12 octobre 2024) adaptée de (Gadd et Goddard, 2011; Savransky, 2000)

Shulyak (1998) définit un système technique comme « tout ce qui effectue une fonction » et donne en exemple une voiture, un stylo ou même un livre. Ces deux derniers exemples correspondent à des systèmes très simples, mais chaque système peut lui-même être composé de plusieurs sous-systèmes. Ces systèmes et sous-systèmes s'organisent en hiérarchie des systèmes techniques en fonction du nombre d'éléments qui interagissent. L'auteur prend l'exemple de la voiture et des sous-systèmes qui la composent tels que le moteur, le mécanisme de pilotage, les freins... Dans la Table 2.1 traduite de Shulyak (1998), nous pouvons observer la hiérarchie du système technique "Mode de transport" et sa décomposition en sous-systèmes avec l'exemple de la voiture et de ses freins[16]. Le sous-système du frein est décomposé jusqu'au système le plus élémentaire de la « liaison chimique » qui implique l'interaction de deux éléments uniquement.

En relation avec ces systèmes techniques, Altshuller définit trois principes fondateurs de TRIZ :

1. les lois d'évolution ;
2. « l'idéalité » ;
3. la contradiction.

Selon Altshuller, l'évolution des systèmes techniques est gouvernée par des « **lois d'inventions objectives** ». Ces lois ont pour objectif de « prédire, conduire, suggérer et initier des pistes de création, en parfaite cohérence avec la dynamique qui a toujours présidé au progrès technique ». Elles ont pour rôle d'aider l'ingénieur dans le choix des directions à

16. Cet exemple est incomplet et sert surtout à illustre la décomposition en sous-système et la hiérarchie des systèmes techniques

Systèmes techniques	**Sous-systèmes des systèmes techniques**				
Mode de transport	Voitures	Routes	Cartes	Conducteur	Stations services
Voiture	Motorisation	Freins	Chauffage	Direction	Éléctrique
Freins	Pédale de frein	Cylindres hydrauliques	Fluide	Plaquettes de freins	
Plaquettes de freins	Plaquette	Support	Rivets		
Plaquette	Particule A	Particule B	Liaison chimique		
Liaison chimique	Molécules A	Molécules B			

TABLE 2.1 – Systèmes et sous-systèmes des freins d'une voiture

suivre (Ragab-Zouaoua, 2012). Elles mettent en avant un principe clé de la philosophie de TRIZ selon lequel le procédé d'amélioration d'une partie d'un système qui est déjà performant entraîne systématiquement un conflit avec un autre système technique. Afin de résoudre ce conflit, le système technique le moins évolué des deux se verra amélioré. C'est ce système d'améliorations successives qui mène vers la forme « idéale » du système (Shulyak, 1998) (nous préciserons ci-dessous le principe d'idéalité chez Altshuller).

Le principe d'« **idéalité** » d'un système technique est développé dans Altshuller *et al.* (1999). Pour chaque système technique, il faut pouvoir en définir la « machine idéale », c'est-à-dire la version idéale de ce système, version qui n'existe pas, mais vers laquelle il faut tendre. Selon la loi de l'idéalité, chaque version du système est une amélioration qui le rapproche de cette « machine idéale ». Ainsi, le degré d'idéalité qu'atteint un système dépend de l'écart entre cette machine idéale et la version existante du système, et Ilevbare *et al.* (2013) nous en donne l'équation dans la Figure 2.2. Les bénéfices (« *benefits* ») représentent les fonctions que remplit le système, les dommages (« *Harms* ») sont les résultats non désirés du système et les pertes, et le coût (« *Costs* ») sont les coûts en ressources.

$$Ideality = \frac{\Sigma\ Benefits}{(\Sigma\ Costs + \Sigma\ Harms)}$$

FIGURE 2.2 – Équation de l'idéalité de Ilevbare *et al.* (2013)

Toujours selon Altshuller *et al.* (1999), pour tendre vers cette « machine idéale », il est

nécessaire de résoudre les « **contradictions** » qui existent au sein du système technique. Le principe de contradiction est une caractéristique primordiale de TRIZ et son identification est une étape nécessaire pour la résolution du problème (Moehrle, 2005). Choulier et Draghici (2000), cités par Ragab-Zouaoua (2012), précisent qu' « à l'origine de tout problème d'inventivité, on trouve une contradiction ». Les contradictions se déclinent en deux catégories : les contradictions techniques et les contradictions physiques.

Les contradictions techniques apparaissent lorsque l'amélioration d'un paramètre d'un système technique entraîne une détérioration d'un autre paramètre. Ilevbare *et al.* (2013) donne en exemple une voiture dont on modifie le moteur afin de le rendre plus puissant. L'objectif est d'augmenter la vitesse de la voiture. Or, ce même moteur rend la voiture plus lourde et de ce fait, limite la vitesse de la voiture. Nous obtenons une contradiction technique entre l'augmentation de la vitesse (paramètre 1) et l'augmentation de la masse (paramètre 2).

Les contradictions physiques concernent les cas où une incohérence existe entre les valeurs d'un paramètre physiques (ou techniques) d'un même système. Nous pouvons prendre l'exemple d'une table comme système technique. Nous souhaitons que cette table soit épaisse. Mais elle se doit aussi d'être fine, sous peine de la rendre trop lourde, de détériorer le paramètre « masse » de l'objet. De ce fait, il existe une contradiction physique entre deux valeurs du même paramètre du système [17].

Ces différents principes sont considérés comme fondateurs de la méthode TRIZ. Toute résolution de problème « à la TRIZ » fait appel au moins à l'un d'entre eux (Ilevbare *et al.*, 2013).

2.2.2 Les outils

Afin de résoudre les problèmes techniques, la méthode TRIZ propose tout un ensemble d'outils et de techniques développés au fur et à mesure de l'élaboration de la méthode. Or, ces outils ne sont pas accompagnés d'une procédure d'application, de comment et quand les utiliser (Eversheim (2009) cité par Ilevbare *et al.* (2013)). Ainsi, Moehrle (2005) propose un *framework* qui structure l'utilisation de ces outils.

Le *framework* proposé en vue de guider l'utilisation des techniques développées dans la méthode TRIZ s'organise à travers cinq questions [18] que l'auteur considère comme tous les aspects qu'il faut couvrir lors de l'analyse et la résolution de problèmes :

1. l'état actuel : quel est l'état actuel de la situation ?

17. Une solution dans cet exemple serait d'effectuer des nervures sur le plateau
18. Les cinq questions détaillées ci-dessous sont notre traduction.

2. les ressources : quelles sont les ressources disponibles ?
3. les objectifs : quels sont les objectifs à atteindre ?
4. l'état désiré : à quoi devrait ressembler la situation future ?
5. transformation : de quelle façon l'état actuel peut-il être transformé en l'état désiré ?

Bertoluci (2001) propose aussi une manière de catégoriser les différentes techniques proposées par TRIZ. Cette catégorisation se fait en fonction des cinq étapes constitutives d'une méthode de résolution de problèmes (MRP) telles que définit par Dupont et Verven (1999) : identifier, formuler, résoudre, évaluer, implémenter. La Table 2.2 reprise de Bertoluci (2001)[19] offre un panorama (non exhaustif) des différents principes, outils et techniques proposés par TRIZ en fonction des objectifs visés lors d'une résolution de problème technique. Nous proposons désormais d'explorer plus avant quelques-uns de ces outils les plus reconnus.

Étapes d'une MRP	Étapes de TRIZ	Outils et techniques de TRIZ
Identifier	Établir les causes réelles du problème et les ressources disponibles	RCA (Root Cause Analysis) **Hommes miniatures**
Formuler	Modéliser le problème	Contradiction Technique **Vépoles** Contradiction Physique Lois d'évolution
Solutions	Rechercher des solutions créatives	**Matrice des contradictions** **40 Principes d'invention** **Solutions standards** Principes de séparation **Hommes miniatures** Analyse des ressources du système Lois d'évolution des systèmes Opérateurs STC Bases de données effets
Évaluer	Sélectionner une solution	Résultat Final Idéal
Implémenter		

TABLE 2.2 – Synthèse des étapes de TRIZ de Bertoluci (2001)

19. Nous avons gardé uniquement les éléments propres à la version dite « classique » de TRIZ et n'avons pas repris les outils issus de la méthode, mais développés ultérieurement par des parties tierces.

2.2.2.1 Les 40 principes d'invention et la matrice de contradiction

Comme mentionné précédemment, l'analyse de brevets effectuée par Altshuller a fait état de « principes d'inventions », qu'il considère comme étant à la base de toute invention technique. Malgré la multiplicité des inventions, elles ne dépendraient que de 40 principes basiques (Moehrle, 2005). Parmi ces principes, nous pouvons retrouver par exemple :

- le principe d'universalité : un objet doit remplir plusieurs fonctions afin de limiter le besoin d'autres objets ;
- le principe pneumatique et hydraulique : le remplacement de parties solides par des éléments gazeux ou liquides ;
- le principe de modification de paramètre : l'implémentation d'une modification parmi le changement de l'état physique, la modification de la concentration ou de la consistance, le changement du degré de flexibilité ou la modification de la température (Altshuller et Seredinski, 2004) et (Rantanen et Domb, 2002) à travers (Almansba et Braham Chaouche, 2013).

Ces principes se retrouvent dans un autre outil de TRIZ : la matrice des contradictions [20]. Cette matrice classe les principes d'inventions présentés ci-dessus (Livotov, 2008) et les organisent en réponse aux paramètres qui entrent en contradiction. Elle est composée de 39 lignes et colonnes qui représentent des paramètres de systèmes techniques (la masse, la température, le volume...). Elle fonctionne de la façon suivante :

1. Identification de la propriété à améliorer (lignes) ;
2. Identification de la conséquence néfaste entraînée par 1 (colonnes) ;
3. Étude des solutions proposées dans la case correspondant à la ligne et la colonne associées aux points 1 et 2.

Cet outil permet ainsi d'identifier simplement comment quels sont les principes d'inventions pouvant résoudre le problème technique rencontré (Moehrle, 2005). Pour illustrer l'utilisation de cette matrice, nous pouvons reprendre l'exemple d'une table que l'on souhaite agrandir pour en augmenter la surface utile, mais sans en augmenter le poids [21]. L'effet positif désiré concerne la surface de la table. L'effet négatif en conflit concerne le poids de cette même table. Ainsi, dans la matrice, l'effet désiré correspond à la ligne 5 : « Surface objet mobile ». L'effet négatif correspond à la colonne 1 : « Masse objet mobile ». Dans la Figure 2.3, nous pouvons voir que les solutions proposées dans la cellule au croisement de cette ligne et de cette colonne sont les principes 2, 17, 29 et 4, soit respectivement le principe d'extraction / séparation, le principe autre dimension, le principe hydraulique ou pneumatique et le principe asymétrie.

20. https://www.180-360.net/wp-content/uploads/2017/07/MATRICE-TRIZ-40-PRINCIPES-FRANCAIS.pdf,https://www.180-360.net/resoudre-contraction-avec-matrice-triz/
21. https://innover-malin.com/triz-40-matrice-interactive/

Il appartient au créateur de choisir parmi ces propositions celle convenant le mieux à son problème (on peut aisément supposer que dans le cas présent, le principe d'asymétrie n'est pas le plus approprié).

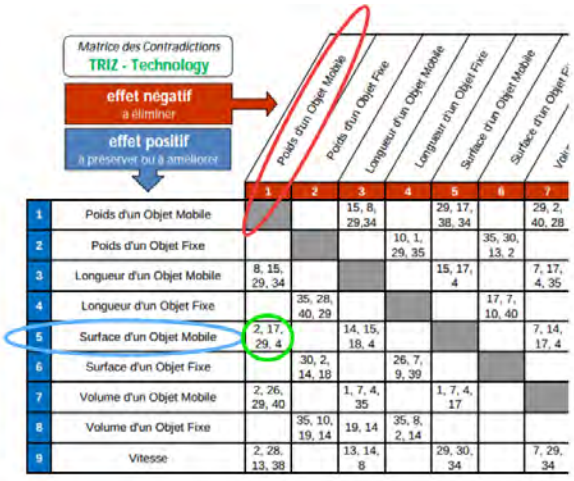


FIGURE 2.3 – Résolution de la contradiction d'agrandissement de la table

2.2.2.2 76 standards inventifs

D'après Livotov (2008), les principes d'inventions présentés ci-dessus sont considérés comme l'un des outils les plus simples de la méthode TRIZ, mais il nécessite de passer par le système des contradictions et s'applique à des problèmes relativement aisés. Lors d'une confrontation avec des problèmes plus complexes, il est recommandé de mettre plutôt en œuvre le système des 76 standards inventifs. Ces standards permettent la transformation d'un système (pour son amélioration) mais aussi la synthèse de nouveaux systèmes. Ils sont, pour cela, basés sur les lois d'évolutions présentées en 2.2.1, mais en représentent l'application pratique. Ces standards sont classés selon cinq catégories :

1. Synthèse et transformation des systèmes techniques ;
2. Amélioration de l'efficacité des systèmes techniques ;
3. Étapes d'évolution des systèmes techniques ;
4. Mesure et détection dans les systèmes techniques ;
5. Aide dans l'application des standards[22].

Ces standards sont étroitement liés au modèle de formalisation des problèmes *S-fields* que nous développons en partie (2.3.1).

22. Nous reprenons ici le nom des classes telles que données par Livotov (2008). En fonction des auteurs, les noms peuvent différer mais le concept reste similaire.

2.2.2.3 Hommes miniatures

Chybowski (2017) décrit la méthode des « hommes miniatures » (en anglais *Smart Little Men (SLM) ou Miniature's Dwarves*) qui consiste à remplacer les différents éléments du système problématique par de petits personnages. Cette méthode s'inspire de la synectique de Gordon (1965) qui envisage la résolution de problèmes selon un mécanisme d'analogie (Michelot, 2015). Celui-ci est défini par Montmain *et al.* (2003) comme un « principe d'empathie [selon lequel] un acteur doit pouvoir s'imaginer comme étant lui-même le système à l'origine du problème ou le problème lui-même. C'est, à la base, le changement de point de vue provoqué par cette approche qui peut effectivement permettre d'exprimer ce problème différemment voire de le solutionner ». Ce principe se heurte à des limites, notamment dans l'incapacité d'un humain à faire preuve de suffisamment d'objectivité pour envisager tous les défauts ou limites d'un système. Altshuller choisi de remanier ce principe et propose d'envisager une armée de petites personnes dans la résolution du problème. Ils peuvent aussi être équipés d'outils sommaires (marteau, lampe...). Le principe derrière cette idée est de mettre de côté les composants du système afin d'amener un nouveau point de vue à la réflexion en se plaçant à la place d'un des hommes miniatures. Ce nouveau point de vue vise à remettre en question la notion de forme et à rappeler que face à un objet qui pose problème « la forme n'est pas rigide » et qu'il est possible de « modifier un objet quand nous en ressentons le besoin »[23]. Il s'agit alors de trouver quel mode de travail doit être mis en place par les hommes miniatures afin de résoudre le problème (Boldrini, 2005) comme démontré dans l'illustration 2.4[24]. Dans celle-ci, le problème rencontré est une fuite d'un fluide. Pour résoudre le problème, ils décident de stopper l'écoulement de ce fluide. Pour ce faire, ils forment donc une barrière qui bouche le trou duquel l'écoulement a lieu. En conséquence, pour résoudre un problème de fuite d'un fluide, l'inventeur peut envisager un dispositif sous forme de barrière qui va bloquer l'écoulement, barrière représentée dans l'illustration par les hommes miniatures.

Cette méthode est décrite comme un moyen de lutter contre l'inertie psychologique de l'inventeur. La notion d'« inertie psychologique » apparaît régulièrement dans les différents travaux d'Altshuller. Nous la retrouvons par la suite notamment dans les travaux sur TRIZ ayant lieu en France, où elle prend le rang d'obstacle à la créativité et de barrière à briser. Cavallucci (1999) décrit cette inertie comme l'état « où les solutions considérées résident principalement dans votre propre expérience et ignorent les technologies alternatives afin de développer de nouveaux concepts ». TRIZ devient ainsi le moyen de « briser cette inertie psychologique en structurant la réflexion de l'ingénieur, afin de l'entraîner au-delà de ses

23. https://www.youtube.com/watch?v=1uFWSk2W3Go&t=1225s
24. https ://sketchplanations.com/

compétences propres, et, surtout, bien au-delà de son centre d'intérêt du moment ».

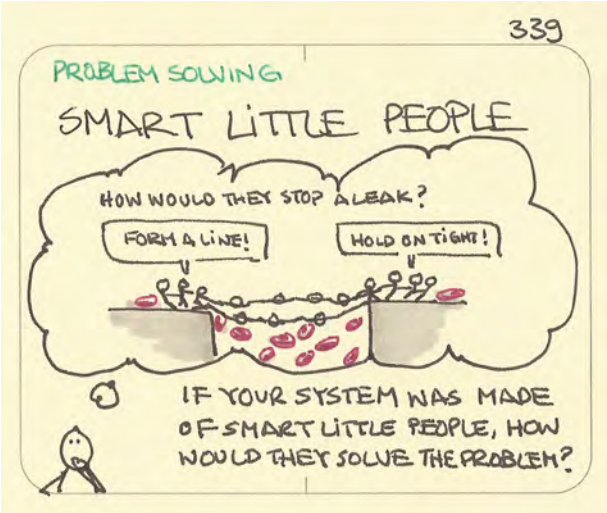


FIGURE 2.4 – Illustration de la méthode des « Hommes Miniatures » par Jono Hey, *Sketchplanations*

2.2.2.4 ARIZ : *Algorithm for Inventive Problem Solving*

ARIZ[25] est le dernier outil que nous présenterons ici. C'est un algorithme développé par Altshuller au fur et à mesure des années qui a fait l'objet de plusieurs versions entre 1961 et 1985. ARIZ-85 est la dernière version développée par Altshuller lui-même et elle comprend neuf sections avec chacune quarante étapes, à réaliser pas à pas (Moehrle, 2005) afin de guider l'inventeur dans l'emploi des différents outils. Par exemple, elle place dans l'ordre suivant : la contradiction technique, suivie de la formulation du résultat final idéal, suivie de la contradiction physique. En effet, d'abord nous établissons quels éléments du système rentrent en contradiction dans le cas d'un changement, puis ce à quoi devrait ressembler le système idéal, et ensuite, au sein de ce résultat idéal, les deux éléments qui sont incohérents.

Cet algorithme est considéré comme adapté aux problèmes les plus complexes (Ilevbare *et al.*, 2013), là où les 40 principes d'inventions ou les 76 standards n'ont pas suffi (Livotov, 2008). Zouaoua *et al.* (2013) déclare que, selon Altshuller, ARIZ est pertinent pour uniquement 5% des inventions, pour les « problèmes inventifs non-standards ».

> « Il est l'outil le mieux adapté parce qu'il assure une exploration plus systématique, réalise des analyses très fouillées à différentes étapes, définit au mieux le Résultat Ultime Idéal dans ses premières étapes, reformule la contradiction technique pour la traduire en contradiction physique privilégiée par ARIZ, mobilise

25. version française : https://altshuller.ru/world/fra/ariz85v.asp

> les ressources (champs et substances) pouvant servir à sa résolution (Ragab-Zouaoua, 2012). »

Dans cette section, nous avons présenté certains des outils que propose la méthode TRIZ dans le cadre de la résolution de problèmes techniques. Ces outils sont les plus courants et s'appuient au moins partiellement sur les principes généraux de TRIZ développés en partie 2.2.1. Dans cette première partie de chapitre, nous avons pu observer la grande diversité des concepts, méthodes et outils développés au sein de cette théorie. Cette multiplicité et cette diversité conduisent néanmoins certains auteurs à qualifier cette méthode comme étant « rigid, difficult to comprehend, had a limited scope of problem-solving, and demanded expert interventions » d'après Ghane *et al.* (2024) citant Verhaegen *et al.* (2011); Wang et Zhao (2021); Wu *et al.* (2021). Ainsi, des travaux ont tenté de simplifier son utilisation, notamment en automatisant certains procédés avec l'aide du TAL, comme nous le verrons dans la section suivante.

2.3 Implication de TRIZ dans nos travaux

Nous avons présenté les aspects fondateurs de la théorie et certains outils qui en ont découlés pour sa mise en application. Maintenant, nous souhaitons présenter dans un premier temps la modélisation proposée par TRIZ que nous visons dans nos travaux. Dans un second temps, nous présentons quelques travaux qui mettent en lien le TAL et TRIZ.

2.3.1 La modélisation en *vépole*

Parmi les outils proposés par TRIZ, la modélisation sous forme de *vépole* est celui qui a trouvé une application directe dans notre travail. Plus précisément, cette modélisation a guidé l'orientation de cette thèse puisqu'elle est le formalisme que nous avons choisi pour représenter les problèmes décrits dans les FA.

Le *vépole*, aussi appelé *Substance - Champ* (*Substance - Field* ou *Su-Field* en anglais) ou encore « Function analysis », est un terme issu de la contraction des termes « Substance » et « Champ » en russe, respectivement « Vechestvo » et « Pole » (Cavallucci, 1999). L'analyse Substance-Champ est composée de trois étapes (Bultey *et al.*, 2007) :

1. la modélisation du problème sous forme de *vépole* ;
2. la résolution du *vépole* par l'utilisation des 76 standards inventifs ;

3. l'interprétation vers des objets physiques.

Dans notre cas, nous nous focaliserons sur la première étape, la modélisation. En effet, c'est lors de cette étape que la relation avec le contenu textuel que nous traitons se fait. Nous cherchons à modéliser les anomalies décrites dans le texte sous forme de *vépole*, de relation fonctionnelle. Cette modélisation est le point d'entrée vers TRIZ que nous visons. Elle est la dernière étape de notre traitement, avant de pouvoir être exploitée avec la méthode TRIZ, ses outils et ses experts.

Le *vépole* peut être défini comme un « système technique minimal » (Cavallucci, 1999). Selon Rousselot *et al.* (2008) le *vépole* repose sur le principe que « les problèmes d'un système peuvent être ramenés à un nombre réduit de parties élémentaires en interaction ». Chacun des éléments du système peut être remplacé par une forme équivalente élémentaire qui interagit avec les autres éléments du système. Les parties du système sont les « substances » et les interactions sont les « champs ». Un *vépole* minimal, pour être considéré comme complet, doit être composé à *minima* de deux substances, un *produit* et un *outil*, et d'un champ. L'*outil* agit sur le *produit* par l'intermédiaire du *champ*. Ce champ (ou action) possède une acception large du terme. Il peut être une force, une énergie physique... (Russo et Duci, 2015).

Lorsqu'on représente un problème technique par le biais de cette modélisation, on se retrouve avec un *vépole problème* (cf. Figure 2.5). En effet, il existe aussi des *vépoles solution*. L'interaction entre les substances peut être qualifiée d'« utile » (*vépole solution*), de « néfaste » ou d'« insuffisante ». Le passage d'un *vépole problème* à un *vépole solution* se produit grâce à l'application d'un des 76 standards inventifs décrits en 2.2.2.2.

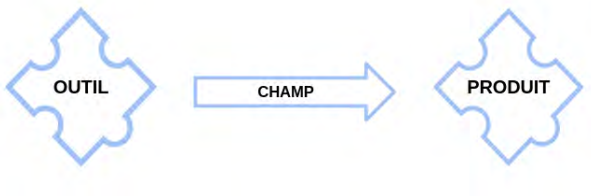


FIGURE 2.5 – Schéma d'un *vépole* type

Prenons pour exemple un pneu de voiture crevé. Le problème ici peut être vu sous plusieurs angles, le pneu n'est pas assez solide, des déchets sur la route ont crevé le pneu, mon manque d'attention m'a entraîné vers un trottoir... Mais ces informations sont déjà de l'ordre du diagnostic et relève plutôt de l'analyse des causes, comme nous avons pu le voir dans la section 1.2.3 du chapitre 1. La description du dysfonctionnement présent se doit de se concentrer sur la situation problématique, c'est-à-dire que de l'air s'échappe du pneu, ou bien que le pneu ne retient pas l'air. Cette relation est illustrée avec le *vépole* en bas de la Figure 2.6.

Il correspond à la modélisation de ce problème. La relation entre les deux substances dans un *vépole* représente toujours une action. Ici, c'est l'action « retenir », mais cette action peut être qualifiée d'insuffisante, elle ne remplit pas entièrement sa fonction. Cette modélisation peut ensuite être traduite sous forme de relation impliquant des substances plus élémentaires comme démontré par le *vépole* en haut de cette même Figure. Ce procédé reprend le principe d'Altshuller selon lequel la résolution d'un problème peut se faire par une abstraction de ce même problème dont nous avons discuté 2.2.1 et montré dans la Figure 2.1.

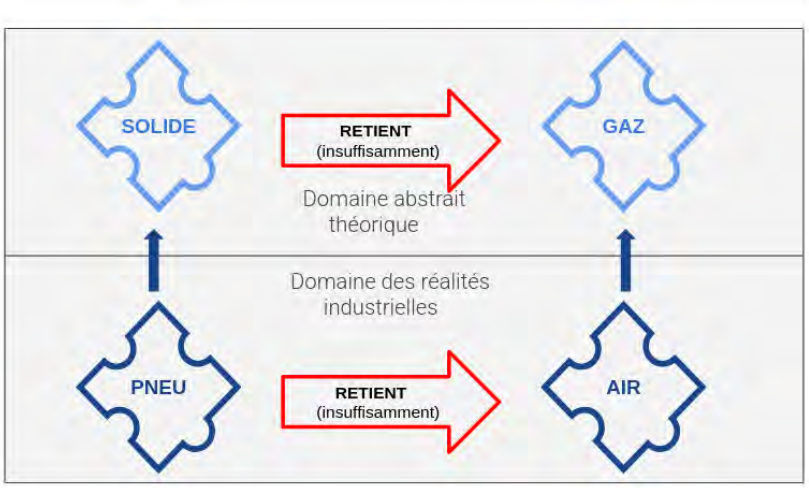


FIGURE 2.6 – *Vépole* du pneu crevé

La description du *vépole* que nous donnons dans cette section correspond à la forme minimale de cette modélisation. Celle-ci fait l'objet de déclinaisons afin de préciser la relation, que ce soit par l'ajout de propriétés aux substances, la présence de *vépoles* successifs, la spécification de relations... Dans le cadre de notre travail, nous faisons uniquement usage de la forme minimale de cette modélisation, telle que nous l'avons décrite ci-présent. Dans les textes que nous traitons, obtenir la forme minimale du *vépole* signifie pouvoir récupérer ces deux éléments et la relation qui les unit, et faire ainsi abstraction d'un ensemble d'informations textuelles qui ne sont pas nécessaires dans cette relation.

2.3.2 La base I4KN

Comme nous avons pu le mentionner précédemment, la modélisation en *vépole* sert de point d'accès vers une résolution de problème grâce aux 76 standards inventifs. La recherche d'une solution parmi ces 76 standards, et à partir des *vépoles* est par ailleurs facilitée par l'algorithme de Smirnov[26]. Celui-ci permet, en partant du *vépole*, de guider l'inventeur vers

26. https://www.triz-conseil.com/modeles-triz/

une sélection de standards à appliquer.

Une autre approche a vu le jour afin de résoudre le problème modélisé par le *vépole* à travers l'utilisation de liste d'effets physiques. L'une des implémentations de liste « d'effets physiques », est la base de connaissances I4KN[27] développée par l'entreprise *MeetSYS*[28], partenaire industriel de cette thèse. Cette base se présente comme un *wiki*, une plateforme collaborative pour capitaliser les connaissances scientifiques et techniques. La plateforme propose une navigation à partir de « fonctions élémentaires » telles qu'elles sont conçues dans le domaine du « *design engineering* » (Hirtz *et al.*, 2002). Dans la Figure 2.7, montrant la page d'accueil de la plateforme, on retrouve donc les fonctions suivantes : produire, éliminer, déplacer, retenir, séparer, protéger, générer, absorber, redistribuer, accumuler, convertir, modifier, détecter et mesurer. À l'exception des fonctions détecter et mesurer qui sont classifiées à part, les fonctions sont catégorisées selon le type d'élément auxquels elles s'appliquent, soit une substance, un champ ou une propriété. Ces derniers reprennent les éléments du *vépole* tels que nous avons pu le voir précédemment.

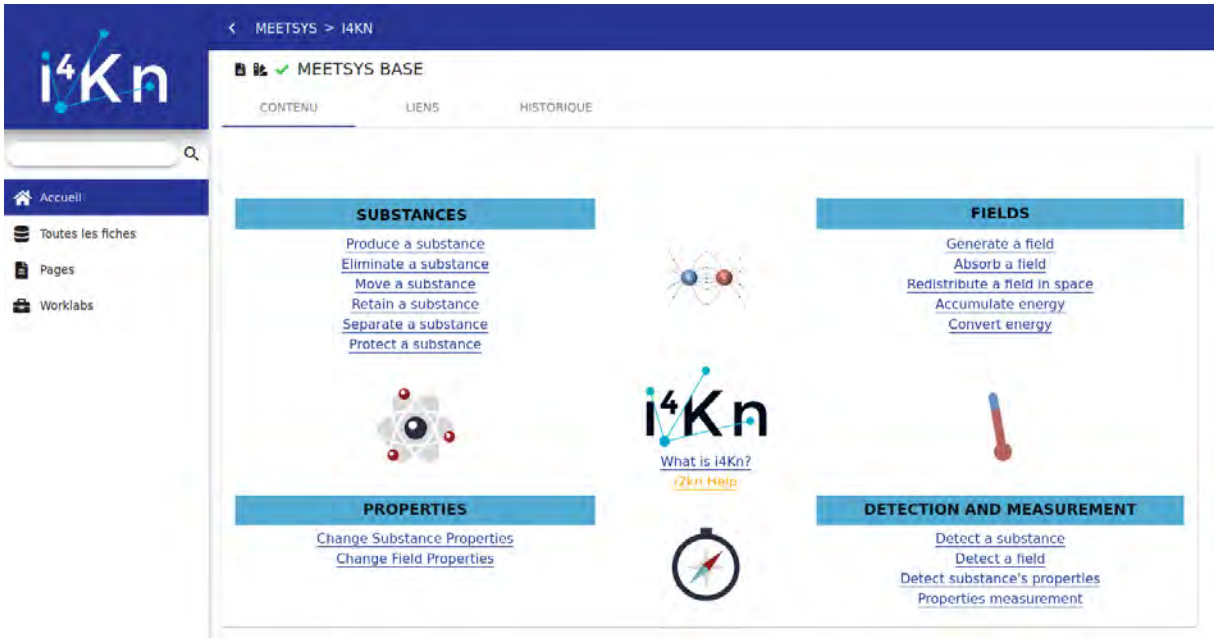


FIGURE 2.7 – Page d'accueil d'I4kn

La Figure 2.8 présente la page correspondant à une de ces fonctions élémentaires, ici `Retain a substance`. Comme nous pouvons le voir, les substances sont catégorisées en « solide », « liquide », « gaz » et « autres ». À nouveau, nous retrouvons des éléments familiers, soit les différentes substances du *vépole* une fois ramenés à un niveau d'abstraction plus

27. https://i4kn.meetsys.i2kn.com/
28. https://www.meetsys-i2kn.com/

élevé. La fonction et la substance, ici `Retain a solid` par exemple, reforment les éléments champ et produit d'un *vépole*.

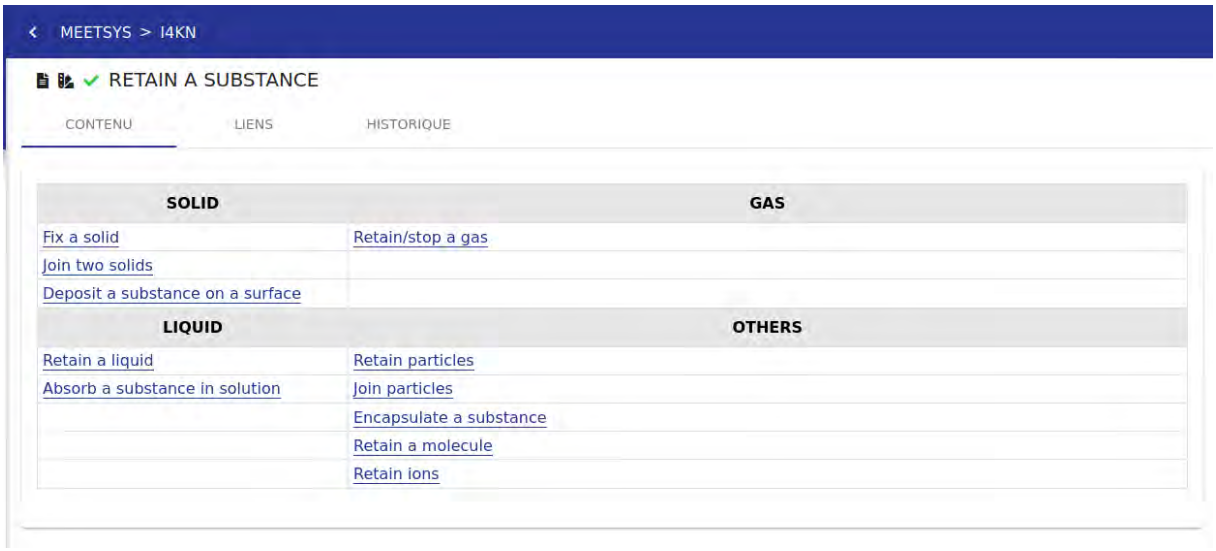


FIGURE 2.8 – Page d'une fonction dans I4KN

En choisissant l'une des propositions d'action dans la page d'une fonction (i.e. Figure 2.8), on accède à un ensemble de solutions proposées pour effectuer l'action en question. La Figure 2.9 présente des solutions pour l'action `Retain_stop a gas` applicables pour retenir une substance gazeuse. Ces mêmes solutions font l'objet (pour la majorité) d'une page distincte sur la plateforme qui contient une définition et éventuellement des exemples et des applications. La plateforme possède aussi une visualisation de type graphe qui permet de naviguer à travers les liens entre fonctions, les concepts et les solutions.

Dans le cadre de notre travail, notre objectif principal est de parvenir à extraire les informations énoncées dans les FA, et de les modéliser sous forme de *vépole*. En effet, cette formalisation sert à représenter un problème technique et parallèlement, les REX de notre corpus rapportent des anomalies techniques / physiques. Ainsi, ce mode de représentation nous a paru particulièrement pertinent comme point d'entrée vers la méthode TRIZ et la perspective de résolution de problème. Il précise quels éléments il est nécessaire de récupérer dans les énoncés afin de représenter le problème en question. Aussi, la base I4KN permet, par le biais du *vépole* d'apporter des propositions de solutions au problème modélisé.

2.3.3 TAL et TRIZ

Plusieurs travaux mêlant TRIZ et le TAL ont déjà été menés en se focalisant sur divers aspects de cette méthode. En revanche, à notre connaissance, les corpus d'études sont es-

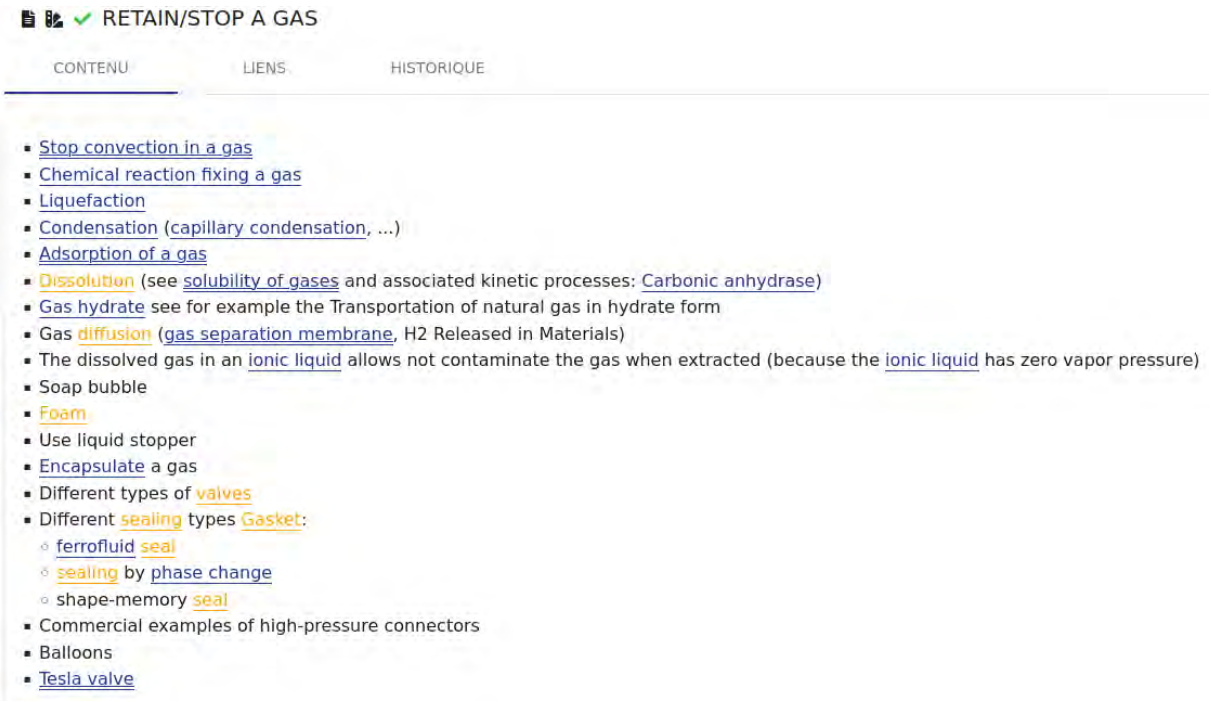


FIGURE 2.9 – Propositions de solutions dans I4KN

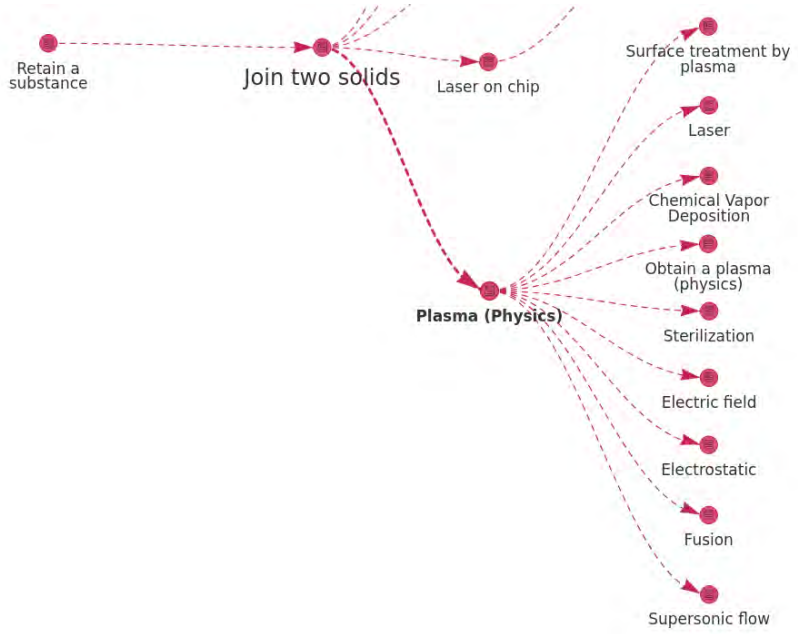


FIGURE 2.10 – Visualisation graphique dans I4KN

sentiellement des brevets (nous rappelons ici que c'est principalement à partir de l'étude de brevets qu'Altshuller a développé TRIZ). Ghane *et al.* (2024) ont effectué une *review* de multiples travaux (57 articles scientifiques sélectionnés) concernant *Semantic TRIZ*, soit l'intersection entre TRIZ et l'IA (se focalisant sur la fouille de texte et le TAL). Selon eux,

« l'application du TAL et du machine learning, dans le contexte de TRIZ, est d'automatiser le processus de compréhension du langage ayant trait aux composants des systèmes d'ingénierie dans des documents textuels pour la résolution de problème et l'innovation de produits »[29]. Cet objectif fait écho aux motivations industrielles de cette thèse.

Verbitsky (2004) est le premier à proposer l'utilisation de techniques de TAL pour automatiser le traitement de brevets afin d'utiliser TRIZ sur une vaste collection de documents. Il nomme cette approche « Semantic TRIZ », qui sera par la suite surnommé « S-TRIZ ». Cette approche a été implémentée dans l'outil *Goldfire Innovator*, présenté dans Verbitsky (2004), qui permet d'effectuer une recherche dans une base de données de brevets. Pour ce faire, une première étape d'extraction de lexique est effectuée à l'aide d'un modèle de Markov caché. Il s'ensuit un *parsing* syntaxique, puis une analyse syntaxique de type SAO (Sujet-Action-Objet). L'auteur prend l'exemple de la phrase « *Electrolytic dissociation can be successfully used to measure air humidity* » dont l'extraction d'items correspond à :

— sujet : Electrolytic dissociation ;
— action : measure ;
— objet : air humidity.

Les méthodes utilisées afin de mettre en place cette approche ne sont pas détaillées dans l'article. En revanche, elle produit deux applications directes. La première est la possibilité de faire une recherche en langage naturel de type « How can I measure temperature of the substrate ? ». L'outil extrait les items Action (*measure*) et Objet (*temperature*) et effectue une correspondance avec les items extraits des brevets. L'outil permet aussi de formuler une recherche formulée sous forme de contradiction similaire à celles lors de l'utilisation de la matrice de contradiction : « How can I increase area, but decrease volume ? ». Cette recherche sera aussi effectuée dans la base de données d'items extraits du corpus de brevets.

L'étude de Ghane *et al.* montre que la structure SAO est une des techniques parmi les plus récurrentes dans les travaux de *Semantic TRIZ*. Selon Choi *et al.* (2011), citant Cascini *et al.* (2004) et Moehrle *et al.* (2005), cette structure permet de représenter au mieux une fonction avec des concepts clés (plutôt que des mots clés). Une fonction est ici entendue telle que définie par Savransky (2000) : « *The action changing a feature of any object* ». Afin d'extraire les structures SAO à partir d'une phrase, le texte est *parsé* syntaxiquement. Un ensemble de règles prédéfinies est ensuite appliqué dans le but d'extraire les structure autour de prédicats (Chen *et al.*, 2017; Yang *et al.*, 2017; Kim *et al.*, 2020). Une correspondance doit ensuite être effectuée entre les fonctions SAO extraites et les sujets d'intérêts, éliminant ainsi les structures non pertinentes. Cela implique généralement l'utilisation de dictionnaires de verbes et de termes autour de sujets spécifiques. Un score de la fonction SAO extraite est

29. Notre traduction.

calculé en fonction de la correspondance des concepts avec les éléments du dictionnaire. Des poids sont attribués aux différents éléments du SAO afin de mettre en évidence la relation fonctionnelle. Le score moyen du SAO (extrait d'une phrase ou d'un paragraphe) permet de mesurer le rapport entre le texte et le sujet d'intérêt (Cascini *et al.*, 2004). D'autres méthodes de calcul de similarité ont été étudiés notamment avec des méthodes vectorielles (Kim *et al.*, 2020). Toutefois, Ghane *et al.* (2024) en évoquent les limites, dont notamment la potentielle perte d'information ou la mauvaise interprétation, dues à l'unique extraction des éléments de la structure SAO, et non de tous les éléments textuels de la phrase. Par ailleurs, l'extraction des structures SAO est très coûteuse en temps puisqu'elle nécessite la construction d'une collection de règles.

L'analyse de Ghane *et al.* (2024) nous indique aussi que 68% des études portent sur des méthodes sémantiques de TAL, notamment des techniques de similarité sémantique comme le *Topic Modeling*. Berdyugina et Cavallucci (2021) font usage de cette technique dans le cadre de l'extraction de contradictions (au sens *trizien* du terme [30]). Parmi les travaux mêlant TAL et TRIZ, nous pouvons aussi citer Ni *et al.* (2021, 2022), qui ont développé un outil de détection de brevets similaires : SAM-IDM. Ce rapprochement entre brevets est basé sur un calcul de similarité de phrases effectué par le biais du modèle MaLSTM (Mueller et Thyagarajan, 2016), qui implique l'utilisation de deux structures LSTMs (*Long Short Term Memory networks*) pour prédire une similarité sémantique entre deux phrases. Les LSTMs (Hochreiter et Schmidhuber, 1997) sont des modèles basés sur des réseaux de neurones récurrents et ont la capacité de traiter des phrases, contrairement à des méthodes de *word embeddings* qui se concentrent sur la représentation d'un seul mot (Bengio *et al.*, 1994). L'outil vise l'extraction de brevets issus de domaine différents afin de proposer des solutions au problème cible. Il cherche à répondre à l'une des problématiques majeures de TRIZ, c'est-à-dire la nécessité de développer une expertise dans la méthode avant de pouvoir la mettre en application, ce qui demande donc un investissement en temps conséquent.

Dans cette section, nous avons présenté quelques travaux qui montrent les possibilités d'utilisation du TAL pour la mise en application de TRIZ. Toutefois, ces travaux semblent montrer certaines limites. Ils confirment la pertinence, voire le besoin, de l'utilisation du TAL afin d'accéder à cette méthode qui demande, non seulement l'acquisition d'une expertise pour son utilisation, mais aussi de capitaliser une grande quantité de données textuelles.

30. Pour rappel, il existe dans TRIZ deux types de contradictions : techniques ou physiques. La contradiction technique a trait à l'amélioration d'un système qui entraîne un nouveau problème. La contradiction physique renvoie à une inadéquation entre les valeurs que doit prendre le même paramètre d'un même système pour résoudre un problème.

En effet, le grain d'analyse textuelle reste large, avec des approches par similarité textuelle au niveau de la phrase ou de l'extraction de concept. Ainsi, notre approche par le biais des *frames* se fait dans l'idée de pousser le traitement textuel vers un grain plus fin en s'ancrant dans un cadre théorique, que nous présentons au chapitre 3, qui nous place plus proche du contenu sémantique. Comme mentionné plus haut, cette approche est aussi permise par la nature de notre corpus de REX qui fait preuve d'une redondance dans les problèmes rencontrés, ce que nous explorons au chapitre 4, et a permis de faire émerger des archétypes de problèmes (chapitre 5). Ces archétypes permettent de faire le lien, à la fois entre les *frames* que nous identifions, mais aussi, comme nous le présentons aux chapitres 8 et 9, vers les *vépoles* types correspondant à ces mêmes archétypes. Nous avons aussi pu voir que les travaux qui mêlent le TAL et TRIZ se concentrent sur certains types de document, à savoir les brevets en majorité et des bases de données d'articles scientifiques et techniques. À notre connaissance, l'utilisation de REX comme corpus cible n'a pas encore été explorée. Cette thèse est donc aussi l'occasion de poser la question de l'adéquation des REX comme corpus avec un traitement « à la TRIZ ».

2.4 Conclusion

Dans ce chapitre, nous avons en premier lieu effectué un panorama des principes généraux de TRIZ et de ses outils. Cela nous a permis de rendre compte d'une part de l'ancrage de cette théorie dans un principe d'abstraction dans sa représentation des systèmes et des problèmes techniques. D'autre part, nous avons constaté la multiplicité des méthodes et des outils qui entrent en jeu. Ainsi, cette abstraction procure une facilité pour un utilisateur à représenter les éléments dans une perspective de résolution de problème ou d'invention. À l'inverse, la prolifération d'outils et de méthodes rend sa mise en œuvre ardue.

Ce dernier constat, à la manière de Bultey *et al.* (2007), incite à envisager l'utilisation de l'IA, et dans notre cas du TAL, afin d'en faciliter l'usage. Dans la troisième section de ce chapitre, nous présentons donc de quelle façon nous interagissons avec TRIZ dans nos travaux. Cette interaction passe notamment par la modélisation sous forme de *vépole*, qui est notre point d'entrée vers TRIZ et vers la recherche de solutions aux problèmes rapportés dans les REX. Comme nous avons pu le voir, cette recherche de solution peut être effectuée par le biais de listes d'effets physiques, qui prend ici la forme de la base de connaissance I4KN.

Si le *vépole* est notre point d'entrée vers TRIZ, il est aussi l'objectif que nous visons, la formalisation du contenu de notre texte. Nous avons pu voir que TRIZ et le TAL ont été l'objet

de nombreux travaux. L'approche que nous proposons dans cette thèse vise, de manière similaire à certaines de ces études, à l'extraction d'informations formalisées sous forme de relation fonctionnelle (le *vépole*) comme nous le verrons dans les chapitres 8 et 9. Les résultats de l'analyse de Ghane *et al.* (2024) révèlent d'ailleurs que l'outil de TRIZ le plus utilisé dans le cadre des travaux en TAL est la modélisation « Substance-Champs » (le *vépole*), vue dans la partie précédente, à hauteur de 19% des articles étudiés. En revanche, notre proposition de typologie (chapitre 5) permet d'une part de cibler les archétypes de problèmes rencontrés, et ainsi de focaliser le travail sur les phrases (ou morceaux de phrases) porteuses de l'information. Ces archétypes sont néanmoins le résultat d'une certaine redondance dans les problèmes exprimés dans ces REX que nous présentons au chapitre 4, ce qui n'est peut-être pas le cas dans un corpus de brevets. D'autre part, le lien entre la typologie et le texte, passant par des prédicats dans une méthode basée sur les *frames* de *FrameNet* (que nous développons aux chapitres 3 et 5) permet de passer outre l'utilisation d'un *parseur* syntaxique dont les performances sont mitigés sur notre corpus spécialisé et bruité.

Chapitre 3

Représenter un énoncé : la sémantique des cadres

Sommaire

3.1 Introduction : pourquoi la sémantique des cadres ?

Dans le chapitre précédent, nous avons présenté TRIZ et notre objectif de modéliser les éléments de l'anomalie évoqués dans le texte par le biais de cette méthode. Pour ce faire, il est donc nécessaire d'obtenir une description fine des éléments présents et de leurs interactions d'un point de vue sémantique. C'est ainsi que la « sémantique des cadres » de Fillmore (1976, 1977b, 1982, 1985) entre en jeu. Nous nous appuyons sur cette théorie, et la ressource *FrameNet* (Ruppenhofer *et al.*, 2016) qui en découle, afin de décrire et de représenter les archétypes d'anomalies, en lien avec la typologie d'expression d'un problème technique que

nous avons construite et que nous développons au chapitre 5. Cette typologie se base sur des marqueurs lexicaux qui prennent le rôle de déclencheurs de la situation prototypique portée par le cadre.

Dans ce chapitre, nous prenons donc le temps de présenter la notion de rôle sémantique et son interaction avec la théorie de la « sémantique des cadres » qui est le cadre théorique dans lequel s'inscrivent nos travaux. Nous présentons par la suite la ressource *FrameNet* et comment elle met en place les notions principales de la « sémantique des cadres ». Enfin, nous explorons quelques travaux qui en résultent par trois aspects : l'adaptation de *FrameNet* vers d'autres langues, son utilisation pour des corpus spécialisés et comment le TAL permet de mettre en œuvre une analyse linguistique *à la FrameNet*.

3.2 Les rôles sémantiques et les *frames*

La théorie de la sémantique des cadres (*« frame semantics »*), à laquelle on s'intéresse ici-présent provient de travaux de Fillmore (1976, 1977b, 1982, 1985). Fillmore se concentre sur l'étude du sens des mots à travers le lien entre langue et connaissance du monde, ou la nécessité de faire appel aux connaissances du monde afin d'obtenir une compréhension du sens d'un mot. Ces connaissances du monde sont structurées en « cadres », soit une situation prototypique évoquée par le mot en question. Cela l'entraîne à redéfinir les rôles sémantiques traditionnels.

Djemaa (2017) définit les premières approches autour de la notion de « rôles sémantiques » (Fillmore, 1967; Gruber, 1965) comme visant à « décrire le sens d'un prédicat par une liste de rôles » et avec des étiquettes assignées à chacun des arguments d'un prédicat. Ces étiquettes prédéfinies sont caractéristiques de l'événement porté par le prédicat en question. La catégorisation de la relation sémantique entre le prédicat et ses participants permet de faciliter la construction de la représentation sémantique du texte « et savoir repérer dans un texte *qui* a fait *quoi* à *qui* » (*Ibid*). Parmi les rôles les plus répandus, nous pouvons retrouver ceux de la liste présentée dans le tableau 3.1 empruntée à Sanacore (2023). Ces différentes étiquettes ne font toutefois pas consensus (Fradin et Winterstein (2012); Levin et Rappaport Hovav (2005)...).

Selon Stalmaszczyk (1996), les postulats communs dans l'approche de Fillmore et Gruber sont les suivants :

1. les rôles sont des étiquettes atomiques ;
2. les étiquettes proviennent d'une liste fixe ;
3. les étiquettes suivent une hiérarchie ;

Rôle sémantique	Définition	Exemple
agent	Initiateur de l'action, capable de volition	Marc a cuisiné une soupe.
bénéficiaire	À qui bénéficie l'action qui est accomplie	Marie a acheté un cadeau pour sa sœur.
cause	Initiateur d'une action, non capable de volition	Le feu a brûlé la cabane
destination	Point d'arrivée d'un déplacement	Nous avons pris le train pour Paris.
experienceur	Perçoit l'action mais ne la contrôle pas	Jean a une migraine.
instrument	Intermediaire utilisé pour réaliser une action	Max a cassé les pierres avec un marteau.
patient	Affecté par l'action, subit un changement d'état	Julie a cassé la vitre
résultat	Produit final d'une action	Claire a fabriqué un pot.
source	Point de départ d'un déplacement	Ce train arrive de Rome.
stimulus	Produit un effet (psychologique) sur un expérienceur	Le bruit a effrayé les passagers.
thème	Entité déplacée ou localisée	Alex a jeté le frisbee.

TABLE 3.1 – Rôles sémantiques les plus utilisés ainsi que leur définition d'après Sanacore (2023)

4. les rôles sont liés à une position syntaxique ;
5. un argument possède exactement un rôle sémantique.

D'après (Djemaa, 2017), ces approches dites « traditionnelles » comportent cependant des limites et font état de désaccords parmi la communauté scientifique, notamment concernant la définition des rôles et de leur granularité, et l'établissement d'un ensemble fini de ces rôles (Dowty, 1991; Cruse, 1973; Levin et Rappaport Hovav, 2005).

Diverses propositions ont été faites afin de circonvenir à ces limites (Dowty (1991), par exemple les rôles prototypiques ou les classes de Levin (Levin, 1993)). Dans notre cas, nous avons choisi de nous intéresser à celle de Fillmore. La sémantique des cadres de Fillmore propose une approche plus fine, avec une granularité plus précise des rôles sémantiques. Cette théorie est cependant basée avant tout autour de la notion de compréhension des textes, plutôt qu'une redéfinition des rôles sémantiques. Pour cela, Fillmore (1976) fait appel à plusieurs notions, parfois issues de domaines différents :

> « There are several notions that make up the background to what I have in mind, and these are : the concept of context, the concept of prototype or paradigm case, the notion “frame” or ”schema” as it is used in recent work in psychology and artificial intelligence, and the notion that sometimes goes by the name of “semantic memory“. »

La première notion est celle de contexte, envisagée sous différents aspects (situation d'énonciation, contexte de la phrase, contexte de l'expérience). Fillmore la considère pertinente de deux façons : le sens des mots dépend du contexte d'expérience, et l'interprétation d'un énoncé peut dépendre de la situation d'énonciation et de nos souvenirs des contextes d'expériences antérieures avec l'énoncé et ses constituants.

La seconde notion invoquée est celle de prototype ou *paradigm case* pour lequel Fillmore renvoie à Wittgenstein (1953) Erdmann (1922) et Rosch (1973). Elle pose comme postulat que, pour percevoir ou atteindre un concept, il est nécessaire d'avoir en mémoire un répertoire de prototypes. La perception ou la conception consiste alors en la reconnaissance de la manière dont un objet peut être vu comme une instance d'un de ces prototypes.

La notion suivante est celle de *frame*, *schema* ou *scenario*, que nous pouvons retrouver dans les domaines de la psychologie ou de l'intelligence artificielle (telle qu'elle était mise en œuvre à cette période, c'est-à-dire à l'aide de règles symboliques). L'idée principale est la suivante : nous possédons un inventaire de schémas en mémoire qui permettent de structurer, classifier et interpréter les expériences. Ces schémas sont accessibles et modifiables de plusieurs manières. Cette notion de *frame* n'est pas dépendante du langage, mais elle implique que des mots et formules du discours sont associés en mémoire à des *frames* particuliers. Ainsi, l'exposition à une forme linguistique dans un certain contexte active un *frame* particulier, ce qui permet l'accès au matériau linguistique associé à ce même *frame*. Pour illustrer cette notion, Fillmore (1976) prend l'exemple, réemployé à nombreuses reprises, de la transaction commerciale commune à tous les langages dont la société possède une économie monétaire. Le scénario pour ce cadre implique les rôles suivants : l'acheteur, le vendeur, les biens et l'argent. Tous les mots relatifs à ce cadre (« acheter », « vendre », « coût »...) peuvent accéder au cadre entier de la transaction commerciale dans lequel *l'acheteur* donne *l'argent* et récupère *les biens* et *le vendeur* donne *les biens* et récupère *l'argent*. Ces mots « activent » le scénario de la transaction commerciale dans l'esprit de la personne qui les rencontre.

Enfin, la dernière notion est celle de « mémoire sémantique ». Elle renvoie au modèle du monde de l'interlocuteur, soit le réseau de relations entre les éléments de connaissance et comment ces derniers sont intégrés dans un modèle (plus ou moins cohérent) d'une certaine image du monde. Parallèlement, par l'acte de communication, une personne affecte le contenu du modèle du monde de son interlocuteur. Cette notion renvoie à celle de « mémoire

sémantique » de (Quillian, 1966) et « l'image » telle qu'elle est entendue dans les travaux de Boulding (1961).

Dans l'exemple de la transaction commerciale mentionnée plus tôt, les différents rôles mentionnés (l'acheteur, le vendeur, les biens et l'argent) sont les participants du cadre en question (tous les participants n'étant pas nécessairement présent dans un énoncé). Ces éléments sont les rôles sémantiques en jeu propres au cadre de la transaction commerciale. Ils seront activés par les mots faisant référence à ces rôles (« somme d'argent », « coût », « client »...)[31]. Ces rôles sont donc bien à un niveau de granularité plus fin que ceux que nous avons pu voir Table 3.1. Dans (Fillmore, 1977a), l'auteur précise aussi une des fonctions de la sémantique des cadres comme étant de proposer un « pont » entre des descriptions de situations et leurs représentations syntaxiques sous-jacentes. Ainsi, les rôles attribués aux participants du cadre sont sémantico-syntaxiques. Cette propriété de la sémantique des cadres est celle qui représente le point de passage entre les énoncés dans les FA et la modélisation TRIZ. En effet, les rôles permettent ainsi de décrire la situation d'un point de vue sémantique à partir des différents éléments dans l'énoncé.

Ainsi, Fillmore propose une façon d'aborder la compréhension d'un texte en se basant sur la théorie cognitive des cadres. Elle implique deux objets, les cadres (des scénarios en mémoire), et les rôles sémantiques spécifiques et activateurs de leur cadre. Cette théorie a donné lieu à de nombreux travaux, dont la création de la ressource *FrameNet* (Ruppenhofer *et al.*, 2016) que nous explorons dans la section suivante.

3.3 *FrameNet*

De nombreuses ressources ont vu le jour afin d'essayer de structurer le lexique (par l'utilisation des rôles sémantiques ou non). Parmi celles-ci, nous pouvons retrouver la base de données *Wordnet* (Miller, 1995; Fellbaum, 1998) qui est basée sur des ensembles de synonymes (*synsets*). Nous pouvons aussi citer le corpus annoté *PropBank* (*Proposition Bank*) (Kingsbury et Palmer, 2002; Palmer *et al.*, 2005; Gildea et Palmer, 2002). Cette ressource se présente comme une couche supplémentaire au corpus arboré *Penn Treebank II* (Marcus *et al.*, 1993) issu d'articles du *Wall Street Journal*. Le corpus est annoté avec des structures

31. Nous utilisons ici le terme « activé » en rapport avec la conception de Fillmore selon laquelle « Particular words or speech formulas, or particular grammatical choices, are associated in memory with particular frames, in such a way that exposure to the linguistic form in an appropriate context activates in the perceiver's mind the particular frame-activation of the frame, by turn, enhancing access to the other linguistic material that is associated with the same frame »

prédicat-argument, en utilisant des étiquettes génériques pour ces derniers (« Arg0 », « Arg1 », « Arg2 »...), mais un rôle attribué au prédicat. D'autres ressources pourraient être encore citées telle que VerbNet (Kipper *et al.*, 2006), le LVF (les Verbes Français) (Dubois et Dubois-Charlier, 1997), LG (Lexique-Grammaire) (Gross, 1975)... Dans le cadre de nos travaux, c'est de *FrameNet* dont nous avons fait usage afin de décrire au niveau local les éléments qui constituent nos énoncés.

3.3.1 Présentation générale

La ressource *FrameNet*[32] (Baker *et al.*, 1998; Ruppenhofer *et al.*, 2016), développée à L'UC Berkeley est basée sur la théorie de la sémantique des cadres de Fillmore. Elle instancie une grande quantité de cadres et est largement utilisée, notamment dans des objectifs d'annotation de corpus. Elle met à disposition :

1. des cadres (situations prototypiques), leurs participants et leurs propriétés ;
2. une liste de lexèmes associés aux cadres instanciés ;
3. un schéma d'annotation pour chacun des cadres.

Un *frame* possède les éléments suivants : un nom représentatif composé de un à trois mots (comme « Grant_Permission ») avec une définition. Il est connecté par des relations inter-*frames* qui peuvent être d'ordre hiérarchique, comme nous le verrons dans la section 3.3.2.

Les participants et les propriétés mentionnés au point 1 sont usuellement appelés les « éléments du cadre » (*frame elements*). Ils correspondent aux rôles sémantiques à grain fin tels que présentés dans la section 3.2. Le point 2 fait référence aux unités lexicales, aussi appelés « déclencheurs » qui activent le cadre. Nous revenons sur chacun des points 1 à 3 dans les sections qui suivent.

La version 1.7 de *FrameNet* contient 1 221 *frames* pour 13 572 lexèmes et 10 505 *frames elements*. Cette ressource a été développée pour l'anglais. Notons que certains *frame elements* peuvent être récurrents entre différents cadres, notamment lorsqu'il est question des propriétés.

Dans Fillmore et Baker (2012), les auteurs décrivent la méthode utilisée dans la création et définition des *frames* instanciés dans *FrameNet*. Elle est constituée de cinq étapes :

1. la caractérisation des *frames* ;
2. la description et le nommage des éléments du cadre ;

32. http://berkeleyfn.framenetbr.ufjf.br

3. la sélection des lexèmes associés au cadre ;
4. l'annotation de corpus pour les cadres en question ;
5. la génération automatique d'entrées lexicales.

Dans la suite de cette section, nous prenons un *frame* en exemple, `Annoyance`, et présentons sa description dans *FrameNet*. Nous reprenons pas à pas les éléments qui composent et dépendent du *frame* tels que nous les retrouvons représentés dans le schéma de la Figure 3.1. Par rapport à ce modèle, esquissé par Lönneker-Rodman et Baker (2009), nous ajouterons la description d'une grille d'annotation et de sa mise en pratique dans la ressource.

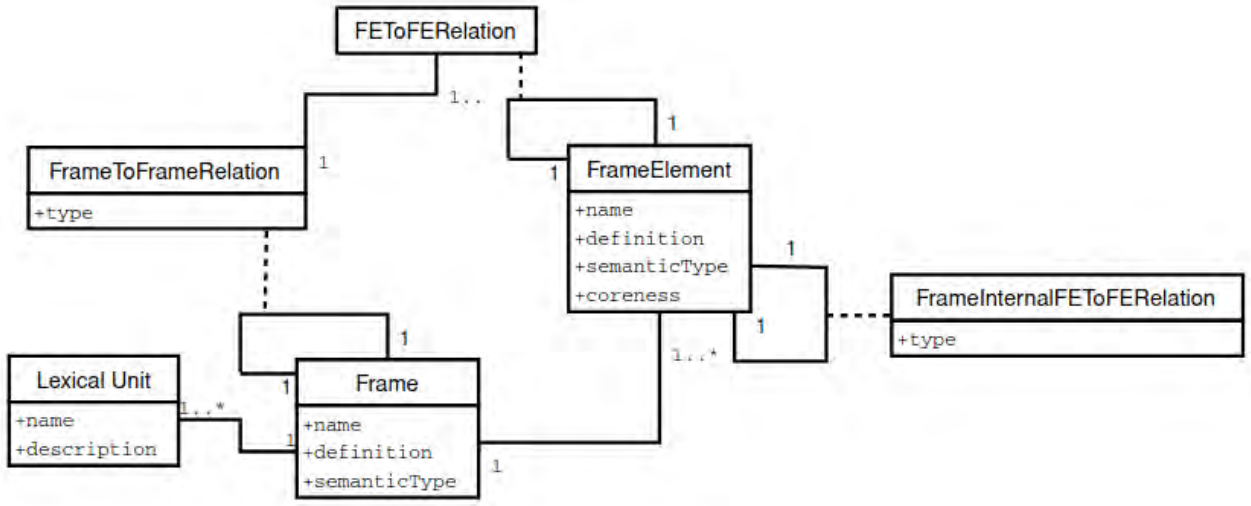


FIGURE 3.1 – Schéma du modèle *FrameNet* issu de Lönneker-Rodman et Baker (2009)

3.3.2 Définition et rôles noyaux

Les Figures 3.2 à 3.5 montrent une page de *FrameNet* pour le *frame* `Annoyance`. Dans la Figure 3.2, nous retrouvons la définition du cadre, qui inclut les éléments du cadre, suivie d'un exemple. Cette définition est donnée sous la forme d'un *scenario*, d'une situation prototypique, plutôt qu'une description formelle. Elle se différencie d'une définition de dictionnaire classique qui, pour notre exemple, s'apparenterait à la suivante issue du *Cambridge Dictionnary*[33] pour l'entrée *annoyance* :

1. the feeling or state of being annoyed ;
2. something that makes you annoyed.

33. https://dictionary.cambridge.org/fr/dictionnaire/anglais/annoyance

À la suite de la définition, nous pouvons retrouver les participants du cadre (les *core elements* ou éléments noyaux) avec, pour chacun, une définition. Ces participants sont des rôles sémantiques propres au cadre. Selon (Ruppenhofer *et al.*, 2016) traduit par (Djemaa, 2017) un rôle noyau « instancie un composant d'un cadre conceptuellement nécessaire, tout en rendant le cadre unique et différent des autres ». Dans l'exemple d'*Annoyance*, les éléments noyaux sont « *Experiencer* », « *Expressor* », « *State* », « *Stimulus* » et « `Topic` » (l'Experienceur, l'Expresseur, l'État, le Stimulus et le Sujet). Nous pouvons noter ici que l'`Experiencer` et le `Stimulus` sont deux rôles qui figurent parmi les rôles sémantiques les plus répandus dans l'approche traditionnelle de la définition des rôles sémantiques (cf. Table 3.1). Ruppenhofer *et al.* (2016) listent un ensemble de caractéristiques des *core elements*. Si une de ces caractéristiques est rencontrée, l'élément est systématiquement noyau, sans que l'inverse soit vrai pour autant. Ces caractéristiques sont les suivantes :

— un rôle qui doit nécessairement être exprimé conjointement à un prédicat (Ruppenhofer *et al.* (2016) prend l'exemple du prédicat *resemble* qui implique nécessairement la présence d'une entité similaire au sujet) ;
— un rôle qui fait l'objet d'une interprétation par défaut lorsqu'il est absent (l'auteur prend ici l'exemple du verbe *arrive* qui implique le rôle de *goal*, c'est-à-dire un point d'arrivée) ;
— un rôle dont la sémantique est imprévisible à partir de sa forme, c'est-à-dire les rôles sans marquage formel ou dont le marquage est idiosyncratique.

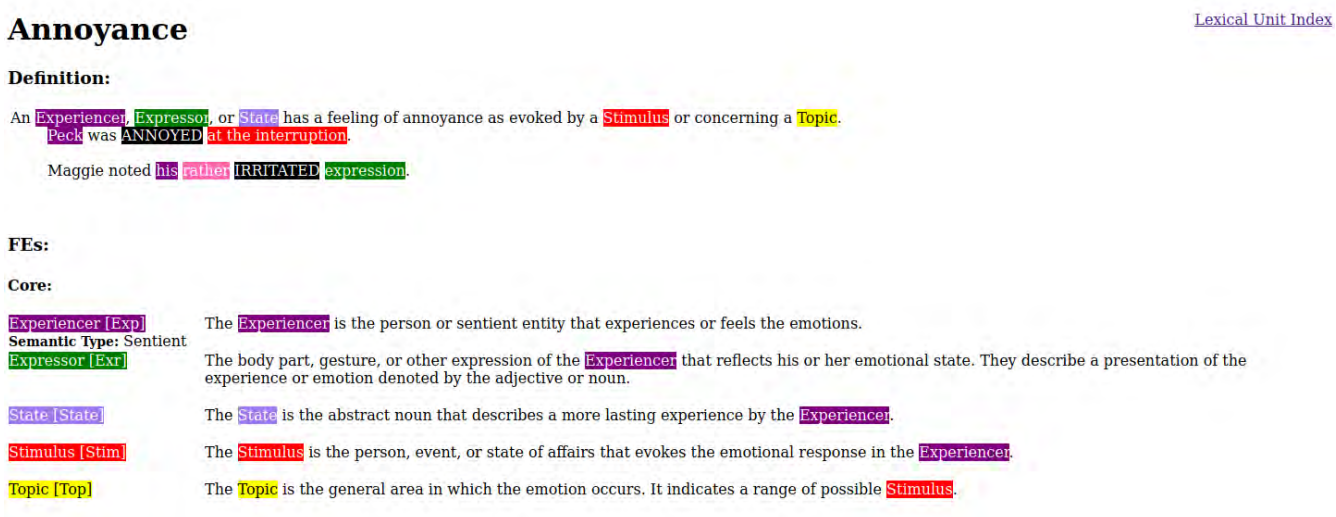


FIGURE 3.2 – Définition et *core elements* du *frame* « *Annoyance* »

Comme nous pouvons le voir dans la Figure 3.2, le rôle « *Experiencer* » possède une caractéristique supplémentaire intitulée « *Semantic Type* » qualifiée de « *Sentient* ». Ce « type sémantique » restreint le constituant du rôle puisqu'il doit obligatoirement être un objet

« *sentient* », soit doué de sensation. Ce type fait référence aux « types ontologiques » de la hiérarchie des types de *FrameNet* décrite par Lönneker-Rodman et Baker (2009) et reprise dans la Figure 3.3. Les auteurs mentionnent deux types sémantiques, les types ontologiques et les *framal* qui servent à caractériser plus précisément les rôles. Les types *framal* font référence à des méta-informations. Il inclut les *frames* non lexicaux et *non-perspectivized*. Pour un *frame* non lexical, nous pouvons prendre l'exemple donnée dans Ruppenhofer *et al.* (2016) qui est le *frame Post-getting*. Celui-ci ne possède pas d'unités lexicales qui lui sont propres. Il connecte les deux *frames* que sont *Getting* et par une relation de « précédence » (en. *Precedes*) et *Getting* par une relation d'« héritage » (en. *Inheritance*).

Seuls les types ontologiques sont organisés selon la hiérarchie de la Figure 3.3. Sanacore (2023) souligne cependant l'absence d'une version complète de cette hiérarchie sur le site de la ressource ou dans les nombreux articles présentant *FrameNet*. Ces types ne sont d'ailleurs ni systématiques, ni majoritaires (Lönneker-Rodman et Baker (2009) décomptent 109 *frames* avec un *framal type* et 29 *frames* dans lesquels on retrouve un ou plusieurs *types ontologiques*).

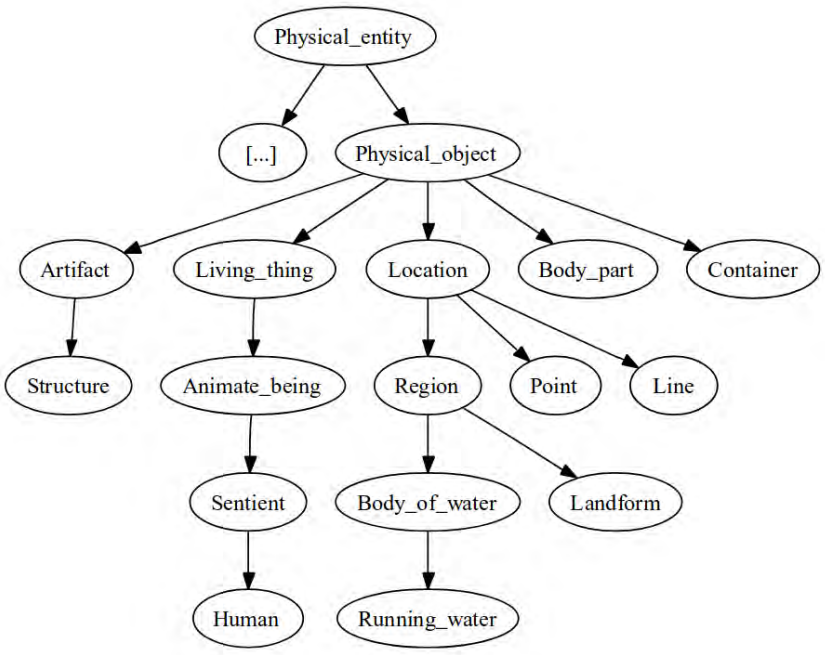


FIGURE 3.3 – Types sémantiques ontologiques de FrameNet (Lönneker-Rodman et Baker, 2009)

3.3.3 Rôles « non-noyaux » et relations frames-frames

Aux participants noyaux du cadre s'ajoutent les *non-core elements* qui se découpent en rôles « périphériques » et « extra-thématiques ». Dans l'exemple de la Figure 3.4, ces deux types de rôles sont regroupés dans la catégorie « *non-core* ». Cette distinction se retrouve néanmoins dans les schémas d'annotation comme dans la Figure 3.6 sur laquelle nous reviendrons plus avant.

Concernant les définitions et conclusions autour de ces rôles non-noyaux, nous reprenons les travaux de Djemaa (2017). L'autrice reprend notamment la distinction que fait Fillmore (2007) entre les *rôles noyaux*, qui, comme nous l'avons vu, font référence au noyau du sens du cadre, et les *rôles périphériques* qui peuvent apparaître dans de nombreux cadres, notamment lorsqu'ils relèvent d'un même type sémantique. Elle prend l'exemple du rôle « *Speed* » récurrent dans tous les *frames* de type « Déplacement ». Les *rôles extra-thématiques* quant à eux, n'appartiennent pas conceptuellement au cadre et n'en dépendent pas. Cependant, elle souligne le manque de clarté dans la distinction entre les rôles *périphériques* et *extra-thématiques*. En effet, la majorité des rôles non-noyaux sont définis comme l'un ou l'autre de ces types de rôle en fonction du cadre dans lesquels ils apparaissent, sans pour autant dénoter d'une interprétation différente dans les cas d'usage. Par ailleurs, l'interprétation des rôles extra-thématiques n'est pas censée dépendre du *frame* comme précédemment mentionné.

L'autre information présente en Figure 3.4 est celle des « *frame-frame relations* ». Celle-ci fait référence aux relations entre les cadres indexés dans *FrameNet* qui permettent de faire état des relations entre les différentes situations cognitives décrites par les cadres (Sanacore, 2023). Baker (2009) propose une classification de ces relations en trois types :

1. relations de généralisation : « l'héritage », « la perspective sur » et « l'usage » ;
2. relations de structure événementielle : « sous-cadre » et « précédence » ;
3. relations aspectuelles : « cause de » et « conséquence de ».

Dans notre exemple du cadre *Annoyance*, seule la relation d'héritage est pertinente. *Annoyance* hérite du cadre *Emotion by stimulus*. Il n'a en revanche pas de cadre héritier. Nous retrouvons bien, dans le cadre *Emotion by stimulus* les rôles périphériques du cadre *Annoyance* que sont « *Degree* », « *Manner* », « *Parameter* » et « *Time* ». Le *frame Annoyance* reste cependant autonome, il ne dépend pas de *Emotion by stimulus* et inversement. La relation spécifie néanmoins le *frame* en spécifiant la situation cognitive avec lequel il est en lien.

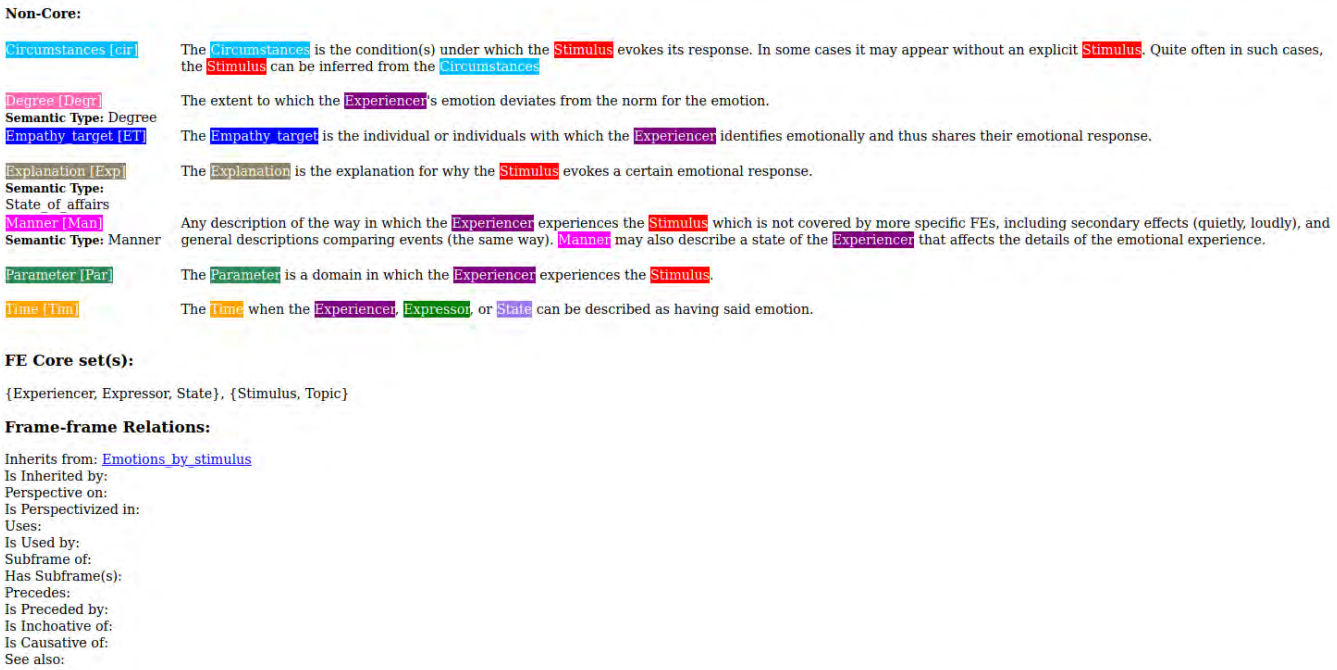


FIGURE 3.4 – *Non-core elements* et liens du *frame Annoyance*

3.3.4 Unités lexicales et schémas d'annotation

La Figure 3.5 correspond à la dernière partie de la page du cadre *Annoyance*. On y retrouve la liste des *Lexical Units* évocateurs du cadre. Ces unités lexicales sont aussi appelées de « déclencheurs » en français. Dans l'exemple d'*Annoyance*, les termes déclencheurs sont peu nombreux et tous des adjectifs. En revanche, si on prend l'exemple du cadre « *Killing*[34] », 75 unités lexicales sont indexées incluant des verbes (*i.e. annihilate, behead, exterminate...*), des noms (*i.e. assassin, murderer, patricide...*), et des adjectifs (*i.e. lethal, deadly, fatal...*).

Ruppenhofer *et al.* (2016) note cependant que les déclencheurs peuvent être polysémiques et ainsi évoquer des cadres différents. Il prend l'exemple du terme *bake* qui est évocateur de trois *frames* : *Apply heat*, *Cooking creation* et *Absorb heat*. Il précise aussi que les expressions multimots peuvent constituer des unités lexicales, tout comme les expressions idiomatiques.

À chaque déclencheur, deux pages peuvent être associées. La première est une « Entrée lexicale » (*Lexical entry report*), et la seconde est une page « Annotation » (*Annotation report*).

Dans la page d'annotation, nous pouvons retrouver les annotations dites *lexicographiques* (df. Figure 3.6). Elle se présente sous la forme d'un tableau présentant les différents rôles sémantiques et leur type (*core, péripheral* et *extra-thematic*) suivi d'un ensemble de phrases issues du corpus BNC (*British National Corpus*) dans lesquelles le déclencheur apparaît. Pour chaque phrase, les rôles thématiques se réalisant sont annotés.

34. https://framenet.icsi.berkeley.edu/frameIndex

Lexical Units:

annoyed.a, frustrated.a, irritated.a, sick and tired.a, tired (of).a

Created by KmG on 04/09/2008 03:10:44 PDT Wed

Lexical Unit	LU Status	Lexical Entry Report	Annotation Report	Annotator ID	Created Date
annoyed.a	**Finished_Initial**	Lexical entry	Annotation	KmG	04/09/2008 03:23:46 PDT Wed
frustrated.a	**Created**	Lexical entry		KmG	04/21/2008 02:43:45 PDT Mon
irritated.a	**Created**	Lexical entry		KmG	04/09/2008 03:23:56 PDT Wed
sick and tired.a	**Created**	Lexical entry		804	09/26/2019 03:54:06 PDT Thu
tired (of).a	**Created**	Lexical entry		804	09/26/2019 03:55:25 PDT Thu

FIGURE 3.5 – Lexèmes associés du *frame « Annoyance »*

Un autre type d'annotation est aussi présent dans *FrameNet*, les annotations dites *full-text*. Elles sont issues d'un processus différent consistant à parcourir des textes continus plutôt que des phrases individuelles. Cela permet d'obtenir une distribution réelle des cadres et des rôles en corpus.

FIGURE 3.6 – Éléments *non noyaux* et liens du *frame « Annoyance »*

Dans la page d'entrée lexicale se trouve une définition simple du déclencheur, une table de *réalisation des rôles* et une table des *patrons de valence*. La première table énumère, pour chaque rôle pouvant être associé, le nombre de fois où il apparaît dans les phrases d'exemples annotées ainsi que ces réalisations syntaxiques (`constituant.fonction grammaticale`). La deuxième table instancie les patrons syntaxiques des rôles apparaissant dans les phrases

d'exemples avec aussi le type de constituants et leur fonction grammaticale.

Dans cette section, nous avons, à travers l'exemple du *frame Annoyance*, présenté la ressource *FrameNet* et les différentes informations qu'elle propose. Comme nous avons pu le voir, c'est une ressource extensive de données autour de la sémantique des cadres telle que définie par Fillmore. Elle présente l'avantage de structurer et de mettre en relation tout un ensemble de cadres et leurs rôles sémantiques associés, et les illustre grâce à un ensemble important d'annotations.

3.4 Adaptation de *Framenet* pour les autres langues

Les travaux de Boas (2009) et Tonelli (2010) dans un second temps ont étudié l'adaptation du modèle *FrameNet* pour d'autres langues. À travers un travail sur corpus en espagnol, allemand et japonais, Boas (2009) démontre que la structure de ce modèle est transposable. L'utilisation des cadres permet de contourner certains problèmes communs des analyses cross-linguistiques comme les structures polysémiques divergentes, les patrons lexicalisés... Les auteurs alignent des lexiques, dictionnaires bilingues et corpus parallèles. Puis, ils comparent des données attestées annotées et créent des entrées parallèles pour les unités lexicales. Ces dernières permettent de relier les unités lexicales en anglais et dans la langue cible, et donc le *frame qui en dépend*.

Dans un premier temps, nous souhaitons présenter les travaux ayant mené au *French-FrameNet*[35] (Djemaa, 2017; Djemaa *et al.*, 2016; Candito *et al.*, 2014). La création de cette ressource s'insère dans le cadre du projet ASFALDA[36] ayant eu cours de 2012 à 2016 et visant à la production d'un corpus d'annotation sémantiques et d'un analyseur sémantique appris sur ces annotations. Djemaa (2017) liste les raisons derrière le choix de *FrameNet* comme base pour ces deux objectifs, notamment par rapport à *PropBank* ou *VerbNet* :

1. l'orientation de *FrameNet* sur la sémantique plus que la syntaxe le rendant transcatégoriel ;
2. le nombre de projets dans d'autres langues qui montrent la portabilité des cadres ;
3. son orientation sémantique qui est avantageuse pour l'étude de l'interface syntaxe-sémantique ;
4. son association à des corpus annotés ;

35. http://asfalda.linguist.univ-paris-diderot.fr/frameIndex.xml
36. https://sites.google.com/site/anrasfalda/

5. son applicabilité prouvée par le nombre de ressources qui en ont découlé.

Dans ce projet, l'objectif cité de produire un corpus annoté en cadres et en rôles sémantiques se veut représentatif de la distribution des phénomènes linguistiques annotés. Ainsi, cela permet d'obtenir une quantification des phénomènes et, par là même, l'extraction d'un lexique quantifié, mais aussi, une meilleure exploitation pour l'apprentissage d'un modèle probabiliste.

Dans ce cadre, la focalisation du projet s'est effectuée sur sept domaines notionnels définis par les auteurs et non présents dans *FrameNet* : la causalité, la communication langagière, le jugement / l'évaluation, les positions cognitives, les relations spatiales, les relations temporelles et les transactions commerciales. Pour l'annotation, les auteurs ont fait usage des deux corpus que sont le *French Treebank* (Abeillé *et al.*, 2003) et le *Sequoia Treebank* (Candito et Seddah, 2012).

D'autres projets ont visé à adapter *FrameNet* pour d'autres langues. Par exemple, Subirats et Petruck (2003); Subirats (2009) présentent le projet *Spanish FrameNet* (SFN) pour l'espagnol issu d'une collaboration entre l'Universitat Autònoma de Barcelona et l'International Computer Science Institute à Berkeley. L'étude s'est faite sur des corpus variés (textes journalistiques, critiques littéraires, articles Wikipédia...) qui ont d'abord été annotés en morphologie, puis en cadres et rôles sémantiques. La méthode employée consiste à s'appuyer sur les cadres déjà définis dans la version anglaise et d'identifier en corpus les déclencheurs potentiels afin de répertorier leurs réalisations syntaxiques.

Un projet de création d'un *FrameNet* japonais a aussi vu le jour au sein de la Keio University et la Tokyo University intitulé *Japanese FrameNet* (Ohara *et al.*, 2004; Ohara, 2009). Dans une approche similaire au *Spanish FrameNet*, les annotateurs se basent sur les cadres préexistants pour l'anglais. Au cours de l'annotation du corpus, ils ont la possibilité de modifier le cadre afin de l'adapter au japonais, ou bien d'en créer un nouveau. Les corpus utilisés sont des articles du *Mainichi Newspaper* et des textes issus de nouvelles et d'essais.

Le projet allemand SALSA a aussi visé le développement d'un FrameNet à l'Universität des Saarlandes (Burchardt *et al.*, 2009). Ce projet a la particularité d'utiliser une méthode différente que les deux précédents puisqu'il procède à une annotation lemme à lemme. Les lemmes à annoter ont d'abord été définis sur la base de leur fréquence, puis leurs occurrences ont été annotées. Il s'avère aussi que pour allemand, deux autres projets pour la création d'un *FrameNet* sont actuellement en cours. D'autres langues ont aussi été sujettes à des projets de ressource à la *FrameNet* comme le suédois ou le néerlandais.

La visée de notre travail est d'obtenir une description sémantique fine des éléments

présents dans le corpus de REX d'Ariane 5 rédigé en français. Pour cela, une approche en rôle sémantique nous a paru adapté à ce cas d'étude. L'utilisation du *FrenchFrameNet* a donc été envisagée. Cependant, les domaines notionnels couverts par cette ressource ne recouvrent pas les classes de problèmes présents dans notre typologie (5) et donc, dans notre corpus, ni dans les situations couvertes par TRIZ. En revanche, les divers travaux autour de FrameNet démontrent la possibilité de transposer le *FrameNet* de Berkeley à d'autres langues. Dans la section 3.5, nous verrons aussi que cette transposition s'applique à d'autres domaines, ce que nous aurons l'occasion d'explorer notamment dans le chapitre 8.

3.5 Utilisation dans un corpus spécialisé

Selon L'Homme *et al.* (2020), la méthode *FrameNet* est régulièrement utilisée dans les domaines spécialisés, que ce soit pour la création de ressources spécifiques ou pour l'enrichissement du *FrameNet* de Berkeley. C'est une méthode privilégiée car elle permet de capturer les concepts propres aux domaines de connaissance. Par ailleurs, elle permet aussi de prendre en compte les propriétés linguistiques des termes et de les relier à leur contexte conceptuel. Ainsi, elle donne accès à une compréhension fine du comportement linguistique à travers un travail de corpus. Plusieurs ressources ont donc été créées pour des domaines spécifiques à partir de la méthode *FrameNet*, et dans des domaines variés tels que la biologie, le foot, le droit, l'environnement ou l'informatique. Dans cette section, nous présentons quelques-unes de ces ressources afin d'illustrer la pertinence de l'utilisation des *frames* de Fillmore et de *FrameNet* dans des corpus spécialisés comme le nôtre.

La première ressource que nous présentons est *BioFrameNet* (Dolbey, 2009), une extension du *FrameNet* originel focalisée sur la biologie moléculaire. L'idée derrière ce projet est d'étudier les propriétés grammaticales du langage de la biologie moléculaire dans des écrits scientifiques. Pour ce faire, l'auteur se focalise sur les unités lexicales et combinaisons sémantiques et syntaxiques qu'elles exhibent. De fait, cela permet aussi d'évaluer la capacité de la ressource *FrameNet* à traiter avec le langage de ce domaine précis. D'autre part, l'auteur souhaite explorer l'interaction entre les ressources lexicales et les ressources additionnelles du domaine telles que les ontologies ou les bases de connaissance, comme il en est l'usage dans l'étude des langues spécialisées (tel que nous avons pu le voir dans le chapitre 1 section 1.3).

Pour ce travail, le corpus d'étude est le GRIF (Gene References Into Function)[37]. Ce sont des textes courts (maximum 254 caractères) décrivant la fonction d'un gène. Cette brièveté

37. https://www.ncbi.nlm.nih.gov/gene/about-generif

imposée a pour conséquence les phénomènes suivants :

— une langue dense et compacte ;
— des noms composés fréquents ;
— des expressions multimots ;
— l'utilisation de la terminologie du domaine et d'entités nommées ;
— des métonymies pour référer à des fonctions et des processus autres.

Ces éléments impliquent donc une difficulté de compréhension pour des non-biologistes, mais aussi pour des biologistes non-experts de ce type de gènes. Ils impliquent aussi des difficultés dans l'optique d'un traitement automatique.

Afin de réaliser ce travail, l'auteur a créé des cadres spécifiques aux domaines reliés à ceux de *FrameNet* par des relations cadre-à-cadre. Il note aussi l'ajout de sens nouveaux à certains lexèmes déjà présents dans la ressource. La méthodologie utilisée reprend celle des annotations lexicographiques effectuées dans le projet originel.

Dans cette étude, nous pouvons noter une similarité avec la nôtre qui est la brièveté des textes (même si dans le cas des FA, nous n'avons pas connaissance d'un maximum imposé) et leur opacité pour un public non-expert. En revanche, Dolbey cite plusieurs ressources (ontologies et bases de connaissances) qui, en conjonction avec une consultation d'experts, lui ont offert une aide non négligeable vers une compréhension de ces textes. Ce type de ressource n'est cependant pas accessible dans le cadre de nos travaux. Le chapitre 8 présentera notre proposition de méthode afin de contourner cet écueil, c'est-à-dire comment appréhender des textes incompréhensibles sans expertise pour une analyse à la *FrameNet*.

Le second projet que nous souhaitons mentionner est le *Framed DicoEnviro* [38] de L'Homme *et al.* (2020). C'est une ressource multilingue (français, chinois, anglais, portugais et espagnol) concentrée sur le domaine de l'environnement. Elle traite plus spécifiquement des champs du changement climatique, des énergies renouvelables, du transport électrique, des espèces en danger et de l'agriculture durable. Les auteurs utilisent une méthodologie *bottom-up* similaire à celle utilisée par Schmidt (2009) dans son projet de *Kicktionnary*, un *FrameNet* pour l'univers du football, mais aussi couramment utilisée dans les projets de terminologie. Cette méthodologie définit huit étapes distinctes réparties en deux catégories recensées Table 3.2. Les cinq premières étapes sont appliquées sur chacune des langues séparément alors que les étapes de définition des *frames* sont réalisées sur l'ensemble des résultats.

Les auteurs expliquent que dans de nombreux cas, les cadres déjà présents dans *FrameNet* ont pu être réutilisés ou adaptés, mais plus de la moitié ont tout de même nécessité la création d'un nouveau *frame*. Au total, 112 nouveaux *frames* ont été créés, 38 sont restés inchangés et 40 ont été légèrement modifiés pour un total de 190 cadres étudiés. Les auteurs prennent

38. https://olst.ling.umontreal.ca/dicoenviro/framed/index.php

	Sélection des entrées terminologiques		**Déterminer les cadres des entrées lexicales**
1	Sélection des corpus spécialisés	6	Définition des cadres
2	Identification des termes	7	Encodage des cadres
3	Sélection et extraction des contextes	8	Définition des relations entre cadres
4	Définition des structures d'arguments		
5	Annotation des contextes		

TABLE 3.2 – Étapes de la méthodologie de création des cadres pour *Frame DiCoEnviro* (L'Homme *et al.*, 2020)

l'exemple de *inhabit* qui dans *FrameNet* est associé au cadre *Residence*. Cependant, dans le domaine des espèces en danger, il prend le sens d'« habitat » pour lequel « *a living entity finds itself in a given location that should provide it with what it needs to feed, reproduce and survive* ». L'élément noyau « Resident » du *frame* original devient le participant « Patient » avec la précision que le *Patient* est une espèce (en. *Species*).

Dans cet article, un ensemble de phénomènes propre aux domaines spécialisés rendent la tâche non triviale. Parmi ceux-ci, les auteurs mentionnent particulièrement le fait que certains termes ont une signification très proche dans le domaine spécialisé et dans la langue standard. En revanche, une étude plus approfondie montre des différences, dues à une conceptualisation différente d'une entité ou d'un événement en domaine de spécialité. Pour illustrer cela, les auteurs prennent l'exemple d'un « lieu de vie spécifique » du point de vue des espèces en voie de disparition. En essayant de rapprocher cette situation du cadre de *Residence*, plusieurs inadéquations apparaissent qui sont dues à la restriction de ce cadre aux êtres humains uniquement :

— certains déclencheurs listés dans *FrameNet* ne s'appliquent pas aux animaux et inversement ;
— les participants noyaux ne sont pas les mêmes en situation de voie de disparition ;
— les relations inter-cadres diffèrent en situation de voie de disparition.

De même que les auteurs, nous observerons dans le chapitre 8 que les *frames* mis à disposition dans *FrameNet*, bien que proches des notions avec lesquelles nous traitons ne sont pas exactement représentatifs du concept en question. Ainsi, un travail d'adaptation du cadre de départ est nécessaire pour appréhender le concept tel qu'il se manifeste dans le domaine de spécialité qui nous concerne.

3.6 La sémantique des cadres et le TAL

De nombreux travaux en TAL ont fait usage des rôles sémantiques et de *FrameNet*. Ces travaux s'appliquent à des domaines variés tels que les systèmes de questions-réponses, l'extraction d'information, le résumé automatique... Ces applications passent par une tâche d'annotation en rôles sémantiques (*Semantique Role Labeling*). Par ailleurs, avec le développement récent des LLM (Large Language Models) génératifs, de nouvelles applications apparaissent comme par exemple l'annotation automatique en *frames* par un LLM. Dans cette section, nous présenterons tout d'abord la tâche de *Semantic Role Labelin*g, soit comment automatiser la méthode *FrameNet*. Puis, nous explorerons quelques-uns des projets récents autour des LLM et des *frames* que nous aurons l'occasion de mettre à l'épreuve dans le chapitre 9 de cette thèse.

3.6.1 *Semantic Role Labeling*

Tout d'abord, il convient de différencier le *Semantic Role Labeling* du *Frame Semantic Role Labeling*. Ce dernier est utilisé pour parler de la tâche qui fait spécifiquement appel aux *frames* tels que conçu par Fillmore et dans *FrameNet*. Le premier est utilisé de façon générique pour parler de cette tâche, qu'elle fasse écho aux *frames* de Fillmore ou à d'autres ressources ou méthodes faisant appel aux rôles sémantiques (notamment la ressource *PropBank* mentionnée section 3.3 de ce chapitre).

La tâche de *Semantic Role Labeling* (SRL) consiste à automatiser l'identification des rôles sémantiques, des arguments et des prédicats d'une phrase (Jurafsky et Martin, 2024). L'utilisation de *FrameNet* et *PropBank* permet de définir les prédicats et les rôles pertinents. Ces ressources peuvent aussi servir pour entraîner un modèle. Les premiers travaux autour de l'automatisation de cette tâche avec des techniques d'apprentissage automatique (supervisé dans le cas présent) appartiennent à Gildea et Jurafsky (2000) et Gildea et Jurafsky (2002), respectivement basés sur *FrameNet* d'une part et sur *PropBank* de l'autre. Avant cela, de nombreuses études sur des systèmes à base de règles ont eu lieu, notamment dans les années 1970 comme ceux Hendrix *et al.* (1973), Simmons (1974), Wilks (1973), etc.

Dans Palmer *et al.* (2010), les auteurs définissent la tâche comme une tâche de classification : pour un verbe et des constituants donnés, sélectionner le rôle sémantique parmi une liste finie de rôles. Après avoir extrait les participants et leurs annotations, il est possible d'entraîner un classifieur afin de prédire l'étiquette à attribuer. L'évaluation se fait sur le *matching* entre les étiquettes *gold* et les étiquettes prédites. La tâche est traditionnellement divisée en deux étapes. La première consiste à identifier les constituants qui sont les arguments du

déclencheur. La seconde vise à l'attribution du rôle sémantique aux participants.

Pour effectuer ces tâches, plusieurs approches ont été mises en place. Parmi ces approches, nous retrouvons l'utilisation d'algorithmes d'apprentissage supervisée comme les arbres de décisions (Surdeanu *et al.*, 2003), les SVM (*Support Vector Machine*) (Pradhan *et al.*, 2005) ou encore la régression logistique (Toutanova *et al.*, 2005). Des études avec des approches semi-supervisées ont aussi été menées, notamment dans Fürstenau (2009) ou Gordon et Swanson (2007). Dans les travaux de Fürstenau (2009), l'idée est d'augmenter les données d'entraînement en sélectionnant des textes non annotés, mais extrêmement similaires aux données annotées. Puis, de projeter les annotations sur les textes non annotés. Dans les travaux de Gordon et Swanson (2007), les auteurs se concentrent sur les prédicats qui ne possèdent pas de données annotées. Ils projettent des étiquettes à partir de ceux des participants dont les prédicats, issus du corpus d'entraînement, sont syntaxiquement similaires aux prédicats sans annotations. Des études avec une approche non supervisée, c'est-à-dire sans utiliser de données annotées, ont aussi été menés. Swier et Stevenson (2004) utilisent la ressource *VerbNet* afin de contraindre un classifieur lors de l'annotation.

Comme déjà mentionné, le *Semantic Role Labeling* n'implique pas nécessairement l'utilisation de *FrameNet*. Cependant, des travaux visant spécifiquement à l'automatisation à partir de cette ressource ont aussi été menés. Ils ont notamment donné lieu à plusieurs outils comme SEMAFOR[39] (Das *et al.*, 2014), FRAMAT[40] (Roth et Woodsend, 2014; Roth et Lapata, 2015) ou OPEN-SESAME[41] (Swayamdipta *et al.*, 2017). Par ailleurs, de nombreuses avancées ont eu lieu lors de la tâche 19 du défi *SemEval 2007*, focalisé sur cette thématique. De même, le défi *CoNLL* de 2005, 2008, 2009 et 2012 fut concentré sur du SRL mais à partir de la ressource PropBank.

3.6.2 Les *frames* et les LLM conversationnels

Les performances récentes des agents conversationnels (*chatbots*) basés sur des LLM impliquent une reconsidération de la place des modèles de langues génératifs dans l'étude des phénomènes linguistiques. De nombreux travaux se sont concentrés jusqu'alors sur la capacité des LLM à représenter des informations ou des connaissances linguistiques (Rogers *et al.*, 2020; Mahowald *et al.*, 2024). Torrent *et al.* (2024) proposent plutôt de poser la question de l'utilité de ces outils pour la pratique de la linguistique. Est-ce que les IA conversation-

39. http://www.cs.cmu.edu/~ark/SEMAFOR/
40. https://github.com/microth/mateplus
41. https://github.com/swabhs/open-sesame

nelles peuvent servir d'assistant/outil au linguiste ? Les auteurs reprennent l'interrogation posée par (Dis *et al.*, 2023) comme objectif : « find a way to benefit from conversational AI without losing the many important aspects that render scientific work one of the most profound and gratifying enterprises : curiosity, imagination and discovery ».

À cette fin, Torrent *et al.* (2024) explorent l'idée d'utiliser les agents conversationnels basés sur des LLM en tant qu' « assistants », ou « co-pilotes », pour des tâches d'analyse linguistique. Parmi celles-ci, plusieurs expérimentations reposent sur l'articulation avec les *frames* et *FrameNet* et notamment sur l'utilisation d'un agent conversationnel dans l'extension et l'affinement de la ressource. Nous choisissons de les présenter pour illustrer ce type de travaux qui émerge.

Le premier point sur lequel se concentre l'étude est la capacité du *chatbot* à assister le linguiste dans l'extension de *FrameNet*. Cette extension peut correspondre à trois aspects différents : la création de nouveaux cadres, l'ajout de nouveaux exemples annotés ou l'identification de nouvelles unités lexicales (déclencheurs) à des cadres déjà existants. C'est cette dernière qui fait l'objet de ce travail. Afin de réaliser cette tâche, une approche comparative de deux *chatbots*, ChatGPT (OpenAI *et al.*, 2023) et OpenAssistant (Köpf *et al.*, 2023), a été menée. Un même prompt a été sélectionné avec une combinaison de trois variables qui sont modifiables, pour un total de douze scénarios testés. Ces trois variables sont le nombre de participants du *frame* (nombreux *vs* peu nombreux) ; le nombre d'unités lexicales du *frame* (nombreux *vs* peu nombreux) ; la « famille » du *frame* (*Event*, *Entity*, *Attribute*), c'est-à-dire si le *frame* hérite d'un de ces trois *frames* susmentionnés. Ces douze scénarios sont appliqués en promptant le *chatbot* pour cinq *frames* différents pour la génération de dix déclencheurs en réponse. Par cette étude, les auteurs souhaitent répondre aux trois questions suivantes :

— Dans quelle mesure un *chatbot*, en tant qu'assistant, est utile pour l'augmentation des unités lexicales de *FrameNet* ?
— À quelle point la quantité d'information influence l'efficacité du *chatbot* ?
— Quel type d'information influe le plus sur la qualité de la réponse ?

Les résultats obtenus permettent d'arriver à la conclusion que l'agent conversationnel est un bon « co-pilote » pour cette tâche. Pour cinq des douze scénarios proposés à ChatGPT, plus de 85% des déclencheurs sont corrects. Seul un des scénarios produit moins de 50% de réponses correctes. Pour ce cas, on indique qu'un utilisateur réel pourrait simplement diminuer le nombre demandé de déclencheurs afin de réduire la quantité de réponses incorrectes.

La seconde tâche que les auteurs explorent est l'utilisation des agents conversationnels pour l'expansion de *FrameNet* à d'autres langues. Comme discuté dans la section 3.4, la ressource a été adaptée dans plusieurs langues et les cadres se sont souvent révélés transposables d'une langue à l'autre. En partant de ce postulat, les auteurs souhaitent évaluer la

capacité d'un *chatbot* à proposer des déclencheurs dans une langue différente pour un cadre déjà existant et décrit en anglais. La langue choisie est le portugais brésilien. La différence majeure avec l'expérience précédente est une modification du prompt. Dans la première expérience, le prompt demandait des unités lexicales additionnelles, différentes de celles déjà présentes dans *FrameNet*. Dans cette expérience, il est seulement demandé des unités lexicales pertinentes par rapport au *frame* donnée. Les résultats sont satisfaisants et valident l'hypothèse. Ils observent tout de même que les cadres avec un grand nombre d'unités lexicales en anglais obtiennent de meilleurs résultats lors de la génération d'unités lexicales dans une autre langue.

La dernière tâche consiste à la construction d'un *frame*, c'est-à-dire la génération de cadres qui héritent d'un cadre donné (*Entity, Gradable_Attribute* et *Event*, dans le cas présent) ou d'un cadre associé d'après un ensemble d'unités lexicales. Les résultats présentés montrent une efficacité de cette tâche par les agents, mais qui est extrêmement dépendante du prompt et du fil de conversation qui a eu lieu (demandes de corrections, questions complémentaires...). Selon les auteurs, cette dépendance au prompt, dont le fait que l'agent puisse répondre à de fausses informations s'il est « prompté » en ce sens, démontre le manque de connaissances du fonctionnement linguistique de ces modèles. Ils démontrent aussi que ces agents sont avant tout des interlocuteurs utiles pour faire émerger des perspectives, des analyses supplémentaires pour le linguiste.

Dans cette section, nous souhaitons aussi aborder la question du *Semantic Role Labeling* présenté dans la section précédente, mais par le biais d'un agent conversationnel. Cette tâche fait directement écho à celle effectuée dans le cadre de notre travail et que nous présentons en détail au chapitre 9.

Devasier *et al.* (2025) ont mené une étude évaluant la capacité d'un agent conversationnel à annoter du texte en rôle sémantique (*Semantic frame parsing*). La tâche est habituellement divisée en trois sous-tâches : l'identification du déclencheur, du *frame* et des participants (dans le cas présent, les deux premières sous-tâches sont traitées conjointement par l'utilisation d'un modèle d'identification de cadre). Les tests sont effectués sur trois modèles différents, Qwen 2.5 (Yang *et al.*, 2024), Llama 3 (Grattafiori *et al.*, 2024) et GPT-4o (OpenAI *et al.*, 2023), mêlant ainsi des modèles de petite et grande taille (de 0.5 milliard de paramètres à 78 milliards). Cette étude se concentre sur trois points : une évaluation des différents formats d'*input* et d'*output* pour le *frame semantic parsing* avec un LLM génératif, la production d'un benchmark avec différentes architectures de LLM, et une nouvelle méthode pour identifier les *frames*.

Pour le premier point, les auteurs testent quatre formats de représentation de données dans leur *prompt* : *Markdown*, *XML* et deux variations de JSON. Ils démontrent que le format

JSON en pour l'*input* produit des résultats significativement meilleurs.

Les différents LLM sélectionnés permettent de faire varier la taille des modèles, leur accessibilité (*open-source* ou non) et l'architecture. Les résultats ont montré que le modèle Qwen 2.5 est plus performant que Llama 3.3, alors que Qwen 2.5 est beaucoup plus petit que Llama 3.3 (respectivement 3 milliards et 70 milliards de paramètres).

Ces travaux ont aussi montré que les approches *zero-shot* et *few-shots* (sans exemples ou avec peu d'exemples) sont moins performants que lorsqu'un plus grand nombre d'exemples sont fournis au modèle.

Par ailleurs, des tests ont été menés sur le corpus YAGS (Yahoo ! Answers Gold Standard) (Hartmann *et al.*, 2017). C'est un *dataset* composé de questions et réponses générées par l'utilisateur annotées en *frame*. Il a été conçu afin de tester des modèles de *Frame Semantic Role Labeling* sur un corpus d'un domaine différent des données de *FrameNet* (en l'occurrence, des données conversationnelles en ligne). Les résultats se sont significativement dégradés lors de la réalisation de la tâche sur ce corpus démontrant les limites du LLM à généraliser et à être appliqué à des domaines de spécialité (comme nous le verrons au chapitre 9).

Parmi les travaux impliquant les LLM conversationnels et leur utilisation pour de l'analyse linguistique, nous pouvons aussi citer Sanacore (2023). L'auteur en fait usage afin de générer des histoires permettant d'étudier les relations sémantiques récurrentes entre les lexèmes des familles dérivationnelles. Par ailleurs, Ettinger *et al.* (2023) utilise aussi ces agents comme outil, et dans le cas présent comme « annotateur expert en linguistique ». Ces travaux se focalisent sur la capacité de ce type de modèles à analyser la structure sémantique d'un énoncé, en faisant notamment appel au formalisme AMR (Abstract Meaning Representation) (Banarescu *et al.*, 2013).

Dans ce chapitre, nous avons présenté les rôles sémantiques et la théorie de la sémantique des cadres. Nous avons aussi exploré la ressource *FrameNet* qui implémente les différentes notions autour de cette théorie et met à disposition tout un ensemble de données formalisées, décrites et mise en action. Cette ressource a servi de cadre, de point d'appui dans l'analyse que nous effectuons des situations d'anomalies décrites dans notre corpus. Par ailleurs, de par les adaptations de *FrameNet* vers d'autres langues, et par la création de ressources « *à la FrameNet* », nous avons pu constater la portabilité de cette méthode vers d'autres corpus, en accord avec l'usage que nous en faisons. C'est-à-dire, une transposition de certains cadres vers le français, et pour le domaine du dysfonctionnement technique dans un contexte de culture de sécurité. Enfin, nous avons vu les différentes interactions entre les *frames*, *FrameNet* et

le TAL. Dans un premier temps, nous nous sommes intéressés à la tâche de *Semantic Role Labeling*, soit comment obtenir un système capable de reconnaître les différents constituants dans un énoncé, leur rôle et le cadre auxquels il se rapporte. Dans un second temps, nous avons pu voir comment les LLM conversationnels, envisagés comme assistants du linguiste, peuvent être utiles dans l'étude de l'analyse sémantique à la *FrameNet*.

Deuxième partie

Le cas d'étude des Fiches Anomalies Ariane 5

Chapitre 4

Un corpus de REX bruité

Sommaire

Dans ce chapitre, nous présentons le corpus de travail avec lequel nous avons travaillé : les Fiches Anomalies produites lors des campagnes de lancement du lanceur Ariane 5. Nous commençons par présenter le contexte de production de ces documents et le format général sous lequel ils nous ont été présentés. Puis, nous observons son contenu lexical à travers trois prismes. Le premier consiste en une analyse de fréquences afin notamment d'obtenir un premier aperçu du contenu des données. Le second présente les différents marqueurs de spécialités que nous y retrouvons. Le dernier se focalise sur le bruit présent dans le corpus. Enfin, nous faisons une brève démonstration de l'impact de ces caractéristiques sur les étapes traditionnelles de prétraitement des données en TAL.

4.1 Genèse et format

Les données avec lesquelles nous travaillons constituent un corpus de REX (Retours d'EXpériences). Il comprend des Fiches Anomalies (FA) rédigées pendant les différentes campagnes de lancement du lanceur Ariane 5 ayant eu lieu entre octobre 1997 et janvier 2011, pour un total de 35 242 FA. Elles ont toutes été produites lors de « campagnes sol ». L'appellation « campagnes sol » désigne spécifiquement le lieu d'apparition des anomalies, c'est-à-dire la base au sol proche du lanceur Ariane 5 au Centre Spatial Guyanais (CSG),

et non lors des vols de la fusée. Ces documents font échos aux procédés de REX dans les environnements industriels et particulièrement dans les domaines à risques. Ils impliquent une pratique du REX systématique devant toute situation non conforme à l'attendu (comme nous avons pu en discuter au chapitre 1). Une FA est donc créée par un opérateur pour tout écart à l'attendu rencontré, aussi minime soit-il. Par exemple, nous pouvons retrouver au sein du corpus de FA la description de l'anomalie suivante : « LA SOURIS N'A PLUS DE BOULE. » pour un problème avec un matériel informatique. Ces constats d'anomalie sont d'ailleurs issus d'une rédaction de rapport sur le vif, dans une pratique systémique.

Une fois le constat de l'anomalie rapporté, une première évaluation par un responsable technique est effectuée afin de déterminer si l'anomalie fait l'objet d'une FA à proprement parlé, c'est-à-dire du processus visible en Figure 4.1. Si le responsable qualité ne considère pas l'anomalie comme constitutive d'un problème, la situation est « laissée en l'état » et la FA est annulée. Ces rapports ne figurent donc pas dans le corpus que nous étudions. Si l'anomalie est considérée constitutive d'un problème, une procédure d'analyse dans le but de déterminer la cause de l'anomalie est enclenchée. Cette procédure peut être plus ou moins longue et ardue en fonction de la nature du problème. Elle peut inclure différents acteurs, mais est effectuée par des spécialistes techniques. Suite à la détermination de la cause, soit une solution est mise en place, soit il est décidé de laisser la situation en l'état. Le rapport original auquel donne lieu la FA peut constituer un document de taille conséquente en fonction de la gravité du problème et de la difficulté posée. Malheureusement, nous n'avons pas accès à ce type de document.

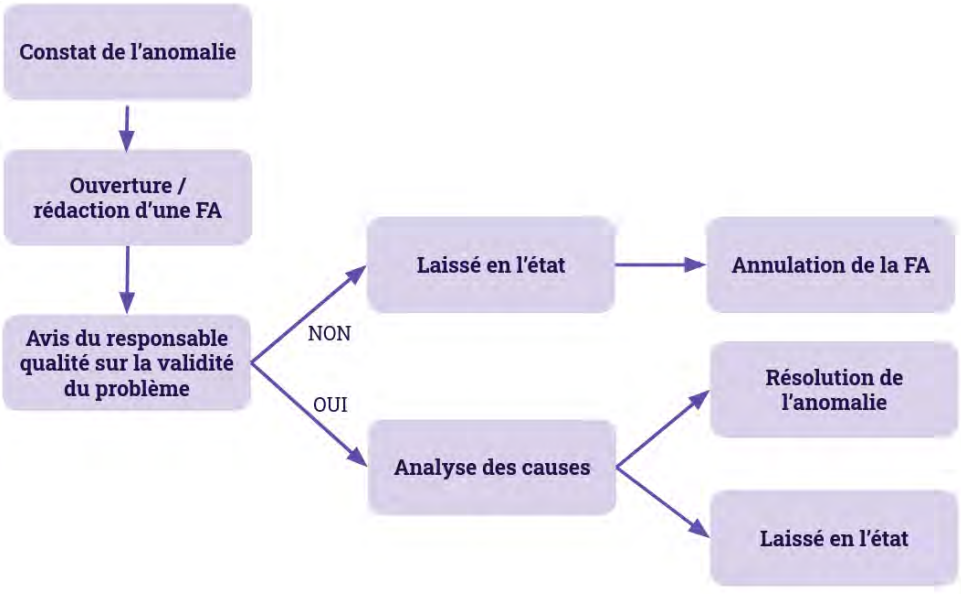


FIGURE 4.1 – Cycle de vie d'une FA

Nous n'avons pas pu obtenir de description plus précise du processus complet des FA, mais le corpus avec lequel nous travaillons s'apparente plutôt à une liste de Fiches Anomalies résumées de façon minimale, automatiquement extraite depuis une base de données. Cette liste prend la forme d'un document de type tableur CSV dans lequel nous retrouvons 19 champs distincts listés en Table 4.1. Toutes les FA sont rédigées en français, et le tableur mêle champs avec du texte libre et champs avec des métadonnées (lieu de l'anomalie, numéro d'opération, numéro de campagne...). Tous ne sont donc pas susceptibles d'être pertinents à un traitement linguistique. Nous pouvons noter cinq champs avec du texte libre (en gras dans le tableau) : Description de l'anomalie, Cause(s), Action(s) Niveau 1 proposée(s), Action(s) Niveau 1 réalisée(s) et Action Observations.

Index	Nom du champ	Description	Champ en texte libre
1	Numéro campagne	Numéro de la campagne de lancement	Non
2	Numéro chrono	Numéro attribué automatiquement dans l'ordre d'apparition lors e la campagne	Non
3	Date du constat	Date du rapport d'anomalie	Non
4	État de la FA	Est-ce qu'elle est fermée ou encore en cours.	Non
5	**Description de l'anomalie**	Courte description de l'anomalie rencontrée	Oui
6	Code ensemble	La pièce ou le système impacté	Non
7	Code Lieu	Lieu physique où s'est déroulée l'anomalie	Non
8	Attribution urgence	Spécialité technique	Non
9	Num d'opération	Code d'opération	Non
10	Libellé de l'opération	Opération pendant laquelle l'anomalie est apparue	Non
11	**Cause(s)**	Diagnostic du problème	Oui
12	**Action(s) Niveau 1 proposée(s) par RT**	Proposition de solution à apporter par le Responsable Technique	Oui
13	**Action(s) Niveau 1 réalisée(s)**	Solution mise en place	Oui
14	Utilisation	Métadonnée de classification	Non
15	Impact sauvegarde	Indicateur de gravité	Non
16	Échéance	Cible temporelle pour la réalisation de l'Action Niveau 1	Non
17	**Action Observations**	Observation faite suite à l'application de la solution	Oui
18	Code Action N1	Classification des Actions proposées et réalisées	Non
19	Date de clôture	Date à laquelle la FA est clôturée	Non

TABLE 4.1 – Liste des différents champs du tableur des FA

Tous les champs en langage naturel ne sont en revanche pas renseignés avec le même degré d'informations. La Table 4.2 nous indique la moyenne du nombre de mots par champ. Comme nous pouvons le voir, la moyenne pour les champs « Cause(s) » et « Observation(s) » est basse avec respectivement 5,4 et 1,9 mots par FA. Ces nombres s'expliquent de différentes

Champ	**Nombre de mots maximum**	**Nombre de mots minimum**	**Nombre de mots moyen**	**Renseignement non descriptif (en %)**
Description de l'anomalie	157	2	19,3	0,01
Cause(s)	75	1	5,4	32,70
Action(s) proposée(s)	146	1	12,7	5,84
Action(s) réalisée(s)	178	1	13,6	2,11
Observation(s)	73	1	1,9	35,12

TABLE 4.2 – Moyenne du nombre de mots par champ des FA

façons. D'une part, si tous les champs sont systématiquement renseignés dans le tableur, cela ne signifie pas qu'ils sont remplis avec des informations détaillées et profitables. En effet, on retrouve de nombreuses FA avec des valeurs telles que « NEANT », « * », « En cours », « RAS », *etc.*. Nous observons 32,70% de cas de cet acabit pour le champ « Cause(s) » et 35,12% pour le champ « Observation(s) ». Les 0,01% du champ « Description de l'anomalie » correspondent à quelques occurrences de « FA ANNULEE ». Ces valeurs ne donnent pas d'information précise et descriptives de la situation. Elles ne nécessitent donc pas de traitement spécifique. Dans des cas comme « * », elles révèlent même simplement de l'obligation d'avoir un champ qui ne soit pas vide, probablement en raison du format initial de la base de données à partir de laquelle le corpus a été extrait. Le champ « Description de l'anomalie » est le seul pour lequel nous ne retrouvons pas ce type de réponse, à raison puisque c'est lui qui donne lieu à la FA. Nous observons aussi une corrélation entre le nombre de mots maximum (relativement) élevé et le faible taux de renseignement du champ pour « Description de l'anomalie », « Action(s) proposée(s) » et « Action(s) réalisée(s) ». Nous faisons donc face à des données parcellaires et dont le texte est peu développé.

La Table 4.3 montre deux FA en exemple. La colonne « FA maximaliste » montre une FA avec tous les champs renseignés « généreusement ». Il est même fait état de deux anomalies dans le constat. Le fait de lister plusieurs anomalies dans une FA est un comportement que nous pouvons observer régulièrement dans le corpus, bien que non conforme à la procédure qui indique qu'une fiche doit être réalisée pour chaque anomalie rencontrée. En conséquence, les autres champs adressent aussi les deux anomalies en description. Nous pouvons aussi observer une répétition du champ « Action(s) proposée (s) » et « Action(s) réalisée(s) », correspondants aux cas où l'action proposée a été adoptée en tant que solution. À l'inverse, la FA de la colonne suivante montre un cas où très peu de texte est renseigné. Les champs « Cause(s) », « Action(s) réalisée(s) » contiennent seulement un astérisque. Le champ « Observation(s) » signale uniquement la fermeture de la FA. Le champ « Action proposée(s) » est utilisé pour proposer de renvoyer le problème. Seule la description de

l'anomalie est véritablement donnée. Ces exemples illustrent l'hétérogénéité des données du corpus. Cette hétérogénéité implique que tous les champs du corpus ne pourront pas être analysés de façon similaire. Elle nous incite aussi à focaliser les analyses sur le champ « Description de l'anomalie ».

Champ	**FA maximaliste**	**FA minimaliste**
Description de l'anomalie	- ABSENCE DE PROTECTION ANTI - PLUIE NIVEAU PC GAT ET CONNECTEURS PC BACCHUS ET TMA - CONDENSATION IMPORTANTE SUR PLAQUE PC GAT- AXE ETRIER DE LARGUAGE PC BACCHUS	POE CASE : DETERIORATION DES SURFACES DE VENTILATION + JOINTS(REF PRISE : 8900-763 A N° DE SERIE 4). OXYDE
Cause(s)	-CONF DE LIVRAISON BIL AU LIEU DE BEAP ?-NON RESPECT DE LA SMO-CLIM-11 SUR LE TRANSITIVE DE LA CLIMATISATION DU BIL 3 HEURES AVANT OUVERTURE DES PORTES ? (A CONFIRMER)	*
Action(s) Niveau 1 proposee (s) par RT	1/ MISE EN PLACE PROTECTION PC GAT.2/ SECHAGE AVANT MISE EN PLACE PROTECTION.3/ VIE EN L'ETAT.	ENVOI EN REVALIDATION CHEZ SOURIAU.
Action(s) Niveau 1 realisee(s)	1/ MISE EN PLACE PROTECTION PC GAT.2/ SECHAGE AVANT MISE EN PLACE PROTECTION.3/ VIE EN L'ETAT.	*
Action Observations	1/ 12/05/98 : PROTECTION DES CONNECTEURS PLIANE+SCOTCH (MS003P2) EN ATTENDANT LA CTTA. CTTA Nr 45 DU 14/5 :3/ AXE A GRAISSER : FAIT LE 19/05/98 AVEC DE LA GRAISSE CONOKO.CONNECTEURS NON ETANCHES EN AMBIANCE GUYANAISE.MISE EN PLACE PROTECTION SUIVANT SMO'S 836.	FA CLOSE.

TABLE 4.3 – Deux exemples de FA

4.2 Les différentes facettes du corpus

Dans cette section, nous abordons l'étude du corpus et de son contenu. Nous faisons un état des lieux des différentes caractéristiques de ces données. Pour cela, nous observons tout d'abord les régularités lexicales des différents champs en langage naturel du corpus. Puis, nous observons les différentes marques du domaine de spécialité et dans quelle mesure elles interfèrent avec l'appréhension des données. Enfin, nous nous arrêtons sur la problématique du bruit dans le corpus.

4.2.1 Quelques fréquences

Afin d'obtenir un premier aperçu du contenu des FA, nous avons choisi d'effectuer un calcul de fréquence pour observer le lexique récurrent dans le corpus. Cela nous a permis d'observer les tokens les plus fréquents et nous présentons dans la Table 4.4, pour chaque

champ avec du langage naturel, les 15 tokens les plus fréquents avec une segmentation simple sur les espaces. Tout le texte est mis en minuscule et les mots vides sont supprimés[42]. Les résultats sont présentés par ordre décroissant.

Nous pouvons d'abord observer que les tokens du champ « Description de l'anomalie » (« lors, », « non », « niveau »...) les plus fréquents apparaissent moins souvent que les tokens les plus fréquents des autres champs. Nous comptons 3330 occurrences pour « lors », token le plus fréquent du champ « Description de l'anomalie », contre 4715 pour « remplacement », token le plus fréquent du champ « Action réalisée(s) ». Cela peut s'expliquer par une plus grande diversité de vocabulaire pour la description de l'anomalie, ce qui serait corrélé au nombre plus élevé de tokens dans ce champ (pour rappel, 19,3 tokens en moyenne par FA). Nous explorons plus en détails les tokens et bigrammes les plus fréquents de ce champ dans un second temps.

Description		**Cause**		**Action proposée**		**Action réalisée**		**Observations**	
Token	**Fréq.**	**Token**	**Fréq.**	**Token**	**Fréq.**	**Token**	**Fréq.**	**Token**	**Fréq.**
lors	3330	neant	5583	remplacement	5759	remplacement	4715	fa	5581
non	3068	cours	3344	etat	3456	ok	4611	close	4942
niveau	2487	non	3327	investigation	2907	etat	3860	note	464
table	2406	hs	1138	remise	2305	fa	2885	clos	365
defaut	2288	erreur	1119	faire	2185	reprise	2590	pm	244
entre	2266	vieillissement	1102	reprise	2131	place	2549	profit	234
cote	2220	defaut	1011	controle	2035	controle	2195	etat	228
hors	2112	anteriorites	1005	place	1939	ete	2130	remplacement	181
perte	1969	lors	893	mise	1862	realise	2046	realise	181
ligne	1842	fa	812	verification	1629	mise	1968	voir	174
vis	1777	indeterminee	736	si	1622	correct	1882	procedure	164
lh2	1730	procedure	662	procedure	1482	fait	1788	impact	162
fuite	1609	conformite	635	vie	1399	pm	1678	sol	157
apres	1549	suite	581	apres	1396	remise	1631	niveau	156
voir	1515	pb	566	etat	1277	procedure	1612	action	155
Nombre total de tokens par champs									
460 901		136 056		290 023		331 489		40 431	

TABLE 4.4 – Table des fréquences des mots par champ

Pour le champ « Cause(s) », Les deux tokens les plus fréquents sont « neant » et « cours », que nous retrouvons régulièrement dans le corpus avec « En cours » (le mot « en » ne figure pas dans la liste à cause de la suppression des mots vides dans lequel il est inclut). Ces deux tokens correspondent aux FA dont la cause n'est pas (encore) renseignée, de même qu'« indeterminee » en onzième position. Nous pouvons aussi noter la présence de termes

42. Nous avons utilisé la librairie NLTK pour la suppression des mots vides

évocateurs de la cause du problème tels que « hs », « erreur », « vieillissement », « defaut », Nous verrons par la suite que ces termes sont aussi fréquemment présents dans le champ de description du problème. Cela illustre la possibilité d'expliquer un problème par un autre problème, une anomalie initiale qui déclenche un problème en bout de chaîne, comme nous avons pu en discuter dans la section 1.2.3 du chapitre 1 portant sur l'analyse des causes.

Dans les champs « Action proposée(s) » et « Action réalisée(s) », nous pouvons observer une récurrence des tokens les plus fréquents. La matrice de corrélation Figure 4.2 montre le nombre de tokens communs entre deux champs du corpus parmi les 50 tokens les plus fréquents. Cette matrice indique clairement une forte cooccurrence des termes les plus fréquents avec 30 tokens communs entre les deux champs. Dans l'exemple Table 4.3, nous avons d'ailleurs pu voir un cas pour lequel les deux champs contiennent exactement le même texte. D'autres cas du corpus contiennent des textes entièrement ou extrêmement similaires. Ils correspondent à la situation dans laquelle l'action qui a été proposée pour résoudre le problème a été mise en place.

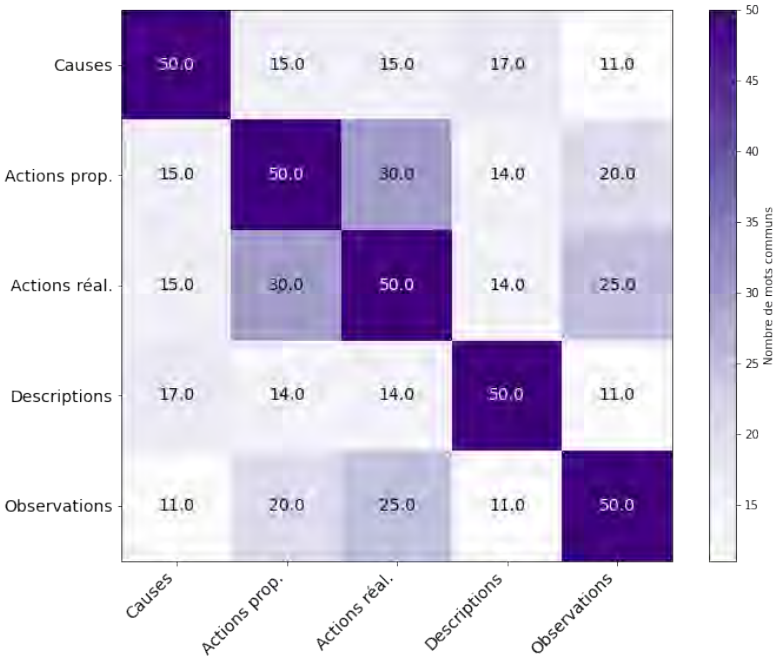


FIGURE 4.2 – Matrice du nombre de tokens communs entre les différents champs

En observant le champ « Observations », nous réalisons que seuls deux tokens sont présents en grande fréquence : « fa » et « close ». Les tokens suivants sont dix fois moins nombreux avec un passage de 4942 occurrences pour « close » et 464 pour « note ». En effet, la majorité des FA pour ce champ sont renseignées avec la mention « FA CLOSE », comme dans l'exemple Table 4.3, et seuls peu de cas font effectivement mention d'observations plus

détaillées. La matrice de confusion montre aussi de nombreux cas de tokens communs entre les champs « Observation(s) » et « Action réalisée(s) ».

En plus d'être plus fourni et systématiquement renseigné, le champ « Description de l'anomalie » est le plus intéressant par rapport à notre objectif de modéliser un problème technique. En effet, c'est dans ce champ qu'est décrite l'anomalie en question et dans lequel nous allons retrouver les informations nécessaires à notre formalisation. De ce fait, notre étude du corpus s'est principalement concentrée sur cette section. C'est pourquoi, nous choisissons ici de détailler plus avant le contenu lexical de ce champ.

La Figure 4.3 nous donne à voir les trente tokens les plus fréquents au sein du champ « Description de l'anomalie ». Parmi eux, plusieurs sont évocateurs de la notion de *problème* ou d'*anomalie* : *defaut*, *perte*, *fuite*, *hs*, *impossibilite*. Nous pouvons aussi observer plusieurs mots qui réfèrent à la localisation de l'anomalie : *table* (pour « table de lancement »), *cote* (« côté » ou « cote »), *BAF* (Bâtiment d'Assemblage Final), *local* et *EPC* (Etage Propulsion Central). Parmi ces derniers, nous pouvons observer que plusieurs font partie du lexique spécialisé des FA, comme BAF et EPC, pour lequel il est nécessaire de faire appel à un expert, ou à de la documentation, pour leur compréhension. Le texte complet de la FA ne suffit pas forcément pour leur compréhension et parfois même, pour saisir s'il s'agit d'une localisation ou non.

En observant les bigrammes Figure 4.4, cette tendance se confirme. En effet, nous observons une récurrence des évocations d'un problème avec les expressions *hors spec*, *hors specification*, *hors tolerance*, *non conforme*, *hors service*. De même, nous retrouvons plusieurs bigrammes qui font référence à des localisations : *tour cazes*, *BAF HE* (Bâtiment d'Assemblage Final Hall d'Encapsulation), *caisson lbs* (Liaison Bord Sol), *carneau EAP* (Étage d'Accélération à Poudre), *cote LOX* (OXygène Liquide), *carneau EPC*, *cote LH2*[43].

Ces deux aspects qui ressortent des tokens et des bigrammes les plus fréquents font écho à la nature des données. En effet, pour une Fiche Anomalie, il est naturel de retrouver dans les données des termes en rapport avec la notion d'anomalie, les opérateurs qualifient le problème qu'ils rencontrent. D'autre part, ces textes visent à rendre compte des problèmes en vue de leur résolution et d'empêcher leur reproduction. Ainsi, il n'est pas étonnant de retrouver mention de l'emplacement de l'anomalie.

4.2.2 Les marques de la spécialité

Au sein du corpus, nous pouvons remarquer des marques équivoques des langues de spécialités. En effet, les sigles, les acronymes ou bien les codes sont omniprésents dans les FA.

43. Nous supposons ici que les deux occurrences de « cote » correspondent à « côté »

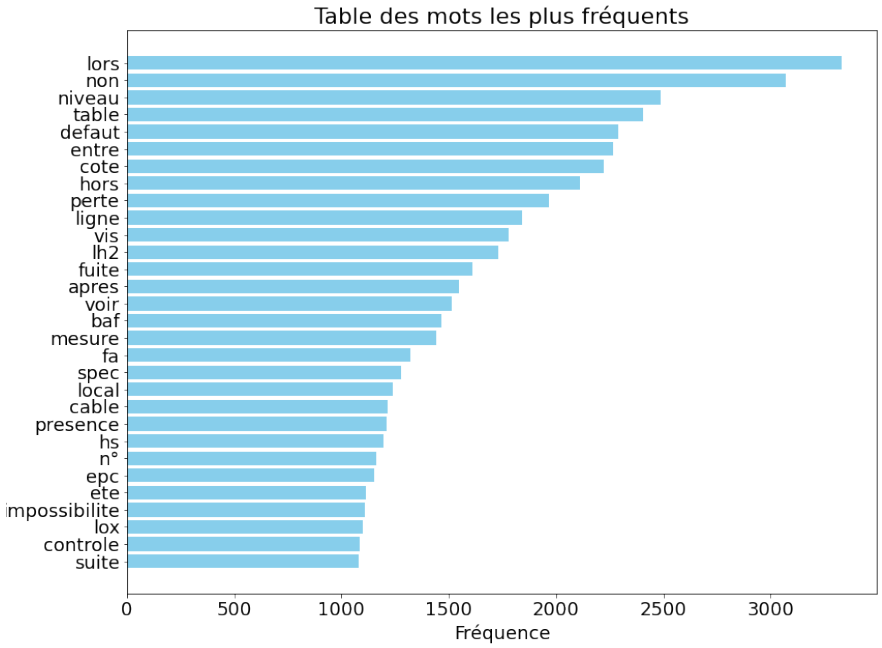


FIGURE 4.3 – Fréquence d'un mot dans le champ « Description d'un problème »

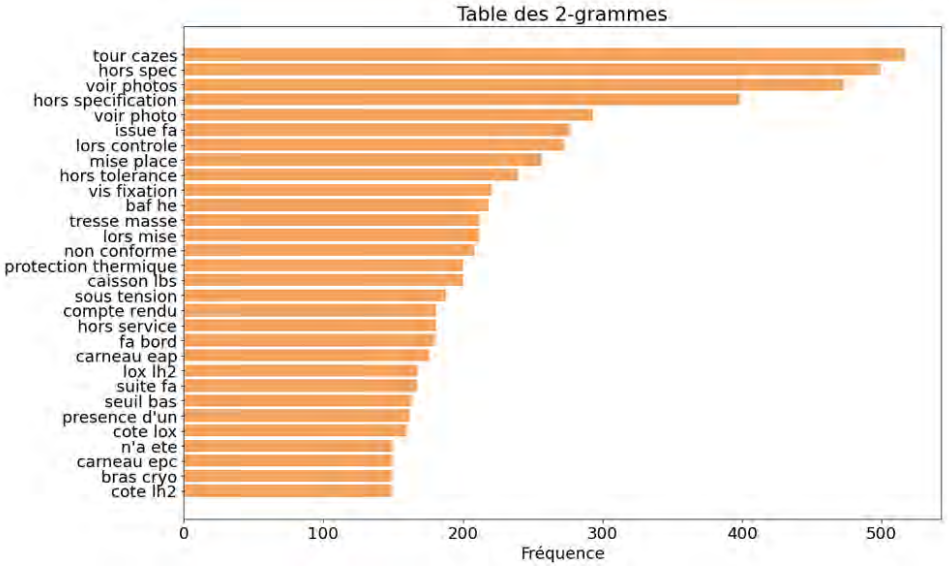


FIGURE 4.4 – Fréquence des bigrammes dans le champ « Description d'un problème »

Il est aussi possible de trouver du lexique usuel en langue standard, mais qui prend un sens particulier dans leur contexte de spécialité, dont nous présentons un exemple ci-dessous. Ce lexique est qualifié de « vocabulaire scientifique général » par Phal (1970) cité par (Vergely,

2004). Ces différents éléments sont récurrents dans les langages de spécialités, et la brièveté des textes du corpus peuvent concourir à leur utilisation. De plus, ils s'adressent à d'autres experts du domaine avec qui ils partagent un savoir commun. Ils rendent cependant la compréhension de ces textes difficile pour un non-expert.

Compréhensible	**Partiellement compréhensible**	**Incompréhensible**
LORS DU REMPLACEMENT DU FILTRE XFT005, IL A ETE TROUVE UN MORCEAU DE TEFLON DANS LA PARTIE AMONT.	A LA RECONNEXION LOGIQUE SCO1/TABLE 1 AU BAF, PERTE DES DEUX FRONTAUX DE SECURITE.	PROBLEME D'ACCROCHAGE TM SUR MAGALI EAP 2.

TABLE 4.5 – Exemple de « Description de l'anomalie » à différents niveaux de compréhension pour un non-expert

Les exemples Table 4.5 illustrent ce problème de compréhension avec trois exemples de description d'une anomalie qui posent une difficulté plus ou moins élevée. Le premier exemple, celui classé comme « Compréhensible »[44], présente une description de l'anomalie claire. Le seul élément qui pourrait poser une difficulté est le code « XFT005 ». Or, ce dernier sert uniquement à désigner le filtre dont il est question et ne modifie pas le sens de l'énoncé.

Le second exemple, qualifié de « Partiellement compréhensible », pose plus de difficulté pour un non-expert. Nous pouvons relever quatre éléments qui interfèrent avec la compréhension du texte :

1. « RECONNEXION LOGIQUE » : qui, sous réserve d'une connaissance en informatique, laisse entendre qu'il est question de réseaux informatiques ;
2. « SCO1/TABLE 1 » : utilisation d'un code ;
3. « BAF » : utilisation d'un acronyme ;
4. « PERTE » : vocabulaire scientifique général. Un entretien avec l'expert nous a en effet indiqué que l'utilisation de « perte » dans le cadre des FA fait systématiquement référence à une « perte de signal » ;
5. « FRONTAUX » : vocabulaire spécifique général à nouveau, qui fait référence à un ordinateur de gestion des communications[45].

Le dernier exemple de la Table 4.5 illustre un cas de description que nous qualifions d'incompréhensible pour des non-experts. Parmi les phénomènes déjà évoqués qui peuvent concourir à cette incompréhension, nous retrouvons les deux acronymes que sont « TM »

44. Cette classification a été opérée par nous-même dans le cadre de cet exemple et ne correspond pas à une classification officielle du corpus.
45. https://www.larousse.fr/dictionnaires/francais/frontal/35401

et « EAP 2 », ainsi que l'appellation : « MAGALI ». Si l'on fait fi de ces unités, le reste de l'énoncé se limite à « PROBLEME D'ACCROCHAGE » (dans ce cas, même le sens d'« ACCROCHAGE » reste flou selon nous). En demandant à un expert, il s'avère qu'il s'agit d'un problème de récupération de télémesures sur la station de décomutation des mesures (MAGALI) dédiée à l'étage EAP2 d'Ariane 5. La décomutation fait référence au traitement d'un signal sortant d'un capteur afin de le rendre traitable par des outils informatiques.

L'utilisation de ces codes, acronymes, sigles, *etc.* est certes commune aux langues de spécialité, mais elle prend une dimension supplémentaire au sein d'un corpus de REX dans un environnement critique comme le nôtre. Effectivement, nous pouvons observer que de nombreux codes viennent qualifier des composants du système. Nous avons pu le voir avec « FILTRE XFT005 » dans le premier exemple Table 4.5 et retrouvons des constructions similaires dans le corpus telles que : « VANNE HVM9019 », « FLEXIBLE FDV1 », « VANNE DE PURGE AVM 3109 », « CABLE ATN E1 J S106-5 »... Cette nécessité de qualifier les composants de façon précise s'explique d'une part par la nature de notre corpus, c'est-à-dire un corpus de REX. L'objectif premier d'un REX est de rendre compte d'une anomalie afin de la résoudre et d'en empêcher la reproduction. Il est donc important de savoir précisément l'endroit où elle est présente et de pouvoir la repérer immédiatement. D'autre part, dans un environnement critique tel que celui d'une base spatiale, chaque élément, chaque vanne, chaque tuyau, chaque câble possède une appellation afin de pouvoir l'identifier et s'y référer avec précision.

Parmi les éléments renvoyant au domaine de spécialité du corpus, nous pouvons aussi noter la présence d'unités de mesures. Ces unités sont variées et apparaissent régulièrement au sein du corpus. La multiplicité des composants (gazs, liquides, mais aussi solides) et la vérification constante de leur adéquation avec la norme nécessaire au bon fonctionnement du système les rend inévitables et variés. Dans le corpus, il est fait référence aux « ohm », « mohms » (miliohm), « mA » (miliampère), « mbar » (milibar), « V » (volt), « ms-2 » (mètre par seconde carrée)...

Dans ce corpus, nous pouvons aussi observer un effacement de l'énonciateur avec une utilisation d'une structure impersonnelle quasi-constante. En effet, sur les 35 242 FA du corpus, le pronom « je » est utilisé au total 15 fois. Le « nous » est employé plus souvent, soit en 234 occurrences, mais ce nombre reste proportionnellement peu élevé. Quelques exemples choisis sont présents en Table 4.6[46] afin d'observer l'emploi de ces pronoms dans le corpus. Les textes ici sont issus de différentes FA et de divers champs. Le premier exemple

46. Les noms propres ont été anonymisés dans cet exemple.

Emploi du "je"	Emploi du "nous"
LORSQUE LES OPERATEURS [NOM ORGANISATION] M'ONT DEMANDE LA FERMETURE DES VANNES REJETS CARNEAUX OVP053 ET OVP055, JE ME SUIS RENDU COMPTE QU'ILS AVAIENT OUVERT LA VANNE OVP056 (QUI N'ETAIT PAS DANS LEUR PROCEDURE) ET ILS ONT DONC REJETTE LES EAUX DANS LE CARNEAU EAP ALORS QU'IL Y AVAIT UN OPERATEUR SUR PLACE. LA PERSONNE [PRÉNOM NOM] A ETE ARROSEE MAIS A REUSSIT A EVACUER.	APRES VISUALISAT° DU FILM VIDEO, IL APPARAIT QUE LORS DE L'OUVERTURE SCR1, LE RUBAN A BOUGE MAIS IL S'EST VITE DECALE DE L'ORIFICE SCR1 ET LE JET SORTANT ETAIT DE PLUS DE TRAVERS DE SORTE QUE LE RUBAN SCR3 A ETE SEUL EN MOUVEMENT PENDANT LES 20 S D'ACTIVAT° ET DONC NOUS AVONS TOUS CONCLU A TORT EN PLATEFORME QUE L'OUVERTURE SCR1 CRACHAIT A LA SORTIE SCR3. EN CONCLUSION, IL N'Y A PAS DE PROBLEME.
USURE DE LA PARTIE AFFECTEE CAR LA VIS NE PENETTRE PAS ASSEZ. JE PROPOSE DE CHANGER TOUTES LA VIS EXISTANTES "CHC 16.60 CLASSE 12-9 PAR DE VIS PLUS LONGUES "CHC 16.75 CLASSE 12-9".	APRES DEPOSE DE LA COSYVE, NOUS CONSTATONS LA PRESENCE DE TRACES NOIRES A LA SORTIE DE CHAQUE INJECTEUR DES RAMPES DE BALAYAGE.
C'EST LE MEME PROBLEME (A L'ORIGINE) QUE CELUI DE LA FA S181/13 JE VOUDRAI CLORE CETTE FA AU PROFIT DE LA S181/13 CAR LE TRAITEMENT EST UNIQUE POUR LES 2 FA	AUCUNE ACTION D'AMELIORATION POUR LE PASSAGE DU TRACTEUR. NOUS ESSAIERONS DE VIVRE EN L'ETAT.
- CES RAYURES SONT CONNUES ET NE GENERENT PAS LES ETANCHEITES. JE PROPOSE DE VIVRE EN L'ETAT.	NOUS AVONS DES TEMOINS D'ENFONCEMENT DES AMORTISSEURS. LES VALEURS RELEVEES SONT CELLE DES TEMOINS.
AUCUN : JE SUIS GARANT DE MES OPERATEURS MAIS PAS DE TOUTES LES PERSONNES POUVANT ACCEDER AU NIVEAU INTERCONNEXION OU EFFECTUANT DES TRAVAUX DANS LE MAT.	NOUS AVONS CONSTATE QU'AU NIVEAU SIGNALISATION LUMINEUSE, IL Y AVAIT 10 AMPOULES H.S + 1 NEON.

TABLE 4.6 – Exemples d'emploi du " je " ou du " nous " dans les FA

d'usage du « je » est particulièrement atypique au sein des données. En effet, nous n'observons habituellement pas de mention d'un nom de personne ou d'organisation comme c'est le cas ici présent, et nous observons une multiplicité de l'emploi de pronoms. Cette FA présente un caractère subjectif peu habituel dans ces rapports, de même que pour le dernier exemple. L'effacement de l'énonciateur est un procédé courant dans les structures formelles. Il peut aussi venir en réponse aux composants d'une culture de sécurité efficace de Reason (1997), présentés au chapitre 1, que sont la *reporting culture* et la *just culture*. La première fait état de la nécessité d'une liberté à rapporter les erreurs, et la seconde au fait d'avoir un climat de confiance dans ce processus, plutôt qu'un recours au blâme. Le fait de se passer du « je » renforce l'impression de distance entre le rapport, et l'opérateur qui le produit. Pour les FA, une procédure cadre la rédaction des rapports. Dans celle-ci, il est indiqué de ne pas utiliser le style personnel. Cet aspect est également présent dans la formation des opérateurs et des inspecteurs qualité.

4.2.3 Un corpus bruité

Une autre des caractéristiques principales de ce corpus est la présence systématique de bruit textuel dans les données. Dans cette section, nous observons quels sont les phénomènes

de bruit que nous retrouvons le corpus.

Al Sharou *et al.* (2021) proposent une taxonomie (cf. Figure 4.5) des différents phénomènes de bruits que nous pouvons retrouver dans des données textuelles. Cette taxonomie propose un premier niveau de classification avec sept classes :

1. les erreurs grammaticales ;
2. l'orthographe ;
3. la répétition de ponctuation (par exemple « ! ! » pour un effet d'emphase) ;
4. les disfluences dans les données transcrites ;
5. le jargon d'internet ;
6. les URLS, liens et balises ;
7. le code-switching.

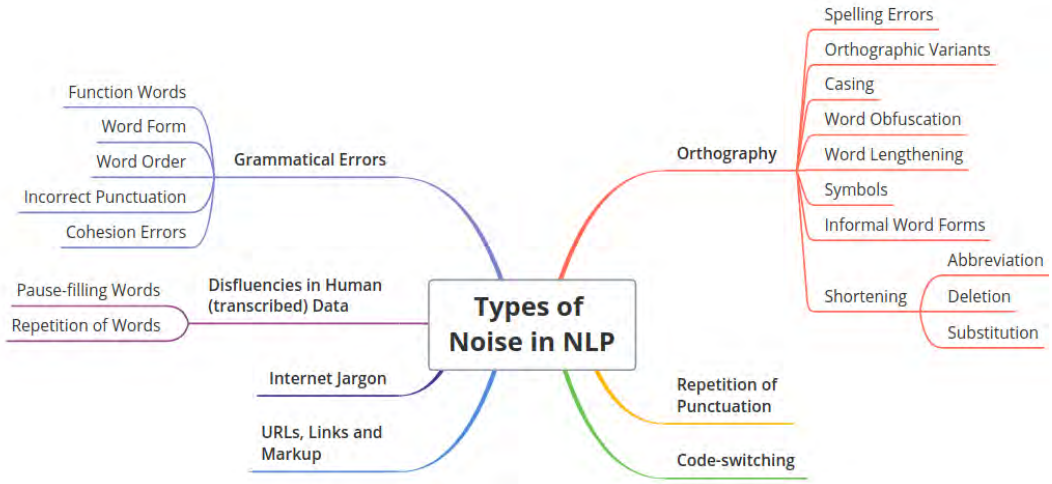


FIGURE 4.5 – Taxonomie des phénomènes de bruit de Al Sharou *et al.* (2021)

Parmi ces différentes classes, seules les deux premières sont pertinentes à notre corpus puisque nous ne travaillons pas avec des données du *web*, ni des données orales, ni des données impliquant différentes langues. Ces deux catégories, les erreurs grammaticales et orthographiques, comprennent plusieurs phénomènes que nous retrouvons dans nos données.

Le phénomène qui saute aux yeux dès le premier abord est la présence quasi-systématique de majuscules au sein du corpus. En effet, le texte est entièrement rédigé en majuscules, à l'exception des unités de mesures (« mm », « b »...) et de quelques sections de texte aléatoires, mais très peu répandues, comme dans l'exemple 4 de la Figure 4.6.

L'autre phénomène que nous retrouvons dans toutes les FA est l'absence d'accents (en rose dans les exemples), et peut être vu comme lié au problème des majuscules. Ceux-ci peuvent entraîner des difficultés de traitement automatique, notamment pour la reconnaissance des temps verbaux (« SPECIFIE / SPÉCIFIÉ » dans le premier exemple), la catégorie grammaticale (« a / à » dans les exemples 1 et 5) ou même le lexème employé (« COTE / CÔTÉ » dans l'exemple 4). Des travaux se sont ainsi focalisés sur la réaccentualisation pour des mots spécialisés comme Zweigenbaum et Grabar (2002) pour le domaine médical.

Parmi les autres phénomènes présents dans le corpus, nous retrouvons particulièrement les symboles (« & », « § », « => »...), les fautes typographiques (quoique relativement rares), et l'absence d'espaces qui peut rendre la compréhension et la segmentation automatique difficiles.

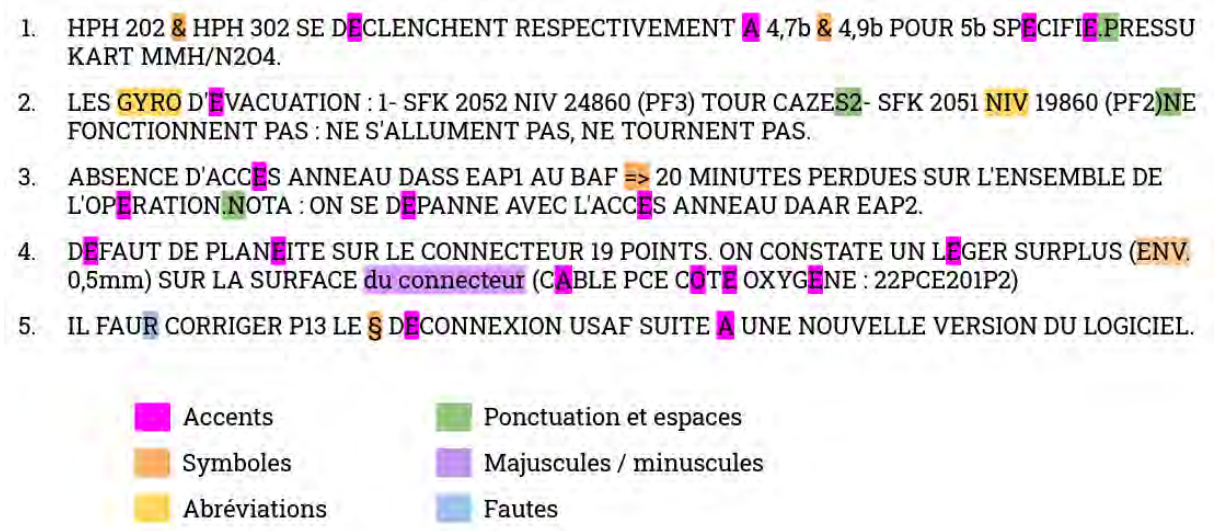


FIGURE 4.6 – Le bruit dans les données

Les textes techniques posent un ensemble de problème pour le TAL. En effet, ils incluent des données de forme non-standard comme une abondance de chiffres, des pièces jointes (photos, compte-rendus... auxquels il est parfois fait référence dans les FA mais auxquels nous n'avons pas accès dans le cadre de notre corpus), des acronymes, des abréviations, des mots mal orthographiés (Andreani *et al.*, 2013). Ces particularités rendent les outils de pré-traitement des données moins performants (Dima *et al.*, 2021) (Brundage *et al.*, 2021). Il en découle le besoin d'adapter le traitement de ces textes, comme nous avons pu le voir dans le chapitre 1 avec notamment la pipeline de TLP (Technical Language Processing) proposé par Dima *et al.* (2021) et Brundage *et al.* (2021). Celle-ci mêle ressources langagières dédiées et le recours à un/des experts.

4.3 Démonstration des impacts du bruit sur des traitements TAL

Al Sharou *et al.* (2021) s'intéressent au bruit dans les données dans une perspective de TAL, et proposent une catégorisation en fonction de ce qu'ils appellent le *harmful noise* et le *useful noise*, soit le « bruit nocif » et le « bruit utile ». Cette différenciation s'effectue principalement en rapport aux performances d'un système automatique. Le « bruit nocif » est défini comme couvrant les phénomènes à supprimer ou à normaliser sous peine de nuire aux performances d'un système automatique. Le « bruit utile » comprend les phénomènes qui améliorent les résultats d'un système automatique.

Pour notre part, dans cette thèse, nous souhaitons aussi opérer une différenciation mais sur un autre critère. Celle-ci se situe entre les marqueurs d'un langage de spécialité et le bruit. Nous avons pu voir dans la section précédente plusieurs caractéristiques propres au corpus et qui s'inscrivent dans son caractère de texte de domaine spécialisé (acronymes, sigles, codes, mesures). Nous choisissons de considérer ces phénomènes comme caractéristiques du corpus inhérents au genre du texte, et potentiellement porteur d'un contenu sémantique. Par contre, le bruit dans les données telles que nous l'avons décrit dans la section précédente est plutôt à caractère nocif. Cependant, dans le cadre d'un traitement automatique, les deux impliquent une répercussion sur les performances du système et distinguer clairement les effets de l'un ou de l'autre de ces phénomènes peut s'avérer délicat. Ainsi, dans cette thèse, nous ne considérons pas les marques de spécialités comme étant du « bruit ». Toutefois, dans cette section, nous démontrons l'impact du bruit et des éléments du domaine sur les performances de traitement automatique, sans pouvoir traiter indépendamment les deux.

Dans Al Sharou *et al.* (2021), les auteurs évaluent l'impact de différents phénomènes de bruit dans trois corpus de réseaux sociaux différents, sur trois tâches de classifications. Pour chacune, ils effectuent une normalisation, une suppression ou aucune modification sur les sept phénomènes de bruit que sont : l'utilisation des majuscules, le *tagging* d'adresse (« @ »), les *hashtags*, les *emojis*, la ponctuation et le *code-switching*. À l'issue de ces expérimentations, les auteurs déterminent que le traitement du bruit dans les données est dépendant de la tâche. Ainsi, le *preprocessing* des données doit être adapté en fonction de la tâche plutôt que d'utiliser une *pipeline* standard systématiquement.

Parallèlement, (Brundage *et al.*, 2021) démontrent dans quelle mesure les outils de TAL entraînés sur des données standards ne sont pas adaptés aux langues de spécialités. Leurs données d'entraînement sont trop différentes de celles que nous retrouvons dans des domaines spécialisés (dans leur cas celui de la maintenance industrielle). De ce fait, les différentes

étapes de pré-traitement des données s'avèrent non performantes.

De façon similaire à Brundage *et al.* (2021), nous proposons d'observer l'impact d'un prétraitement automatique sur la description d'une anomalie trouvée dans une FA. Nous nous concentrons sur les trois étapes classiques que sont : la tokenisation / segmentation, la lemmatisation et l'étiquetage morpho-syntaxique (*Part-Of-Speech tagging*). Pour ce faire, nous avons appliqué l'outil *Stanza*[47] (Qi *et al.*, 2020) sur la phrase 1 des exemples donnés en Figure 4.6. Les résultats de ces trois traitements sont visibles en Table 4.7, les erreurs effectuées par l'outil sont indiquées dans les cases orangées.

HPH 202 & HPH 302 SE DECLENCHENT RESPECTIVEMENT A 4,7b & 4,9b POUR 5b SPECIFIE.PRESSU KART MMH/N2O4.			
Index	**Tokenization**	**Lemmatisation**	**POS tagging**
1	HPH	HPH	PROPN
2	202	202	NUM
3	&	&	SYM
4	HPH	HPH	PROPN
5	302	302	NUM
6	SE	soi	PRON
7	DECLENCHENT	déclércer	VERB
8	RESPECTIVEMENT	respéctifement	VERB
9	A	A	NOUN
10	4,7b	4,7b	NOUN
11	&	&	SYM
12	4,9b	4,9b	NOUN
13	POUR	pour	NOUN
14	5b	5b	NOUN
15	SPECIFIE.PRESSU	spécifienbre	NOUN
16	KART	KART	NOUN
17	MMH/	MMH/	NOUN
18	N2O4	N2O4	NOUN
19	.	.	PUNCT

TABLE 4.7 – Pré-traitement d'une description d'anomalie

47. Version 1.8.2, modules : *tokenize, mwt, pos, lemma* pour le français

Concernant la tokenisation, nous pouvons observer des erreurs lignes 15 et 17. Dans le premier cas, le problème provient de l'absence d'un espace après le point dans la séquence « SPECIFIE.PRESSU ». À cela s'ajoute la présence de l'abréviation « PRESSU » ce qui ne permet pas au système d'identifier deux mots distincts. Le second cas est « MMH/N204 » dont la séparation est effectuée après le « / ».

Au niveau de la lemmatisation, nous remarquons trois erreurs. La première concerne « DECLENCHENT » et peut provenir de l'absence d'accent. De même, le token « A » ligne 9 devrait correspondre au lemme « à ». L'absence d'accent dans le corpus entraîne cette confusion. La seconde concerne « RESPECTIVEMENT » qui ne présente pas de faute particulière. Cela laisse supposer que le fait d'avoir un texte entièrement en majuscules empêche l'outil de reconnaître correctement le token. Qi *et al.* (2020) précisent avoir utilisé un système neuronal, en plus d'un système à base de dictionnaire pour la lemmatisation. Ces résultats peuvent donc aussi être le produit d'hallucinations. Nous avons relancé l'outil sur la phrase, mais en utilisant des minuscules pour le terme « respectivement » (aucune autre modification n'a été appliquée sur le reste de la phrase). Cette fois, la lemmatisation est correctement effectuée. L'utilisation des minuscules sur tout le corpus n'est en revanche pas envisageable. D'une part, l'absence de majuscule en début de phrase empêche l'outil de segmenter les phrases correctement. D'autre part, nous avons effectué ce test sur cette même phrase d'exemple. Si la lemmatisation s'en voit améliorée, l'étiquetage morpho-syntaxique ne suit pas. Nous observons une multiplication de l'emploi de la catégorie « X » pour désigner les éléments « autres », que l'outil n'arrive pas à catégoriser. Le dernier cas concerne à nouveau la séquence « SPECIFIE.PRESSU ». Étant donné que la tokenisation est incorrecte, la lemmatisation ne l'est pas plus et produit le lemme « spécifienbre ».

L'étiquetage morpho-syntaxique compte cinq erreurs. Nous y retrouvons « RESPECTIVEMENT », « A » et « SPECIFIE.PRESSU » qui ont déjà posé un problème dans les étapes précédentes. Ainsi, « respectivement » est catégorisé comme un verbe, la préposition « à » devient un nom, de même que la séquence « SPECIFIE.PRESSU ». Par ailleurs le code « MMH/N204 » est catégorisé comme nom. Selon nous, la catégorie « nom propre » aurait été plus appropriée.

Nous avons appliqué l'outil sur un échantillon plus conséquent des données afin de quantifier les erreurs. Le taux d'erreur obtenu pour l'étiquetage morpho-syntaxique est à hauteur de 20%.

Dans la Figure 4.7, nous avons repris le deuxième exemple de la Figure 4.6 et appliqué deux outils de correction automatique. Les caractères surlignés dans le premier rapport sont les éléments que nous considérons porteurs d'un des phénomènes de bruit préalablement

FIGURE 4.7 – Essais de correction automatique

identifié. Le deuxième rapport est le même mais une fois corrigé par *spellChecker*[48], un outil de correction automatique. En rouge sont surlignés tous les cas où l'outil n'a, soit pas corrigé les erreurs dans le texte, soit ajouté des erreurs. En effet, nous pouvons voir trois occurrences supplémentaires pour lesquelles l'outil a supprimé des espaces qui auraient dû rester présents. Dans le troisième exemple, nous avons utilisé l'outil *Language Tool* sur le même rapport[49]. La couleur rouge est utilisée de façon similaire, et la couleur bleue représente les cas de correction corrects. En deux occasions, *Language Tool* se révèle efficace. En revanche, il modifie le code « SFK » par « SFR », l'abréviation « NIV » par « NI », et il ajoute un « S » à » GYRO » (ce dernier phénomène pourrait être considéré comme correct). Pour les deux outils, l'utilisation de majuscules pour tout le texte n'a pas non plus été modifiée. L'utilisation d'outils de correction automatique « classiques » montre donc rapidement ses limites face à l'abondance des phénomènes présents dans les FA. Dans le chapitre 7, nous explorons ainsi l'utilisation de LLMs pour normaliser les données. La méthode utilisée vise à adresser les différents phénomènes de bruit tout en limitant l'impact sur les marques de spécialité.

Ainsi, le prétraitement de données telles que les nôtres produit des résultats peu satisfaisants et loin des performances actuelles de l'état de l'art. Le manque d'adaptation des outils pour des données techniques et très bruitées s'avère un obstacle qu'il convient de contourner. Brundage *et al.* (2021) et Dima *et al.* (2021) proposent ainsi la création d'un *Technical*

48. https://pypi.org/project/pyspellchecker/
49. *Language Tool* est un correcteur d'orthographe utilisable sur divers outils de traitement de texte, comme extension de navigateur, mais aussi sous la forme d'une librairie python que nous avons utilisée dans cet exemple. https://languagetool.org/fr

Language Processing, une approche qui consiste en la création et l'adaptation d'outils et de ressources pour les textes techniques. Bikaun *et al.* (2024a) s'inscrivent dans cette lignée avec la publication en accès libre du corpus *MaintNorm*, composé de textes de maintenance industrielle et entièrement annoté par des experts. Ce corpus a aussi été nettoyé afin d'en corriger les erreurs typographiques et la mauvaise utilisation des majuscules. Les abréviations ont été modifiées pour faire réapparaître le mot d'origine, et des espaces ont été ajoutés dans les cas de mauvaise concaténation afin d'améliorer la tokenisation (« 250hr » devient « 250 hours »). Toutefois, cette ressource est pour la langue anglaise et notre corpus est en français. En outre, la création d'une ressource de cette taille nécessiterait un temps conséquent, y compris pour les experts, et n'est pas l'objectif de cette thèse. Les anomalies du corpus des FA ne sont aussi pas nécessairement du même type, et les erreurs que nous constatons dans le corpus peuvent venir de sources autres que de simples erreurs typographiques ou des abréviations. Les majuscules et l'absence d'accents peuvent provenir du format des données, ou de la modalité d'extraction initiale du tableur depuis la base de données. Ainsi, nous proposons d'explorer plutôt comment appréhender ce corpus tout en contournant les écueils exposés dans ces deux dernières sections. Nous posons la question de la possibilité de se passer de ces étapes traditionnelles de prétraitement, que nous choisissons de ne pas employer dans un premier temps dans l'étude du corpus.

À ces écueils s'ajoute aussi le problème de la compréhension des données du corpus qui, comme vu précédemment, présentent une certaine opacité pour un public non expert. Aussi, afin d'appréhender ce corpus, il nous est apparu nécessaire de trouver un point d'ancrage, de fortes récurrences partagées, dans le but de mettre de l'ordre dans des données de première apparence disparates. Cette mise en ordre passe par une typologie basée sur l'expression d'un problème dans la description de l'anomalie. En effet, les calculs de fréquence observés dans ce chapitre ont révélé la récurrence de termes faisant référence à la notion de problème : « défaut », « perte », « hs »... Nous avons choisi d'explorer cette piste afin de construire une typologie d'expression d'un problème technique. Cette typologie fait l'objet du chapitre suivant.

Chapitre 5

Comment exprimer un problème technique : une typologie

Sommaire

5.1 Introduction : Pourquoi une typologie ?

Le corpus des FA comprend 35 242 rapports. Malgré leur brièveté, une analyse linguistique manuelle de chacun de ses rapports demanderait un temps considérable. Par ailleurs, nous avons pu voir dans le chapitre 4 qu'un prétraitement usuel avec des outils de TAL traditionnels montre rapidement des limites lorsque confronté à des textes techniques et bruités.

Face à ces limites, nous avons choisi de concentrer notre étude sur un des champs du corpus : « Description de l'anomalie ». C'est au sein de ce champ qu'est décrite l'anomalie en question, et c'est ici qu'est exposée une partie des informations nécessaires à la formalisation. Ce champ possède aussi l'avantage d'être systématiquement renseigné et de façon la plus « prolifique » (pour rappel, avec une moyenne 19,3 mots par rapport). Pour appréhender le contenu de ce champ, nous avons aussi choisi de concentrer notre étude sur les éléments récurrents, et compréhensibles sans savoir expert, de ces rapports que sont les termes évocateurs du problème rencontré. Les analyses de fréquence des *tokens* et *bigrammes* du chapitre 4 ont déjà permis de faire ressortir certains de ces termes (« défaut », « perte », « fuite »...) et leur récurrence. Dans ce chapitre, nous commencerons par explorer plus avant cet aspect du corpus.

Ces démarches ont été réalisées en vue d'établir une typologie d'expression d'un problème technique basée sur ces indices lexicaux. Les typologies dans le domaine des REX sont choses communes, mais visent généralement d'autres types de classification. Nous avons vu notamment dans le chapitre 1 les travaux de Kurela *et al.* (2020) sur l'utilisation du TAL pour la classification des REX en niveaux de gravité. D'autres classifications peuvent concerner les conséquences des anomalies ou les éléments du système impactés. Dans nos travaux, nous choisissons d'avoir une vision plus générale du problème et de nous focaliser sur la façon dont il est exprimé. Par ailleurs, nous avons vu lors de la présentation de TRIZ au chapitre 2 que l'abstraction du problème fait que les différents éléments sont généralisés. Ainsi, des dysfonctionnements qui impliquent les mêmes types d'éléments et d'interaction à un niveau abstrait peuvent être représenté par le même *vépole*, même si les éléments dans le domaine industriel ne sont pas similaires et implique un niveau de gravité différent. Il nous est cependant nécessaire de faire le tri dans les différentes anomalies qui constituent notre corpus, et nous nous basons pour cela sur leur forme afin de nous mener vers la modélisation en *vépole*. Notre démarche est donc inductive (en. *corpus-driven*).

Dans cette typologie, chaque catégorie de cette typologie représente un schéma d'expression d'un problème. Ces schémas d'expression correspondent à des *frames* issus de *Framenet*. Nous aurons l'occasion de présenter dans ce chapitre les *frame* auquel sont associés chaque catégorie, mais qui n'ont pas été envisagés pour ce type de corpus et peuvent donc présenter

certaines limites.

5.2 Un corpus de problèmes et des problèmes de corpus

5.2.1 Quelques fréquences

Dans cette section, nous présentons les différentes méthodes que nous avons utilisées afin d'explorer le corpus. Nous nous focalisons principalement sur la notion de problème et comment il se manifeste d'un point de vue lexical. La première étape a été d'effectuer une analyse du lexique en rapport avec le sens de problème ou d'anomalie. Pour ce faire, nous avons effectué un calcul de fréquences afin de relever les occurrences pertinentes de tokens, bigrammes et trigrammes. Ce relevé est présenté dans la Table 5.1. Nous avons choisi d'inclure les bigrammes et trigrammes afin de repérer les expressions composées évocatrices du problème pour lequel le token seul n'en porte pas le sens comme « ne fonctionne pas » ou « hors specification »[50]. Nous avons utilisé l'outil AntConc[51] et nous avons fixé un seuil minimal de 50 occurrences. Cet outil nous a aussi permis d'aller vérifier certaines occurrences en contexte dans le cas de doute sur leur utilisation. À ce stade de l'analyse, nous avons considéré les occurrences dans une acception large ce qui explique la présence dans cette liste d'expressions telles que « presence d'eau », « n'est pas ». Le texte que nous étudions est essentiellement composé de descriptions d'anomalies, et cela, dans un style minimaliste. Nous faisons donc ce relevé en prenant en compte cet aspect et en incluant tout segment qui semble évoquer que *quelque chose ne va pas*. À partir de ce relevé, il est déjà possible d'effectuer des regroupements autour de notions plus précises :

— l'impossibilité : *impossibilite*, *impossible* ;
— l'absence : *perte*, *absence*, *manque*, *pas de* ;
— le non fonctionnement : *hs*, *ne fonctionne pas*, *a l'arret* ;
— le non respect d'un standard : *attendu*, *ecart*, *hors specification*, *pour une spec*, *hors spec sur*.

Parmi les autres séquences les plus fréquentes que nous ne pouvons pas regrouper dans les catégories ci-dessus, nous pouvons noter les suivants : *message d'erreur*, *presence (de/d'/d'eau)*, *fuite (sur)* et *defaut*. De ces listes de fréquences, nous pouvons noter l'absence de termes tels que *casse*, *abime* ou *dechire*. Ils apparaissent pourtant fréquemment lors d'une analyse manuelle des données.

50. Nous choisissons de garder l'absence d'accent lorsque nous faisons références à des exemples issus des données et des sorties d'outils dans le texte.
51. https://www.laurenceanthony.net/software/antconc/

Token	Fréq.	Bigramme	Fréq.	Trigramme	Fréq.
defaut	2346	pas de	1155	n est pas	679
hors	2128	n est	870	au lieu de	435
perte	1998	impossibilite de	926	ne sont pas	372
fuite	1654	absence de	581	ne fonctionne pas	296
hs	1253	hors spec	541	n'a pas	271
impossibilité	1146	perte de	536	a l'arret	167
absence	1118	au lieu	516	pour une spec	167
impossible	1027	impossible de	500	présence d'un	165
erreur	914	présence d	499	perte de la	148
attendu	897	fuite sur	482	présence d'eau	139
arret	791	perte du	469	perte de l	112
manque	679	présence de	447	presence d'une	109
sans	574	hors specification	438	impossibilité de monter	106
ecart	535	ne fonctionne	383	hors spec sur	100
problème	487	fonctionne pas	296	message d'erreur	95

TABLE 5.1 – Indices lexicaux d'un problème les plus fréquents

5.2.2 Topic modeling

Afin d'approfondir les observations effectuées avec les listes de fréquences, nous avons choisi d'effectuer un *clustering* en utilisant la technique du *Topic Modeling* (Blei, 2012). Le *Topic modeling* est une méthode probabiliste de représentation de documents dans un espace vectoriel. Les dimensions de l'espace vectoriel correspondent à un document ou à un *topic*/ thématique. Les *topics* sont calculés en fonction de la distribution des mots dans le document, sans information supplémentaire. Ils sont ainsi censés représenter les thématiques sous-jacentes de la collection de documents. Cette méthode peut être utilisée avec différentes techniques et celle dont nous avons fait usage est la LDA (*Latent Dirichlet Allocation*) (Blei *et al.*, 2003).

L'utilisation du *Topic modeling* dans le cadre de cette étude ne vise pas une représentation parfaite des *topics*, ni une catégorisation du corpus ou autre. Nous l'utilisons ici à des fins d'exploration de corpus. Nous cherchons principalement à observer si les indices lexicaux identifiés précédemment se démarquent aussi par ce biais, si des regroupements autour du lexique des problèmes apparaissent. De ce fait, nous ne cherchons pas à faire une évaluation quantitative des résultats obtenus.

Comme lors de l'étape précédente, nous nous focalisons sur le champ « Description de l'anomalie » du corpus. Le seul prétraitement effectué est une tokenization. Celle-ci découpe les mots sur la base d'une suite de séquence alphanumérique [52]. Nous avons fait le choix de conserver les mots vides puisque l'aspect technique du corpus et l'utilisation des majuscules pour tout le texte empêche de distinguer des mots vides d'éventuels acronymes (dans le corpus, « LA », « SI »... pourraient faire référence à des termes techniques). Pour chaque anomalie, nous avons récupéré le champ *Description de l'anomalie*. Celui-ci équivaut à un document, ce qui fait donc un total de 35242 documents. Pour rappel, lors de l'étape précédente, nous avons pu observer quatre catégories qui regroupent plusieurs termes parmi les plus fréquents (l'impossibilité, l'absence, le non-fonctionnement, le non-respect d'un standard), quatre indices de l'expression d'un problème technique isolés (*message d'erreur, presence d', fuite, defaut*) et une catégorie qui ne semble pas apparaître (*casse, abime...*), pour un total de neuf catégories. Le nombre de *topics* que le modèle fait émerger est au choix de l'utilisateur. En nous basant sur ces observations, nous avons décidé d'observer les *clusters* produits dans le cas de 9, 10 et 11 *topics*. Les résultats présentés ci-dessous correspondent à 11 *topics*.

En Table 5.2, nous pouvons observer, pour chacun des 11 *topics*, les 30 mots les plus saillants. La saillance d'un mot correspond au degré selon lequel un mot est informatif pour déterminer un *topic*, comparé à un mot quelconque. Si un mot apparaît dans tous les *topics*, le mot n'est pas *distinctif* du *topic* (Chuang *et al.*, 2012).

Dans les résultats de la Table 5.2, nous retrouvons des mots en rapport avec la notion de problème (en gras), dont une majeure partie était déjà présente dans les listes de fréquences présentées précédemment. Ainsi, dans le premier *topic*, nous retrouvons les mots *alarme, erreur, perte, arret* et *defaut*. Cependant, ces mots ne semblent pas pouvoir être regroupés dans une catégorie commune.

Dans le topic 2, nous pouvons observer deux des catégories préalablement identifiées, *l'impossibilite* et *l'absence* avec, respectivement, les mots *impossible* et *impossibilite*, et *manque* et *absence*.

Dans le topic 3, les deux mots identifiés sont *presence* et *casse* et ne paraissent pas à l'abord présenter de points communs. Une analyse manuelle des données révèlera cependant l'existence de séquence telles que *presence de debris/d'insectes/de poussiere*... Ces séquences, tout comme *casse* font référence à une dégradation physique.

Le topic 4 agrège les mots *ecart* et *derive* que nous ne retrouvons pas dans les listes de fréquences et qui semble cooccurrer avec des unités de mesures et mots en rapport avec des

52. Nous avons utilisé la librairie NLTK et le module *RegexpTokenizer* en nous basant sur le symbole *\w* pour effectuer le découpage

Topic	Mots les plus saillants
1	la, lors, table, **defaut**, le, **perte**, fa, apres, suite, de, plus, **erreur**, issue, **alarme**, mise, **arret**, ccs, alimentation, commande, tension, cr, passage, etat, alim, cco, message, zl, camera, fermeture, declenchement
2	la, le, les, entre, **impossible**, po, cote, **absence**, vis, pas, **impossibilite**, pc, en, place, sont, cco1, lieu, connecteur, fixation, **manque**, plan, procedure, prise, cryo, partie, liaison, de, axe, vers, sur
3	niveau, local, eau, cable, cote, **presence**, carneau, bil, porte, epc, protection, bras, massif, acces, eap, mat, coffret, sud, cables, entree, zl3, support, deluge, nord, **casse**, dessus, tres, passif, zone, zse
4	pression, sortie, esc, mb, module, capteur, environ, sol, reseau, bar, flexible, cdl3, helium, bars, raccord, temperature, air, connexion, **ecart**, niv, etancheite, lo2, debit, test, service, **derive**, dsm, detente, ventilation, delta
5	non, baie, ete, t2, sn, avant, t1, fonctionne, constate, operation, systeme, carte, demarrage, ela2, apparition, conforme, panneau, fin, retour, charge, led, utilisation, durant, rack, situe, **deterioree**, conteneur, relais, levage, bus
6	ligne, **fuite**, vanne, tour, cazes, ouverture, vannes, ouverte, lignes, azote, balayage, salle, importante, event, troncon, analyse, purge, fermee, vide, verin, panne, synoptique, projet, clapet, interne, decale, courant, fae, labo, manoeuvre
7	**hors**, **spec**, mesure, valeur, une, **attendu**, voie, mesures, **specification**, max, mm, **attendue**, ok, ohm, tolerance, faux, centrale, bagottement, valeurs, compteur, bruit, ohms, incendie, vrai, controle, maintenance, hc, rapport, entre, fusible
8	baf, **hs**, he, pompe, frontal, sonde, fonctionnement, temps, tm, api, clim, remontee, circuit, enregistrement, gaz, telephone, transbordeur, lancement, indique, cellule, affichage, video, cameras, lanceur, bse, sdo, depassement, manometre, ms, roulage
9	voir, photos, joint, lbs, photo, druck, sec, soupape, ci, annexe, flexibles, caissons, echelle, onduleur, secours, evacuation, nombreux, iii, locaux, jointe, important, ventilateur, cm, met, cat, tf, boucle, pf, als, **arrete**
10	lh2, lox, caisson, stockage, o2, seuil, bas, communication, thermique, haut, pac, n2, disque, h2, usaf, ppm, ouvert, rupture, piscine, hygrometre, boite, matelas, appareil, reservoir, sdc, refoulement, galerie, gromelle, exterieure, retrouve
11	**probleme**, fbf, palette, actif, eap2, rs, crf, medium, eap1, sas, presse, ombilical, pdp, cro, exterieur, ute, ues, etoupe, rc, messages, demarre, inferieur, checksum, bielle, sup, baies, equipement, colliers, bielles, process

TABLE 5.2 – Résultats du Topic modeling

gaz ou des liquides (*heliums, bars, pression, air, etancheite, lo2, debit, ventilation*).

Les topics 5, 6, 8, et 9 ne sont composés chacun que d'un seul mot avec respectivement *deterioree, fuite, hs et arrete*. Nous retrouvons les catégories *fuite* (avec le mot *fuite*), et *non fonctionnement* avec *hs* et *arrete* que nous avions déjà identifiés avec les fréquences.

Le topic 7 regroupe plusieurs mots en rapport avec la catégorie *non-respect d'un standard*. En effet, il s'y trouve les mots *hors, spec, attendu(e)* et *specification*.

Le topic 10 ne contient aucun indice lexical que nous recherchons.

Enfin, le topic 11 contient uniquement le mot *probleme*, qui est trop générique par rapport à notre objectif de typologie d'expression d'un problème technique.

Nous retrouvons donc ainsi les premières catégories préalablement identifiées avec l'analyse de fréquences dans les topics 2 et 7. Nous identifions aussi des catégories et des mots supplémentaires faisant références à la notion de problème.

5.2.3 Word2Vec

De façon similaire à notre utilisation du *Topic modeling*, nous avons aussi utilisé *Word2Vec* afin de poursuivre notre exploration du corpus en rapport avec les indices lexicaux d'un problème technique. Nous faisons appel à cette méthode afin de repérer les différents termes qui sont proches d'un mot donné, soit ici les indices que nous avons déjà repérés avec l'analyse lexicale des fréquences et le *Topic Modeling*. Par ce biais, nous cherchons à observer si plusieurs et quels indices sont utilisés pour exprimer un problème similaire.

Word2Vec (Mikolov *et al.*, 2013) est une méthode de représentation des mots sous la forme d'*embeddings*. Ces *embeddings* (ou *word embeddings*) sont des représentations vectorielles « denses » et avec peu de dimensions, caractéristiques qui diffèrent des vecteurs « *sparse* » (*one-hot* ou « sacs-de-mots ») qui sont de taille plus conséquente et peuvent inclure des dimensions vides. De part leurs faibles dimensions et leur densité, les *embeddings* se montrent plus efficaces et plus performants (Jurafsky et Martin, 2024). Ces *embeddings* sont aussi qualifiés de « statiques », un *embedding* correspond à un mot du vocabulaire des données. La manière dont le vecteur est appris fonctionne sur un système de paires. Lors de l'entraînement, pour chaque mot, le modèle prédit dans quelle mesure le mot risque d'apparaître avec un autre mot du contexte (Miletic, 2022).

Pour cette tâche, nous avons donc choisi d'utiliser *Word2Vec* afin d'observer les mots les plus proches de ceux identifiés avec le *Topic modeling* et à travers les calculs de fréquence. Nous avons observés les mots les plus fréquents et présentons dans la Table 5.3, pour chaque mot de la première colonne, les 15 mots les plus proches. Ces résultats nous permettent de voir quels sont les termes proches des indices lexicaux de l'expression d'un problème que nous avons identifiés précédemment. Nous avons choisi de ne pas présenter tous les mots préalablement identifiés, mais plutôt un ou deux mots par « catégories » (séparées dans le tableau par une ligne vide).

Les résultats montrent que, pour le mot *impossible*, nous retrouvons *impossibilité* déjà vu auparavant, mais aussi *difficulte, difficile* et *empeche*. Cela suggère que, pour l'élaboration de la typologie, nous pouvons regrouper les cas ou une *impossibilité* est exprimée avec les cas ou une *difficulté* est constatée. De même, pour *hs*, nous voyons apparaître les termes

Mot	Mots les plus proches
impossible	**impossibilite**, afin, facon, necessite, **difficulte**, **difficile**, permettant, permet, maniere, pouvoir, plus, **empeche**, peut, espace, fallu
hs	**defectueux**, **defaillant**, **defectueuse**, **defaillance**, chgt, fusion, scm, acs, convecteur, changee, bruyant, **defaillante**, matrox, tsx, ta
fonctionne	**allume**, **demarre**, **remonte**, ouvre, **fonctionnent**, **remontent**, **fonctionnait**, **repond**, tient, **declenche**, voit, **allument**, **deteriorer**, remet, **descend**
manque	poser, **manquait**, placer, enlever, faudrait, demonter, fallu, fixer, mep, agit, aluminium, plastique, renfort, rap, faudra
absence	presence, **manque**, couleur, aide, jaune, violette, verte, reacheminement, **casser**, gamelle, secu, largage, nez, enver, zipper
specification	**spec**, **specif**, famille, **tolerance**, gamme, limite, **tolerances**, critere, gabarit, **specifications**, plage, moyenne, criteres, precision, valeur
attendu	obtenu, acquis, **attendus**, **attendue**, vrai, mohm, lsb, ma, kohms, codes, kohm, acquise, ohm, obtient, faux
alarme	detection, appel, astreinte, apparition, bagottement, **alarmes**, saturation, declenche, tu, surveillance, declenchement, **panne**, perturbation, automate, tache
deteriore	**arrache**, **deterioree**, **abime**, **endommage**, central, **casse**, hublot, wel, **plie**, **corrode**, verre, **arrachee**, fortement, eclat, **tordue**
casse	**desserre**, **deteriore**, **endommage**, zjs, etoupe, **abime**, **devisse**, **cassee**, **arrache**, presse, **coince**, **deterioree**, central, coupe, grippe
fuite	bulles, pollution, **fuyard**, **fuites**, helium, kellog, bulle, aval, raccord, mamelon, surpression, bouteille, snoop, chasse, tuyauterie
presence	poche, pluviale, ecoulement, penetration, infiltration, trace, traces, **absence**, couleur, fissure, plastique, flaque, interieur, pluie, morceau

TABLE 5.3 – Word2Vec : mots les plus proches

defectueux et *defaillant*. Pour *deteriore*, plusieurs mots cooccurrents apparaissent autour de la notion de dégradation / détérioration comme *arrache, abime, endommage, corrode...*.

Quelques anomalies peuvent être notées cependant. Par exemple, la présence de *casser* en compagnie d'*absence*. De même, la présence de *panne* dans les proches d'*alarme* plutôt que de *hs*.

Pour le mot *fonctionne*, nous constatons de nombreux verbes supplémentaires proches. De même que *fonctionne* est issu de *ne fonctionne pas*, tel que nous avons pu le voir dans le calcul de fréquence des trigrammes, ces verbes apparaissent régulièrement sous forme de négation. Ainsi, après vérification dans un concordancier, nous pouvons observer les locutions *ne s'allume pas*, *ne demarre pas / plus*, *n'a pas demarre*...

Lors des étapes précédentes, nous avons pu constater à plusieurs reprises l'apparition du mot *presence*. Nous avons donc ajouté les résultats de *Word2Vec* pour ce mot dans le tableau. Ce mot semble être proche de la notion d'écoulement que nous retrouvons dans la liste et de termes du même domaine : *trace, penetration, infiltration, poche, pluviale, flaque, pluie*. En revanche, le seul mot en lien direct avec la notion de problème présent dans cette liste est son antonyme *absence*.

5.3 Une typologie d'expressions d'un problème technique

Les différentes explorations du champ « Description de l'anomalie » en rapport avec la notion de problème nous ont permis d'en déduire une typologie. Celle-ci est basée sur la façon dont sont exprimés les problèmes, c'est-à-dire les indices lexicaux préalablement repérés. Ces derniers deviennent des marqueurs d'une classe d'expression d'un problème. L'élaboration de la typologie permet de donner une première aide pour la classification des données. Dans un second rôle, elle sert aussi de guide dans l'identification des éléments que nous souhaitons extraire des données afin de les modéliser. Notre hypothèse est que, pour chaque type, il existe un ou des *patterns* d'expression d'un problème technique.

Dans cette section, nous commençons par présenter la typologie dans sa version stable et définitive, en présentant chacune des catégories. Dans un second temps, nous revenons sur le procédé de validation de cette typologie, de ses différentes versions et de certaines questions que nous nous sommes posées. Enfin, nous mettons en lien les différentes catégories avec les *frames* issus de *FrameNet* qui nous semblent correspondre aux situations décrites.

5.3.1 Présentation de la typologie

La typologie d'expressions d'un problème technique est composée de neuf classes visibles dans la Table 5.4. Elle est aussi dégressive, c'est-à-dire que les différents types ne sont pas tous fournisseurs de la même quantité d'information. Par quantité d'information, nous faisons référence ici aux informations relatives au problème spécifique directement, et non aux éléments contextuels comme les circonstances spatio-temporelles. Cette démarche s'appuie sur la méthode utilisée dans la constitution des grilles d'annotation de *FrameNet* et

notamment la distinction entre éléments noyaux et participant non-noyaux. Plus la catégorie est haute dans la typologie, plus la situation décrite peut être précise (cela dépend aussi de la quantité de participants noyaux présents dans l'énoncé). Ces degrés restent une indication générale et approximative. Pour chaque classe, nous indiquons aussi quelques exemples de marqueurs, à l'exception de la catégorie *État du monde* qui par définition n'en possède pas, ce que nous détaillons dans la section 5.3.1.9. Nous présentons dans les sections suivantes chacune des catégories avec sa définition et des exemples de rapports.

Degré	**N°**	**Type**	**Exemples de marqueurs**
Degré 1	1	Fuite d'un liquide ou d'un gaz	fuite, fuyard, écoulement
Degré 2	2	Signal qui s'est déclenché	témoin, alerte, alarme
Degré 3	3	Présence d'un obstacle	gêne, empêche, bloque
	4	Dégradation - Usure - Saleté	cassé, marques, corrosion
	5	Élément manquant	absence, sans, manque
Degré 4	6	Configuration hors spécification	hors spécification
Degré 5	7	Dispositif qui ne fonctionne pas	HS, panne
	8	Action difficile ou impossible	impossible, difficulté
Degré 6	9	État du monde	

TABLE 5.4 – Typologie d'expressions d'un problème technique

5.3.1.1 Fuite d'un liquide ou d'un gaz

La catégorie *Fuite d'un liquide ou d'un gaz* correspond à tous les cas où les opérateurs font part de la présence d'une fuite, comme dans les exemples 5.5 Pour cette typologie, nous nous focalisons sur les éléments exprimés dans le texte, et ici sur les marqueurs plus spécifiquement. De ce fait, la présence de tâches d'un liquide ou de flaques, sans marqueur qui est propre à la fuite, n'est pas considérée comme appartenant à cette catégorie. Par exemple, une FA telle que « Présence de taches d'huile sur le sol » n'est pas du type *Fuite d'un liquide ou d'un gaz*. En revanche, « Présence d'une tache d'huile sur le sol à cause d'une fuite » relève bien de la catégorie. Les marqueurs typiques de cette catégorie sont : « fuite », « fuyard », « écoulement ». Cette catégorie est placée en haut d la typologie puisqu'elle renvoie à une situation bien précise et qui implique de nombreux acteurs : le fluide, la source et la destination de la fuite... Nous revenons sur ce cas particulier dans le chapitre 8.

N°	Exemples de fuites
1	SOUPAPE DE L'OBTURATEUR COL DE CHAMBRE FUYARDE.
2	FUITE INTERNE FDV1 RELEVEE A 5.2 10-1 A L'HELITEST.
3	CHUTE DE GOUTTES D'EAU DEPUIS UNE FUITE DANS LE TOIT DU SAS DU BAF SUR LA BRIDE SUPERIEURE DE LA CASE AVEC DES PROJECTIONS DE GOUTTELETTES SUR L'UCAF11 ET LE CONE.DETECTE VERS 18H15 AVEC UNE FLAQUE D'ENVIRON 1M DE DIAMETRE AU DROIT DE LA PORTE V.01 CASE.

TABLE 5.5 – Exemples de Descriptions de la catégorie *Fuite d'un liquide ou d'un gaz*

5.3.1.2 Signal qui s'est déclenché

La classe *Signal qui s'est déclenché* est particulière puisqu'elle n'indique pas la présence d'un problème. En effet, le texte ne décrit pas quel est le problème, uniquement qu'un indicateur d'un problème s'est déclenché. Ce seul déclenchement d'un signal est une anomalie suffisante pour faire l'objet d'un rapport. Nous retrouverons comme marqueurs : "alerte", "témoin", "alarme"... Elle est ainsi peu susceptible d'être pertinente à une modélisation en *vépole* puisque ce dernier traitre des dysfonctionnements techniques et non du signal de la présence d'un dysfonctionnement.

Cette catégorie est au deuxième degré selon la typologie. En effet, nous y retrouvons de nombreuses informations renvoyant au type d'alerte, au message qu'il véhicule, à la façon dont il se manifeste... C'est une catégorie qui peut présenter un haut degré de précision dans son expression.

N°	Exemples
1	MESSAGE D'ERREUR LORS DE LA GENERATION DES TABLES IMAGES. LES COEFFICIENTS DES MESURES M2TK600 ET M2TK604 NE SONT PAS CORRECTS : LES SEGMENTS S1 ET S2 SONT DISJOINTS. CF ANNEXE.
2	ALARME BASSE DES BALAYAGES INTERPLAQUE AI (HPB167/571) ET BALAYAGE J1 HPB569.
3	LORS DU REGLAGE DE LA CAMERA CRYO LOX, L'OPERATEUR VIDEO A DECLENCHE L'ALARME EN DONNANT UN COUP DE TETE DANS UN DETECTEUR. CE PROBLEME SE REPRODUIT A CHAQUE FOIS, COMPTE-TENU DE LA PROXIMITE DE CE DETECTEUR PAR RAPPORT A LA CAMERA. DG CAISSON LOX ESC N° 1.
4	LORS DE LA CONNEXION AU RESEAU AIR RESPIRABLE EN ZSE STOCKAGE N, LES OPERATEURS CEGELEC ONT CONSTATE QUE L'ALARME S'EST DECLENCHEE. LIBELLE : SDC 1004 TAR SEUIL BAS RESEAU AIR.

TABLE 5.6 – Exemples de Descriptions de la catégorie *Signal qui s'est déclenchée*

5.3.1.3 Présence d'un obstacle

La classe *Présence d'un obstacle* présente les cas où il est question d'un obstacle qui empêche de réaliser une action. Les marqueurs sont : « bloque », « gêné », « empêche »... Cette classe est la seule pour laquelle nous n'avons pas évoqué de catégorisation jusqu'à présent. Le mot *empeche* est apparu lors de l'utilisation de *Word2Vec* mais en compagnie de *impossible*. Cette classe est apparue au fur et à mesure de l'analyse manuelle du corpus. Nous aurons cependant l'occasion de voir par la suite qu'elle est moins présente que les autres au sein du corpus.

Cette catégorie est aussi très proche de la classe *Action difficile ou impossible*. Elle est plus précise que cette dernière puisqu'elle donne une indication sur l'élément qui cause l'impossibilité ou la difficulté à accomplir une action. Nous pouvons d'ailleurs observer dans le second exemple Table 5.7 un cas dans lequel se trouve un marqueur des deux types « BLOQUE » et « IMPOSSIBLE ». Le « BATTANT DE LA PORTE D'ACCESS [...] BLOQUE » est la raison pour laquelle l'ouverture est impossible.

Elle est au degré numéro 3 de la typologie puisqu'elle implique la présence de nombreux éléments dans le texte. Elle se place au même statut que les catégories *Dégradation - Usure - Saleté* et *Élément manquant* qui apportent une précision par rapport à la catégorie *Configuration hors spécification*.

N°	Exemples
1	LA VIS DE FIXATION DES BAND STOP EMPECHE LA MISE EN PLACE DE L'OUTILLAGE DE COMPRESSION DE LA LAME RESSORT.
2	2EME BATTANT DE LA PORTE D'ACCES EN PFCU BAF BLOQUE. OUVERTURE IMPOSSIBLE PAR ACTION SUR BARRE ANTI-PANIQUE.

TABLE 5.7 – Exemples de Descriptions de la catégorie *Présence d'un obstacle*

5.3.1.4 Dégradation - Usure - Saleté

La classe *Dégradation - Usure - Saleté* concerne les textes dans lesquels il est fait état d'une dégradation d'un équipement de quelque nature. Cela recouvre tous les problèmes pour lesquels un élément est « cassé », « déchiré », mais aussi la présence de « poussière », d'« insectes » ou de « corrosion ». Cette classe agrège de nombreux marqueurs. Des exemples de descriptions sont présents en Table 5.8.

La classe *Dégradation - Usure - Saleté* est proche de la classe *Configuration hors specification* telle que nous la concevons et que nous décrivons dans la section 5.3.1.6. Elle apporte une précision sur l'anomalie décrite.

N°	Exemples
1	LA PROTECTION THERMIQUE DE LA LDT N°2 EST ENDOMMAGEE A 10cm DU RACCORDEMENT RMV. (IDEM 504 LIGNE 01)
2	BRAS LH2 :HOUSSE VENTILATION ISS BRULLE ET DECHIREE.FLEXIBLE VENTILIATION ISS BRULE.

TABLE 5.8 – Exemples de Descriptions de la catégorie *Dégradation - Usure - Saleté*

5.3.1.5 Élément manquant

La classe *Élément manquant* concerne tous les cas où un objet, dispositif, document ou tout autre élément est noté absent. Les marqueurs sont de l'ordre de : « sans », « absence »... De même que pour la catégorie *Dégradation - Usure - Saleté*, cette classe est similaire à la catégorie *Configuration Hors Spécification*, mais apporte un niveau de précision supplémentaire dans les marqueurs utilisés.

N°	Exemples
1	ABSENCE CHAINETTE "GARDE CORPS" AU NIVEAU DE LA NICHE NIV 6 TOUR CAZE
2	LES CLES DES CADENAS POUR VERROUILLER LES DISJONCTEURS DANS LA BAIE AMEF SONT INTROUVABLES.

TABLE 5.9 – Exemples de Descriptions de la catégorie *Élément manquant*

5.3.1.6 Configuration hors spécification

Pour la catégorie *Configuration hors spécification*, deux aspects sont pris en compte. Le premier correspond à tous les cas où une mesure dépasse ou n'atteint pas un standard, comme dans l'exemple 1 de la Table 5.10. Nous retrouvons dans cette catégorie le marqueur « ATTENDU » que nous avons pu apercevoir dans les différentes étapes de l'étude du corpus vu précédemment dans ce chapitre. Dans l'exemple 2, nous retrouvons ainsi le marqueur « ATTENDU » pour spécifier le standard auquel la mesure aurait dû correspondre.

Dans un second temps, nous retrouvons dans cette catégorie toute évocation d'un état qui est décrit spécifiquement comme ne correspondant pas à l'état « normal », mais dont le type d'écart n'est pas spécifié, c'est-à-dire que nous n'avons pas de précision sur le fait que l'objet soit dégradé ou manquant. En revanche, il est fait part d'une situation qui est décrite comme incorrecte. Les marqueurs peuvent être très variés : « mal », « n'a pas été », « incorrecte »... Par exemple, dans le cas numéro 3 de la Table 5.10, nous avons le mot « MAL » suivi du participe passé « POSITIONNE » qui constituent un marqueur pour indiquer un écart par

rapport à la situation attendue. L'acception pour la catégorie *Configuration hors spécification* est donc très large et recouvre un grand nombre des descriptions d'anomalies du corpus.

Cette catégorie est donc moins précise que les deux qui la précède. Pour ces raisons, elle est à un degré inférieur. Elle donne cependant un certain nombre d'informations quant à la situation qui se déroule avec notamment la mention d'un standard (parfois implicite).

N°	Exemples
1	LA PRESSION RELATIVE RELEVEE SUR M9PA509 AU COURS DE L'ESSAI EST HORS SPECIFICATION.RELEVE : 370mbSPECIFIE : 1 barNOTA : PRESSION D'ALIMENTATION SOL (DRUCK) : 6,7 bar.
2	TM EAP1M1AC506 (ACCELERATION JAV AXE S) - ACQUIS A 33,3 m.s.2 (217q)- ATTENDU ENTRE +6 ET +14 m.s.2
3	APRES SERRAGE AU COUPLE DE LA VIS DE FIXATION DU RETAINER SET, LA TRESSE DE MASSE NORMALEMENT SERREE PAR LA TETE DE VIS EST LIBRE. L'HELICOIL DOIT ETRE MAL POSITIONNE.

TABLE 5.10 – Exemples de Descriptions de la catégorie *Configuration hors spécification*

5.3.1.7 Dispositif qui ne fonctionne pas

La catégorie *Dispositif qui ne fonctionne pas* concerne les cas dans lesquels le rapport décrit un dispositif, objet ou composant qui ne fonctionne pas ou plus. Il est souvent fait mention d'une panne, d'un défaut, d'un élément défectueux. Seul un constat du problème qui est exposé ici, sans information supplémentaire sur la raison de la panne. Les marqueurs récurrents : « panne » (*cf.* exemple 1 de la Table 5.11), « HS » (Hors Service) (*cf.* exemple 2 de la Table 5.11), « ne fonctionne pas »... Concernant le marqueur « HS », nous pouvons noter que celui-ci peut porter à confusion puisque dans les FA, il peut signifier « Hors Service » ou « Hors Spécification ». La distinction peut se faire grâce au contexte, en général, il sera fait mention de mesures lorsque nous sommes dans un cas correspondant à la catégorie *Configuration hors spécification*. Cependant, en l'absence de ce type d'information, la compréhension de « HS » se fait notamment par les composants énoncés et la connaissance de leur fonction, c'est-à-dire s'ils font référence à des prises de mesures. En ce cas, en l'absence de savoir expert, il peut être difficile de trancher. L'exemple 2 de la Table 5.11 est clair puisqu'un non-expert sait ce qu'est un écran. Dans le rapport « UPR002 **HS** SUR KART MMH (TRANSMETTEUR). UTR001 **HS** SUR KART MMH (CONVERTISSEUR DE TEMPERATURE). UPR001 HORS TOLERANCE IMPOSSIBLE A REGLER.IDEM POUR NPR001 ET NPR201 », les deux premières mentions de « HS » portent à confusion. Dans cet exemple en particulier, seul l'ajout d'une troisième phrase avec la mention de « HORS TOLERANCE » permet de clarifier la catégorisation pour

un non-expert.

Cette catégorie et la suivante sont du même degré. Elles impliquent toutes deux la présence de peu d'information dans l'énoncé. En effet, il est le plus souvent uniquement mention d'un composant et d'un indice lexical indiquant son non-fonctionnement.

N°	Exemples
1	PANNE DE L'OCAM SPARE DURANT DES TESTS DE CAPACITE BATTERIES AU LBC3 DU S5C (EN ATELIER).
2	L'ECRAN PRS2 (PAS 2) EST HS.

TABLE 5.11 – Exemples de Descriptions de la catégorie *Dispositif qui ne fonctionne pas*

5.3.1.8 Action difficile ou 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 » (comme dans l'exemple 1 de la Table 5.12), « dur », « difficulté » (exemple 2)... Nous avions pu voir les marqueurs « impossib(le)(ilité) » apparaître lors des calculs de fréquence et l'utilisation de *Word2Vec* a permis d'effectuer le regroupement avec la notion de « difficulté ».

De même que pour la catégorie précédente, elle implique généralement peu d'information avec uniquement l'action en question et la mention qu'elle est difficile ou impossible à réaliser.

N°	Exemples
1	IMPOSSIBILITE DE MONTER LE CAPTEUR DE VIBRATIONS SUR L'AXE Y(22CVM9037T)
2	1)DIFFICULTE POUR MONTER LE JOINT SOUPLE REF. A511112A0451A02, CELUI-CI TOUCHE L'EPC (BIELLE F3).2)IDEM BIELLE F2. JOINTE).

TABLE 5.12 – Exemples de Descriptions de la catégorie *Action difficile ou impossible*

5.3.1.9 État du monde

La catégorie *État du Monde* se différencie de toutes les précédentes puisqu'elle n'implique pas de marqueurs lexicaux. Lors de l'étude du corpus, nous avons rencontré des cas pour lesquels seule la présence de l'énoncé dans un corpus de REX permet de le considérer comme

énonçant une anomalie. La notion de problème est implicite et réside dans le genre auquel appartient le corpus, sans être marquée lexicalement. Nous pouvons voir trois exemples issus du corpus dans les descriptions d'anomalies 1 à 3 de la Table 5.13. La lecture de ces énoncés hors contexte du corpus et sans connaissance experte ne permet pas de déterminer qu'il est question d'une situation « anormale ».

Dans cette catégorie, nous avons aussi fait le choix d'inclure les énoncés qui font le constat d'un *risque* d'anomalie ou de problème, comme dans l'exemple 4 de la Table 5.13 Puisque l'objet de notre travail est la représentation des anomalies, nous choisissons de ne pas considérer le risque d'une anomalie comme son expression. L'anomalie ne s'est pas (encore) produite et nous considérons que son éventualité dépasse la portée de notre travail. La visée de ce travail est de formaliser un dysfonctionnement technique d'après son énoncé. Or, dans le cas d'un risque, le dysfonctionnement n'est pas directement exposé puisque la situation ne s'est pas produite.

Dans la mesure où aucun marqueur n'existe pour cette catégorie, nous ne pouvons pas repérer d'éléments propres au problème exprimé, la plaçant ainsi au bas de la typologie. Tout événement ou situation quelle qu'elle soit pourrait constituer un rapport de type *État du monde*.

N°	Exemples
1	PRESENCE DES 2 OBTURATEURS AINSI QUE LES FLAMMES CORRESPONDANTES - 5032-0096 - FT518- 5031-0096 - FT518
2	FERMETURE PROVISOIRE DES PORTES BME AXE Y ET Z AVEC 6 VIS A LA DEMANDE VT.
3	CONSTAT D'INSPECTION LANCEUR AVANT TRANSFERT BIL/BAF. VOIR ANNEXE.
4	RISQUE DE FROTTEMENT EN DYNAMIQUE DES CABLES EN SORTIE DE L'INSERT SUR L'ACCESSOIRE ARRIERE DU CONNECTEUR 52PCL32P1 EN AXE TN.

TABLE 5.13 – Exemples de Descriptions de la catégorie *État du Monde*

5.3.1.10 Coexistence des catégories

Dans les exemples que nous avons présentés jusqu'à présent pour les différentes catégories, nous avons choisi des rapports qui font mention d'une anomalie correspondant à la catégorie en question. Or, comme nous avons pu le mentionner dans le chapitre 4, bon nombre de FA font la description de plusieurs anomalies dans un seul rapport. Ces anomalies sont habituellement en rapport, que ce soit par une relation de cause-conséquence, une même localisation... Cependant, cela signifie que plusieurs problèmes de la même catégorie, mais avec des indices différents, peuvent coexister au sein d'un même rapport, mais aussi

qu'il peut y avoir plusieurs marqueurs de catégories différentes. Cette hétérogénéité complique le traitement automatique des rapports, dans la mesure où ils ne peuvent pas être assignés à une seule catégorie de manière univoque. Lors de la présentation de la typologie, nous avons exposé notre objectif d'effectuer, par ce biais, une première classification des rapports en fonction de leur mode d'expression. Cette coexistence de plusieurs catégories vient complexifier cette classification et pose la question de la segmentation du corpus, en-dehors du découpage en FA.

Pour illustrer ces cas de figure, nous présentons des exemples dans la Table 5.14. Dans le premier exemple, nous pouvons observer deux marqueurs, « HS » et « sectionne » renvoyant respectivement aux catégories *Dispositif qui ne fonctionne pas* et *Dégradation - Usure - Saleté*. Nous sommes ici dans une relation cause-conséquence. La dégradation (le fil sectionné) est la cause du non-fonctionnement du dispositif, ici la « sonde ». Il peut être noté que les deux marqueurs n'apparaissent pas dans deux phrases distinctes à proprement parler, le second fait l'objet d'une précision dans des parenthèses seulement.

Dans le deuxième exemple, nous retrouvons trois marqueurs de la catégorie *Fuite d'un liquide ou d'un gaz*. Dans le cas présent, il s'agit d'une énumération de fuites présentes (potentiellement localisées au même endroit). La segmentation de ces trois cas est difficile puisque nous notons l'absence de point ou virgule pour séparer les énoncés. Nous observons en effet deux occurrences de mots collés entre eux là où devrait avoir lieu la segmentation. Pour rappel, nous n'avons pas pu avoir d'information quant à cette absence d'espace que nous retrouvons régulièrement dans les données. Celle-ci peut se produire pendant la saisie des rapports, l'enregistrement dans la base de donnée ou l'extraction depuis cette base...

Dans le troisième exemple, quatre anomalies sont rapportées. Ces différentes anomalies se sont produites durant la même opération d'inspection comme il est indiqué au débit de la description : « INSPECTION EAP AU BEAP »[53]. Nous trouvons ici une énumération segmentée par l'utilisation d'un chiffre suivi d'une parenthèse. Cette fois, trois catégories de la typologie coexistent, *Configuration hors spécification* avec « INCORRECTE » et « NON FIXEES », *Élément manquant* avec « MANQUANTE » et *Dégradation - Usure - Saleté* avec « DECOLLEES ».

Cette coexistence de plusieurs catégories pose la question de la segmentation du corpus. Or, une segmentation « traditionnelle » sur les signes de ponctuation est mise à mal puisque le corpus présente des cas où celle-ci est manquante, comme dans l'exemple 2 de la Table 5.14. De même, dans l'exemple 1, la ponctuation diffère de la norme avec une seconde phrase

53. Les EAP (Étages d'Accélération à Poudre), aussi appelé *booster*, sont les fusées disposées sur les lanceurs. Le BEAP est le banc d'essai où est testé le bon fonctionnement des EAP

N°	Exemples
1	EAP 2 :LA SONDE IDENTIFIEE M3 TC 601 DU FAISCEAU 11WAM972 QUI DOIT ETRE COLLEE SUR LE CODEUR EST **HS**. (FIL METALLIQUE **SECTIONNE**).
2	**FUITE** (environ 10-2 Nm 3/s) sur CTPCH2**MICRO- FUITE** SUR CVIRH (environ 10-3)**MICRO- FUITE** SUR EAPO (environ 10-3)
3	INSPECTION EAP AU BEAP : 1)ETANCHEITE **INCORRECTE** DE LA PROTECTION POLIANE AUTOUR DE LA LIAISON SEPARATION CONE/JAV AXE T. 2)ETANCHEITE AU SCOTCH **MANQUANTE** AUTOUR DE L'OBTURATEUR CAMERA AXE Tn. 3)FLAMMES **NON FIXEES** (DEROULEES) NIVEAU DAS ET JAR. 4)LES 2 ETIQUETTES MPS-CPC SUR MPS, NIVEAU JAV SONT A MOITIE **DECOLLEES**.

TABLE 5.14 – Exemples de rapports avec plusieurs catégories

entièrement entre parenthèse. Ainsi, nous retrouvons ici les problématiques liées au bruit soulevées dans le chapitre 4 sur l'efficacité d'utiliser des outils actuels pour la segmentation étant donné les particularités des données. Cependant, nous repérons diverses relations entre les différents problèmes exposés. Une analyse linguistique pour repérer des patrons de types relationnels entre les différentes catégories pourrait être envisagée.

Pour la suite, cette coexistence des catégories entraîne plusieurs points à prendre en compte pour la suite. En terme d'annotation et de repérage des catégories, cela signifie pouvoir distinguer les catégories au sein d'un même rapport. De plus, notre cible est l'expression du problème. Si plusieurs problèmes sont exposés, il est nécessaire de les cibler indépendamment les uns des autres.

5.3.2 Validation de la typologie

Avant d'arriver à la typologie que nous venons de présenter, celle-ci a fait l'objet de plusieurs versions. En effet, chaque version a été confrontée aux données de façon itérative. De même, la typologie a été présentée à des experts à plusieurs reprises afin d'obtenir leur opinion sur les catégories et leur champ d'actions.

Parmi les catégories qui n'ont pas été retenues, nous pouvons citer : la procédure inadaptée, l'action mal ou pas effectuée, l'intervention d'un élément extérieur ou bien la relance d'un problème déjà rapporté.

Les énoncés avec lesquels il est question d'une procédure inadaptée sont régulièrement présents dans le corpus. Nous pouvons retrouver les exemples suivants : « PROCEDURE RECONDUITE DEPUIS V175 N'EST PAS A JOUR SUITE A LA REUNION MESURE DEBIT. » ou « DEROULEMENT D'UNE OPERATION SUPPLEMENTAIRE SPECIFIQUE AVEC UNE PROCEDURE STANDARD NON DEDIEE A L'OPERATION. LA PROCEDURE DEROULEE

ETAIT ECRITE POUR L'OPERATION STANDARD DE J-5. ». Or, ces cas ne font pas état d'un dysfonctionnement technique à proprement parler. Ces situations ne rejoignent pas les types de problèmes que nous souhaitons modéliser. Par ailleurs, nous n'observons pas d'indices lexicaux qui font écho au type de problème rencontré, seul le terme « procédure(s) » est identifiant de cette situation particulière. Ainsi, les rapports faisant état de procédure sont plutôt à catégoriser en fonction du repérage d'autres marqueurs potentiels. Ainsi, dans le premier exemple, « N'EST PAS À JOUR » est un marqueur de la catégorie *Configuration ors spécification*. De même pour le second exemple avec « NON DEDIEE ». La notion de procédure n'est pas à prendre en compte dans l'attribution d'une catégorie et la catégorisation se fait en rapport au reste de l'anomalie rapportée.

Les situations qui renvoient à une action qui a mal ou pas été effectuée ont été intégrés à la catégorie *Configuration Hors Spécification*. Nous avons pu voir certains exemples dans la section 5.3.1.6 avec le marqueur « MAL POSITIONNE ». Nous retrouvons ici les cas comme « MAL SERRE, PAS EFFECTUE »... En effet, nous considérons qu'il est aussi question ici de situations où un standard n'est pas respecté, même s'il n'est pas question d'un standard de mesure.

Dans le corpus, nous retrouvons aussi certains cas de relance de problèmes déjà rencontrés comme « MAUVAIS POSITIONNEMENT HPR521 SUR SYNO SCO2. (VOIR FA S157/186 DU 05/09/02). ». Ces mentions font suite à un premier rapport. Elles ne sont pas présentes seules dans le rapport. De ce fait, ce que nous visons ici est plutôt l'information sur la situation problématique, décrite avant ou après la mention d'une FA antérieure. Le fait que c'est une reproduction ou la suite d'une autre FA n'est pas nécessaire pour la description et la formalisation du problème. Nous faisons donc abstraction de cette information qui n'entre pas, *a priori*, en conflit avec la typologie ou les marqueurs.

Pour les cas où il est question d'un élément extérieur, nous pouvons mentionner le rapport « BENNE A ORDURE NON ADAPTEE. SOUS EFFET DU VENT LES DECHETS PARTENT DANS LA CUVE DE RETENTION DU BULLEUR NRT 202 ET BOUCHE L'EVACUATION DE LA CUVE. ». Dans cet exemple, et dans les autres cas que nous avons observé, l'élément extérieur (vent, pluie...) est plutôt la cause que la description de l'anomalie. Ils sont pertinents pour la modélisation mais ne sont pas marqueurs du problème du point de vue lexical, tel que nous l'entendons dans la typologie. Dans cet exemple, nous retrouvons donc un marqueur de la catégorie *Configuration hors spécification* avec « NON ADAPTEE » et un marqueur *Présence d'un obstacle* qui est « BOUCHE ».

Ces choix ont aussi été effectués par une confrontation aux données de façon itérative. En effet, à différentes étapes, un échantillon a été sélectionné afin de vérifier, cas par cas, les catégories redondantes, les inadéquations et les situations non représentées dans la typologie. De plus, plusieurs entretiens avec les experts ont été réalisés afin d'effectuer

cette confrontation avec des exemples choisis et d'obtenir un avis sur les éléments manquants.

Une fois que nous avons obtenu une version stable, une première annotation d'un échantillon a été effectuée. À ce moment, la typologie était déjà proche de la version définitive. Elle comportait dix classes. La catégorie *Mal effectué* était alors une classe distincte.

Pour la classification, 100 descriptions d'anomalies ont été sélectionnées. Pour chaque description, l'annotateur pouvait choisir jusqu'à quatre catégories différentes, de manière à attester des cas où plusieurs types de problèmes sont évoqués dans un seul rapport. Les annotateurs choisis ont consisté en deux linguistes et un annotateur expert du domaine. L'accord inter-annotateur a été calculé à l'aide du Kappa de Cohen. Pour le calcul, nous avons uniquement sélectionné la première catégorie (parmi les quatre que les annotateurs pouvaient sélectionner) puisqu'elle est systématiquement renseignée et nous avons calculé type par type. Les résultats de cet accord sont présentés en Table 5.15. Toutes les paires d'annotateurs obtiennent un résultat supérieur à 0,6 ce qui démontre un accord « acceptable ». Nous n'observons pas d'écart important entre les scores obtenus avec l'annotateur expert et les scores obtenus avec les linguistes uniquement (même si l'accord entre l'annotateur 2 et l'annotateur expert obtient des résultats légèrement inférieurs). La typologie est basée sur l'expression plus que sur l'interprétation du problème, ce qui semble rendre l'utilisation d'annotateurs non-experts envisageable pour cette tâche. Les annotateurs étaient aussi invités à laisser des commentaires. Les résultats obtenus ainsi que les commentaires effectués ont permis d'élaborer la version définitive de la typologie.

	Kappa de Cohen
Annotateur 1 et 2	0.67
Annotateur 1 et expert	0.67
Annotateur 2 et expert	0,63

TABLE 5.15 – Kappa de Cohen - Classification selon la typologie

À l'issue de cette étape, nous avons pu faire plusieurs constats. Tout d'abord, étant donné la coexistence de plusieurs catégories dans une FA, une classification du rapport n'est pas appropriée. Nous avons laissé la possibilité aux annotateurs de sélectionner jusqu'à quatre catégories. Cependant, nous ne pouvons exclure la présence de rapports avec plus de quatre catégories exprimées. De plus, les annotateurs peuvent placer les annotations dans un ordre différent, les résultats manquent alors de précision. Ils n'ont aussi aucune obligation quant aux nombres de catégories à renseigner, provoquant des différences dans le nombre d'annotations pour un même rapport en fonction des annotateurs. De ce fait, il nous est apparu plus approprié d'annoter les marqueurs spécifiquement pour la campagne

d'annotation que nous présentons dans le chapitre 6.

Pour la catégorie *Configuration hors spécification*, les annotateurs ont aussi montré une tendance à étendre sa définition. Des confusions sont apparues entre cette dernière et les cas avec des marqueurs tels « mal serrés, mal placé ... », mais aussi la catégorie « Action difficile ou impossible ». Cela nous a amené à reconsidérer les frontières entre ces différentes catégories, entraînant la disparition de la classe « Mal effectué ».

5.3.3 Un type, un frame

L'un des objectifs que nous visons est de repérer les différents éléments qui composent l'énoncé afin d'extraire ceux qui sont pertinents à la modélisation en *vépole*. Dans cette perspective, nous avons étudié la possibilité de rapprocher chacun des types d'un *frame* issu de *FrameNet*. Comme nous avons pu le voir dans le chapitre 3, *FrameNet* propose un schéma d'analyse et d'annotation pour chacun des *frames*. Ces différentes étiquettes permettent de qualifier les éléments de l'énoncé. Nous envisageons cette qualification comme essentielle pour définir les éléments que nous souhaitons extraire. Comme nous avons pu le voir dans le chapitre 3, il existe un *FrameNet* pour le français nommé de *FrenchFrameNet*. Toutefois, les domaines décrient par celui-ci ne recouvrent pas le champ de nos données [54]. Dans ce même chapitre, nous avons aussi pu voir que cette ressource est aisément transposable à d'autres langues. Ainsi, nous choisissons d'utiliser le *Framenet* original fait pour l'anglais. Dans cette section, nous présentons donc les différentes catégories de notre typologie d'expression d'un problème technique, et pour chacune, le *frame* qui lui est associé.

Index	**Type**	***Frame***
1	Fuite	Fluidic_motion
2	Signal qui s'est déclenché	Warning
3	Présence d'un obstacle	Hindering
4	Dégradation - Usure - Saleté	Damaging
5	Élément manquant	Presence
6	Configuration hors spécification	Meet_Specifications
7	Dispositif qui ne fonctionne pas	Being_operational
8	Action difficile ou impossible	Difficulty
9	État du monde	-

TABLE 5.16 – Correspondance entre les catégories et les *frames*

54. Les domaines que couvre le *FrenchFrameNet* sont : la causalité, la communication langagière, le jugement / l'évaluation, les positions cognitives, les relations spatiales, les relations temporelles et les transactions commerciales.

Pour chacune des catégories, nous avons recherché dans *FrameNet* des *frames* similaires. Pour ce faire, nous sommes partis des marqueurs repérés pour chaque classe et avons essayé de retrouver des unités lexicales (déclencheurs) correspondantes en anglais dans la base. C'est à partir de ces unités lexicales que nous avons sélectionné les *frames* auxquels ils se rapportent. Une fois les *frames* sélectionnés, nous avons annoté quelques exemples afin de valider la sélection.

En revanche, nous n'attendons pas une correspondance parfaite entre les *frames* et les catégories. En effet, comme nous avons pu le voir avec les travaux de L'Homme (2023) dans le chapitre 3, l'utilisation de *FrameNet* pour des domaines spécialisés peut amener à modifier les *frames* originaux afin de refléter plus fidèlement la situation en contexte.

Pour illustrer cela, nous pouvons prendre l'exemple de la catégorie *Signal qui s'est déclenché*. Pour cette classe, c'est le cadre *Warning*[55] qui semble le plus adapté. La définition de ce *frame* est la suivante :

> Un Locuteur (*Speaker*) averti un Destinataire (*Adresse*) en donnant un Message qui décrit une situation indésirable qui peut affecter le Destinataire. L'avertissement peut alternativement être décrit comme un Sujet (*Topic*) ayant trait à une situation indésirable ou délivré par un Medium spécifique.

Comme il transparaît dans cette définition et dans le schéma d'annotation Figure 5.1, il existe cinq éléments noyaux à ce *frame* : *Speaker, Adressee, Topic, Message* et *Topic*. Dans le cas des FA, le *Speaker* (ou locuteur) et l'*Adressee* (Destinataire) ne sont jamais mentionnés. Ces deux éléments du *frame*, selon *Framenet*, correspondent à des entités *sentient* (« douées de sensations »). Dans le chapitre 4, nous avons pu voir que le « je » était quasi totalement absent du corpus, ce qui est commun dans les textes techniques. De même, il est très rare qu'il soit fait mention d'un autre « opérateur », d'un Destinataire. Dans l'exemple 3 de la Table 5.6, nous pouvons voir un cas où un *Speaker* est effectivement présent : « L'OPERATEUR VIDEO » qui a « DECLENCHE L'ALARME »[56]. En revanche, dans l'exemple 4 « LES OPERATEURS » ne sont pas déclencheurs de l'alarme, mais en constatent simplement l'activation. Nous observons ici un élément pour lequel une modification pourrait être envisagée pour correspondre à notre contexte. Parmi les *frames* proposés par *FrameNet*, le cadre *Triggering* pourrait s'appliquer. Ce dernier relève des cas où une *Cause* entraîne un *System* à opérer selon son but visé. Ce cadre ne prend toutefois pas en compte la notion d'alerte, de prévention et s'éloigne trop de la situation que nous cherchons à décrire.

55. https://berkeleyfn.framenetbr.ufjf.br/frameIndex

56. Pour rappel, la phrase entière est « LORS DU REGLAGE DE LA CAMERA CRYO LOX, L'OPERATEUR VIDEO A DECLENCHE L'ALARME EN DONNANT UN COUP DE TETE DANS UN DETECTEUR. »

Frame Element	Core Type
Addressee	Core
Communicative_force	Extra-Thematic
Descriptor	Extra-Thematic
Iterations	Extra-Thematic
Manner	Peripheral
Means	Peripheral
Medium	Core
Message	Core
Place	Peripheral
Speaker	Core
Time	Peripheral
Topic	Core

FIGURE 5.1 – Schéma d'annotation de *Warning*

Dans les *frames*, nous retrouvons aussi les degrés de la typologie. Le bas du classement renvoie ainsi à des *frames* impliquant que très peu de *frame elements*, notamment noyaux, et inversement. La catégorie *Action difficile ou impossible* par exemple ne comprend que deux *frame elements*, l'*Experienceur*, celui qui éprouve la difficulté, et *Activity*, soit l'activité qu'il cherche à réaliser.

Dans les chapitres 8 et 9, nous aurons l'occasion d'analyser ces cas plus en détails. Nous reviendrons à cette occasion sur les *frames* correspondant et leur adéquation avec le contexte de spécialité dans lequel les énoncés prennent lieu.

5.4 Conclusion

Dans ce chapitre, nous avons fait usage de plusieurs méthodes d'analyse de corpus (le calcul de fréquences, le *Topic modeling* et *Word2Vec*) afin d'explorer le champ *Description de l'anomalie* de notre corpus. Cette exploration s'est focalisée autour de la notion de problème à travers des marqueurs lexicaux en relation avec cette notion. L'utilisation de ces trois méthodes a permis d'observer notre corpus sous différents angles. L'analyse des fréquences lexicales permet une approche d'analyse de corpus classique et quantitative afin de faire ressortir des régularités au niveau du mot. Le *Topic modeling* et *Word2Vec* sont deux approches d'apprentissage automatique. Le premier nous a permis d'observer des régularités au niveau du mot, mais selon une notion de proximité sémantique entre les documents, et ainsi d'effectuer des regroupements à un niveau différent. L'utilisation de *Word2Vec* s'est faite en focalisant notre étude sur des marqueurs préalablement identifiés et a ainsi permis

de repérer des mots similaires en fonction de leur proximité sémantique.

Ces trois approches nous ont guidés à la construction d'une typologie fondée sur l'expression d'un problème technique. Cette expression se manifeste par les marqueurs identifiés à travers les trois méthodes mentionnées ci-dessus, couplée avec une mise en regard systématique avec le corpus. L'identification des classes et des distinctions entre elles ont aussi été permises par notre familiarisation avec le corpus pendant ce travail de thèse et l'intervention d'un expert du domaine technique pour validation et annotation. Au-delà de ces marqueurs, l'hypothèse qui sous-tend cette typologie est celle de l'existence *patterns* propres à chacun des types. Pour représenter ces *patterns*, nous avons fait le choix d'explorer les *frames* et la ressource *FrameNet*, et dans quelle mesure nos catégories peuvent être rapprochées de cadres déjà existants, définis et démontrés dans des exemples (cependant parfois loin du domaine des FA). Nous avons pu voir que la correspondance entre les *frames* de *FrameNet* et les classes de notre typologie n'est pas systématique. Une correspondance parfaite n'était cependant pas attendue étant donné la différence de langue et le domaine de notre corpus.

Cette typologie nous permet d'ordonner et d'obtenir une vision plus éclairée d'un corpus vaste et difficilement accessible pour des non-experts. Dans le chapitre 6, nous présentons les résultats d'une classification automatique du champ « Description de l'anomalie » du corpus en fonction de cette typologie par le biais d'une tâche d'étiquetage des marqueurs. Cela nous permet d'obtenir tout un ensemble de marqueurs propres à chaque classe, ainsi qu'une classification entière du corpus. Cette classification facilite une étude approfondie de chaque catégorie. Ainsi, dans les chapitres 8 et 9, nous étudions en détail deux de ces classes : la *Fuite d'un solide ou d'un gaz* et la *Présence d'un obstacle*.

Troisième partie

Analyse à gros grains : marqueurs et bruit

Chapitre 6

Opérationnaliser la typologie des problèmes : le repérage des marqueurs

Sommaire

6.1 Introduction : une approche à gros grains

Dans ce chapitre, nous abordons le traitement automatique du corpus, et plus spécifiquement, le repérage et la catégorisation de marqueurs dans le champ « Description de l'anomalie » du corpus. Cette catégorisation s'est faite en fonction de la typologie d'expression d'un problème technique exposée dans le chapitre 5.

Nous souhaitons ici opérer une première étape d'étiquetage en fonction de la typologie, à travers le repérage des marqueurs lexicaux propres aux différentes catégories. Ces marqueurs

correspondent de fait aux déclencheurs du *frame*. Ces déclencheurs sont les indices lexicaux autour desquels nous avons bâtis la typologie. Ils se rapprochent des unités lexicales des *frames* mais restent issus du corpus uniquement. Cette tâche d'étiquetage est, du point de vue technique, une tâche de délimitation et de classification des segments, c'est-à-dire de classification de tokens. Elle nous permet, en définitive, d'obtenir des classes lexicales pour chacun des types, mais aussi une classification entière du corpus. Par ailleurs, il est apparu que toutes les catégories ne peuvent pas bénéficier du même traitement en raison des particularités qu'elles abritent et de l'adéquation avec le *frame* qui peut présenter des limites. En effet, en raison des niveaux variable d'exploitabilité pour la modélisation et de la diversité dans les types et la quantité d'information exprimée, les traitements à employer diffèrent. Lors des chapitres 8 et 9, nous pourrons effectivement constater les catégories *Fuite d'un liquide ou d'un gaz* et *Présence d'un obstacle* ne se prêtent pas au même traitement. Cette classification nous permet donc d'isoler les rapports en fonction de leur catégorie afin de les étudier de façon indépendante et de guider le repérage des participants du cadre. L'hypothèse derrière ces catégories est que chacune correspond à un / des patrons d'expression distincts. En opérant cette catégorisation, cela nous permet donc de distinguer les traitements à mettre en place. La question se pose cependant des rapports avec plusieurs marqueurs, de même catégorie ou non. Ces derniers soulèvent des questions quant à la segmentation du corpus qui, en raisons des problèmes de bruit, est difficile à effectuer pour ce corpus.

Par ailleurs, cette approche graduelle se base aussi sur l'hypothèse qu'un repérage des déclencheurs est plus accessible, notamment dans un texte aussi bruité et technique. En effet, le vocabulaire spécialisé peut poser problème pour la compréhension des rapports, et en conséquence pour leur annotation pour un public non-expert du domaine. Les marqueurs sont cependant plus accessibles et ouvrent ainsi la possibilité de faire appel à des annotateurs non experts du domaine, comme nous le faisons pour cette tâche. Cela peut se révéler utile face à la problématique de la disponibilité des experts pour des tâches aussi coûteuses en temps, problématique courante lors de travaux en domaine de spécialité. Par ailleurs, étant donné le bruit dans les données, le fait de s'appuyer en particulier sur certaines séquences du texte permettent de focaliser l'entraînement sur celles-ci. Ainsi, un apprentissage des marqueurs seulement suggère plus de robustesse. Nous aurons l'occasion, dans le chapitre 7 de constater et de mesurer la robustesse de cette approche par rapport au bruit.

D'un point de vue pratique, il est aussi apparu qu'une annotation complète en *frame elements* serait une charge trop lourde pour des annotateurs, et impossible pour des non-experts, comme nous avons pu le voir dans les chapitres précédents, et qui se confirmera dans les chapitres 8 et 9. Une approche en deux temps est plus raisonnable. Par ailleurs, dans le chapitre 4, nous avons pu voir les différents obstacles à l'utilisation de méthodes de prétraitement afin de nettoyer ou préparer le corpus à des traitements automatiques. De

ce fait, nous décidons de ne pas effectuer de prétraitement ou de nettoyage du corpus pour cette tâche.

Dans ce chapitre, nous définissons dans un premier temps la méthode employée et le processus d'annotation manuelle qui a été mis en place. Dans un second temps, nous présentons les résultats obtenus pour cette classification.

6.2 La question de l'annotation

La première étape de cette classification automatique a été la mise en application d'une annotation manuelle d'un échantillon du corpus pour l'entraînement et la validation d'un système par apprentissage. Pour ce faire, nous avons procédé en plusieurs étapes (Pustejovsky *et al.*, 2017) qui sont :

1. la définition de la méthode et la rédaction d'un guide d'annotation ;
2. une première étape d'annotation manuelle par des non experts du domaine afin de valider le guide ;
3. une annotation par un expert du domaine afin de valider les annotations non-expertes ;
4. l'annotation du *dataset* complet, réparti entre plusieurs annotateurs.

Nous allons désormais développer chacun de ces points de manière plus détaillée.

6.2.1 La méthode et le guide d'annotation

L'annotation manuelle que nous avons choisi de faire effectuer présente deux aspects :

— la sélection libre des segments correspondants aux marqueurs (ou déclencheurs) dans le texte ;

— la catégorisation des segments en attribuant une classe issue de la typologie.

Pour rappel, la typologie est composée des neuf classes suivantes :

1. *Fuite d'un liquide ou d'un gaz* ;
2. *Signal qui s'est déclenché* ;
3. *Présence d'un obstacle* ;
4. *Dégradation - Usure - Saleté* ;
5. *Élément manquant* ;
6. *Configuration Hors Spécification* ;
7. *Dispositif qui ne fonctionne pas* ;
8. *Action difficile ou impossible* ;

9. *État du monde.*

La tâche d'annotation que nous décrivons dans cette section est définie dans la version définitive du guide d'annotation en Annexe A. Dans ce guide, nous retrouvons les consignes techniques et méthodologiques à adopter. Nous avons aussi inclus une définition pour chacune des catégories accompagnée d'exemples. Nous présentons la conduite à tenir pour les cas avec plusieurs déclencheurs, mais aussi en cas d'hésitation entre deux catégories (c'est-à-dire prioriser en fonction des degrés de la typologie), et comment gérer la catégorie *État du Monde*.

Cette annotation n'est pas effectuée par des experts du domaine, mais par des linguistes. Ces derniers sont non-experts du domaine, mais possèdent cependant une expertise en annotation de données langagières. Nous avons choisi de procéder à une sélection libre des segments, ouvrant ainsi la possibilité d'avoir des segments discontinus, plusieurs segments par rapports, et des segments dont les frontières ne sont pas alignées sur des unités prédéfinies (l'étiquetage se fait au niveau du caractère). En effet, comme nous avons pu le voir dans le chapitre 5, les marqueurs sont variés, peuvent être constitués de plusieurs mots, voire être discontinus, et plusieurs peuvent être présents dans un même rapport. Dans l'exemple 1 de la Table 6.1, nous pouvons observer un cas où un marqueur est constitué de trois éléments discontinus « MESUREE A...POUR...MAXIMUM ». Pour ce type de cas, les annotateurs avaient la possibilité d'indiquer un lien afin de connecter les différents éléments du même marqueur. Dans l'exemple 2, nous avons deux marqueurs, mais qui sont relatifs à deux catégories distinctes. Dans ce cas, les annotateurs avaient pour instruction d'annoter les deux. Par ailleurs, la sélection libre permet de contourner les cas éventuels où un marqueur est collé à un autre mot en raison du manque d'un espace dans le texte. Dans l'exemple 3, pour le marqueur « HORS ECHELLE », le mot « HORS » est collé à « AIR » qui le précède, un espace est manquant.

De façon à pouvoir effectuer ce type d'annotation, nous avons utilisé la plateforme *Glozz* (Widlöcher et Mathet, 2009). Celle-ci a été conçue pour l'annotation discursive avec des schémas complexes. Elle est configurable à de nombreux niveaux, permet de définir son propre schéma d'annotation et d'effectuer une annotation à différents niveaux de granularité. Par ailleurs, la plateforme permet de créer des liens définissables entre les différents segments, palliant la problématique des segments discontinus.

Lors de la sélection des marqueurs, les annotateurs doivent aussi attribuer une catégorie au segment, parmi les neuf de la typologie. Concernant ces catégories, la question s'est posée pour *État du monde* qui ne possède pas de marqueur distinctif par définition. Est-ce qu'il est pertinent d'apposer une annotation s'il n'y a pas de marqueurs ? Sans segment sélectionné, cela signifie d'avoir des FA sans aucune annotation pouvant soulever une ambiguïté. Est-ce

N°	Exemples	Catégories
1	ASSIETTE ENTRE LES DEUX PAIRES DE VERINS, COTE MAT DE LA PALETTE 5/8 DU EAP 537.1, MESUREE A 10 mm POUR 6 mm MAXIMUM SPECIFIES DANS LA FICHE SMO MECA 24/0 (D)	Configuration hors spécification
2	LA VISSERIE N'EST PAS CONFORME A LA PROCEDURE. CERTIFICAT MANQUANT .VOIR ANNEXE A532 POA163 2952 CO (0) PAGE 5 ECRITE A LA MAIN.LIAISON BOULONNEE ACU/ACY 2624.	Configuration hors spécification & Élément manquant
3	ALIMENTATION EN AIR HORS ECHELLE PUPITRE- TABLE AU BIL ROUGE- TABLE AU BAF :* RPR 0018 ROUGE* RPD 004 ROUGE	Configuration hors spécification

TABLE 6.1 – Exemples de variations de marqueurs

que l'annotateur n'a pas su annoter le rapport ? Est-ce un oubli ? Nous avons donc fait le choix de demander aux annotateurs, si le cas d'*État du monde* se présente, d'annoter le premier token du rapport. Ainsi, cette information reste à disposition si nécessaire. Les rapports sont identifiés et lors de la classification automatique, nous avons testé les deux situations : avec et sans ces annotations. En revanche, cela entraîne un manque de cohérence sémantique sur ces segments.

L'une des questions à aborder dans le guide était les cas d'hésitation entre deux catégories. Dans cette situation, l'aspect graduel de la typologie entre en jeu. En effet, il était indiqué dans le guide, en cas d'hésitation entre deux marqueurs, de privilégier la catégorie la plus haut placée dans la typologie (*Fuite d'un liquide ou d'un gaz* étant la plus élevée et *État du monde* la plus basse). Cette instruction est basée sur le fait que, plus la catégorie est élevée, plus l'information est précise. De ce fait, en cas d'hésitation, on indique de choisir la catégorie à même de nous renvoyer le plus d'information possible. Dans le guide, nous retrouvons l'exemple suivant : « ABSENCE DE PROCEDURE AU FORMAT ARIANESPACE POUR LE RETOUR EN SECURITE (PROCEDURE FORMAT CSG) ». Si le marqueur "ABSENCE" peut évidemment indiquer l'appartenance au type "Absence d'un élément", la présence de standards dans la FA ("FORMAT ARIANESPACE" par opposition à "FORMAT CSG") peut faire pencher en faveur du type "Hors spécification". Le problème est-il l'absence d'une procédure au format ARIANESPACE ou le fait que la procédure soit au format CSG ? En l'absence de marqueur propre à la notion de hors spécification et étant donné que le type "Absence" possède un degré de précision plus élevé, on choisira de catégoriser la FA comme étant de type "Absence".

Enfin, l'aspect majeur demandé de la part des annotateurs lors de cette tâche était de *ne pas inférer* au sujet de la situation, la nature ou l'origine du problème en question. L'objectif de cette annotation est de repérer comment le problème se manifeste dans le texte, et non

de décrire ou de catégoriser le problème. Il est donc important de rester au plus proche du texte possible. La façon dont sont exprimés les rapports peut amener à supposer la nature du problème d'après les informations présentes, notamment pour un expert. Ce n'est pas ce qui est demandé pour cette tâche et seule la manière dont est exprimé le problème nous importe à ce stade. Nous pouvons voir dans la Figure 6.1 un extrait du guide d'annotation. Cet extrait s'accompagne d'un exemple de possible ambiguïté pour l'assignation du type. Il est demandé aux annotateurs d'éviter d'inférer sur les conséquences d'« INSTABLE » lors de la catégorisation. Tel que le rapport est exprimé, il n'est pas fait mention des conséquences de l'instabilité de la sonde. Dans cet exemple, cela correspond d'ailleurs aux règles de priorité qu'il est demandé d'appliquer.

— "DRUCK DPI610 N° 6280 SONDE INTERNE INSTABLE." : Ici, l'énoncé peut soit correspondre au type "Dispositif qui ne fonctionne pas", soit au type "Hors Spécification". Le marqueur "Instable" n'indique pas un empêchement à faire fonctionner la sonde et sans information en ce sens, la catégorisation de cet énoncé dépendra plutôt du type "Hors spécification", la sonde devrait être stable mais ne l'est pas.

FIGURE 6.1 – Extrait du guide : ne pas inférer

6.2.2 Pré-test et accord inter-annotateur

Afin de valider le guide, un pré-test d'annotation sur un échantillon du corpus a été organisé. Celui-ci a consisté en l'annotation d'un échantillon de 50 descriptions d'anomalies. Les rapports ont été choisis manuellement de façon à avoir une répartition minimale *a priori* de chaque étiquette (au minimum 5 occurrences par catégorie, d'après notre opinion). Les annotations ont été effectuées par trois annotateurs non-experts, mais aguerris à l'annotation. En effet, en raison d'un manque de disponibilité d'experts pour une annotation du *dataset* final pour l'entraînement du modèle, nous avons fait le choix de sélectionner trois linguistes en tant qu'annotateurs pour le pré-test. Une première évaluation a été effectuée entre ces trois linguistes, puis une seconde en comparant les résultats avec un annotateur expert, ce dont nous discuterons dans la section suivante. Cette première annotation nous a permis de valider et d'améliorer le guide d'annotation et de vérifier la viabilité des consignes.

En plus des consignes décrites dans la section précédente, il a aussi été donné la possibilité de signaler une FA comme *Incompréhensible* (en sélectionnant le premier token du rapport et en le typant avec cette valeur). En ce cas, toute la FA est concernée et aucune autre annotation n'est effectuée en son sein. Par ce biais, nous pouvons évaluer la capacité des annotateurs non-experts (les linguistes) à appréhender ces données sans connaissance

experte.

En Figure 6.2, nous pouvons observer la répartition des différentes catégories, tous annotateurs confondus, pour le pré-test avec un total de 217 segments identifiés. Nous pouvons observer une prépondérance des marqueurs de la catégorie *Configuration hors spécifications*. Nous rappelons ici que, pour chaque catégorie, nous nous attendons à trouver au minimum 5 segments, mais les segments supplémentaires reflètent la répartition des segments dans le corpus. Comme indiqué lors de la description de cette catégorie en 5.3.1.6, l'acception que nous avons en est large. De ce fait, ces résultats ne sont pas surprenants. En revanche, nous pouvons noter une forte propension des marqueurs de la catégorie *État du monde*. Les résultats par annotateur visibles dans la Figure 6.3 nous indiquent que l'annotateur 3 a une utilisation de *Configuration hors spécification* plus rare, avec 9 segments annotés contre 18 pour l'annotateur 1 et 15 pour l'annotateur 2, mais plus prolifique de la catégorie *État du monde* avec 14 segments contre 10 pour l'annotateur 1 et 8 pour l'annotateur 2[57].

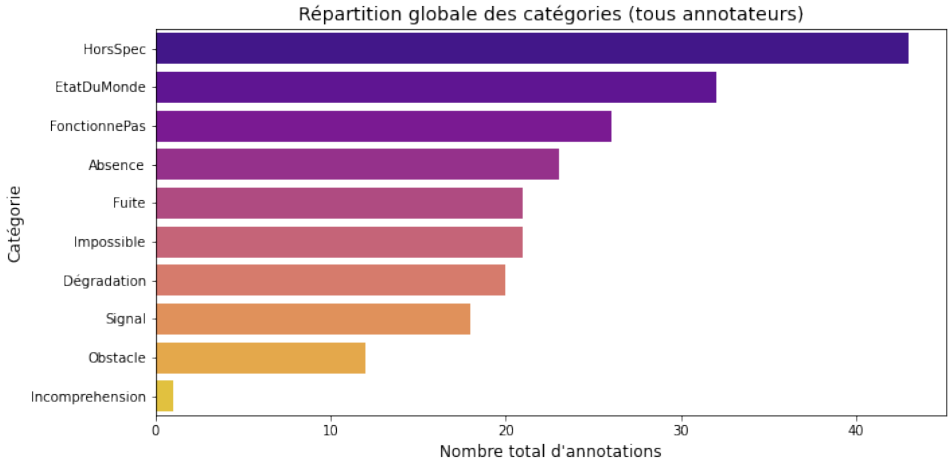


FIGURE 6.2 – Répartition de l'attribution des catégories

Pour les autres catégories, les résultats sont plus uniformes avec une différence de maximum 3 segments entre les différents annotateurs.

La décision de procéder à une sélection libre des segments entraîne des spécificités qu'il est nécessaire de prendre en compte lors du calcul de l'accord inter-annotateur (nous faisons abstraction des catégories dans les explications qui suivent pour simplifier notre propos). En effet, les différents annotateurs peuvent se retrouver à annoter le même segment mais avec

57. Les chiffres donnés sont regroupés dans le tableau 6.2

FIGURE 6.3 – Répartition de l'attribution des catégories par annotateur

des frontières différentes. Ainsi, dans le premier exemple de la Figure 6.4, nous pouvons voir un rapport avec trois segments annotés. Pour les deux premiers segments, nous sommes dans la situation où deux annotateurs identifient les mêmes frontières mais un troisième annotateur diffère. Pour le premier segment, l'annotateur 1 a sélectionné uniquement le mot « MESSAGE » alors que les annotateurs 2 et 3 ont sélectionné « MESSAGE D'ERREUR ». Pour le segment suivant, ce sont cette fois les annotateurs 1 et 3 qui sont en accord sur les frontières avec « NE SONT PAS CORRECTS » alors que l'annotateur 2 s'est contenté de « PAS CORRECTS ». Enfin, pour le dernier segment « DISJOINTS », seuls les annotateurs 1 et 3 ont annoté ce segment. Ici, nous sommes dans un cas de figure différent puisque les trois annotateurs ne sont pas d'accord sur le fait de sélectionner ce segment.

Pour le second exemple, nous pouvons observer un cas ou les trois annotateurs ont identifié un segment au même endroit, mais aucun des trois n'a sélectionné les mêmes frontières. L'annotateur 1 a sélectionné « APPARITION D'UN MESSAGE », l'annotateur 2 « APPARITION D'UN MESSAGE D'INFORMATION » et l'annotateur 3 « MESSAGE D'INFORMATION ». Ils ont tous en commun le mot « MESSAGE » mais les frontières avant et après diffèrent. Pour ces différents exemples, il n'est pas question de déterminer si l'une ou l'autre des réponses est « correcte ». Les frontières sont différentes, mais nous ne considérons pas que la sélection de l'une ou de l'autre taille de segment est plus exacte. De ce fait, lors du calcul de l'accord inter-annotateur pour l'étiquetage en catégories, nous prenons en compte le segment, par annotateur, si un chevauchement existe.

Suite à cette décision, nous pouvons désormais identifier tous les cas où les annotateurs ont repéré le « même » segment. Nous avons récupéré les index de début et de fin pour chaque segment et observons s'il existe un chevauchement entre ces index. Dans la Figure 6.5, nous

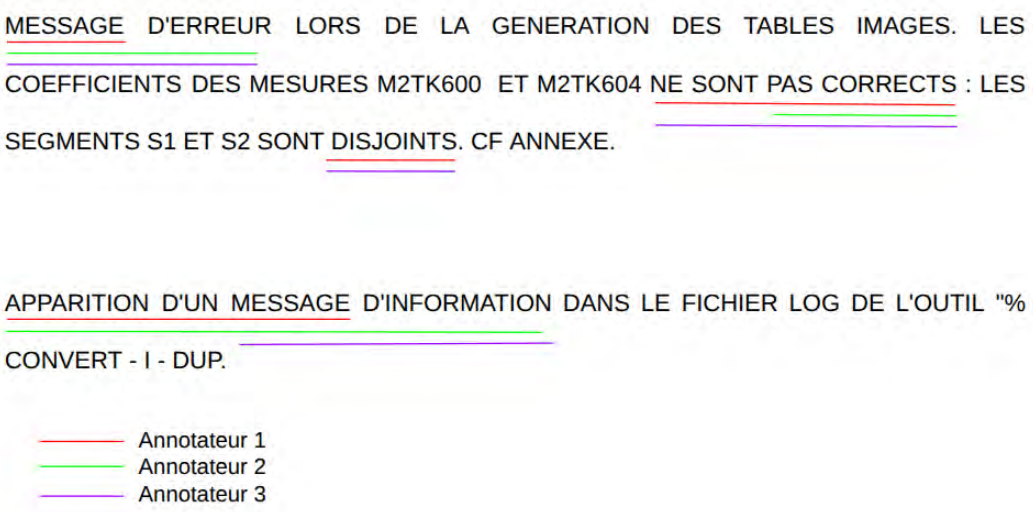


FIGURE 6.4 – Chevauchement des segments

présentons un diagramme représentant la répartition des segments entre les annotateurs. Les annotations montrent que sur l'ensemble des 217 segments annotés, 153 sont des segments communs entre les trois annotateurs. Les annotateurs 1 et 2 ont un nombre plus élevé de segments en commun, soit 122, c'est-à-dire avec un chevauchement même partiel, probablement en raison du plus fort nombre de types *Configuration Hors Spécification*. En revanche, étant donné que l'annotateur 3 a attribué plus de segments *État du monde* et que pour cette catégorie, il était demandé d'annoter le premier token du rapport, il est peu probable d'avoir un chevauchement pour ces segments avec une classe différente. Nous pouvons cependant noter que l'annotateur 3 a presque le même nombre de segments communs avec l'annotateur 1 et l'annotateur 2 (110 et 102 segments). Les annotateurs montrent un fort accord les uns avec les autres par rapport aux zones à annoter.

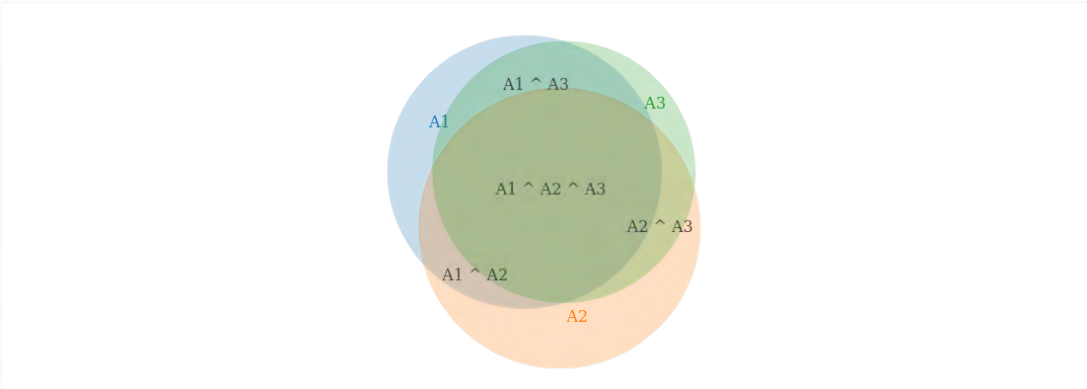


FIGURE 6.5 – Répartition des segments communs

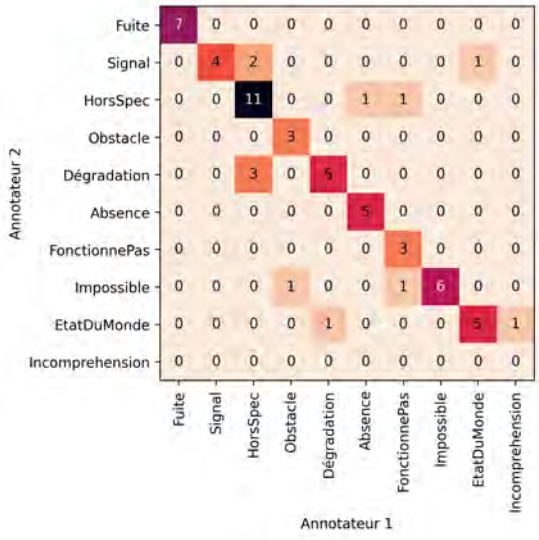
La question qui s'est alors posée est d'utiliser une mesure qui permet d'évaluer l'accord

inter-annotateur, mais en prenant en compte le chevauchement partiel des annotations. Le *kappa de Cohen*, couramment utilisé pour calculer l'accord inter-annotateur montre des limites puisqu'il ne prend pas en compte les différences de frontières, ni les cas où un segment n'est pas sélectionné par un des annotateurs. Il compare uniquement l'accord sur les catégories pour des items qui sont prédéfinis, communs pour tous les annotateurs. De ce fait, notre choix s'est porté sur le *gamma* de Mathet *et al.* (2015). Cette mesure correspond à nos besoins puisqu'elle prend en compte à la fois la classification des segments en fonction des catégories, mais aussi l'alignement des segments (le chevauchement des unités annotées malgré des frontières qui diffèrent). Le score obtenu, pour les trois annotateurs confondus est : **0,64** (sur une échelle de 0 à 1). Nous avons donc un accord inter-annotateur relativement fort entre les trois annotateurs sur cette tâche. Ces mesures restent complexes à interpréter, notamment étant donné les particularités de la tâche.

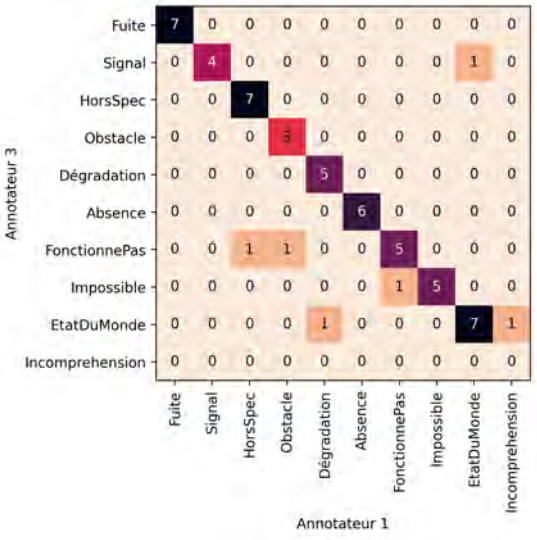
Afin d'observer plus en détail les résultats de la catégorisation, indépendamment de la sélection des segments, nous avons évalué les résultats pour les segments avec un recouvrement partiel. S'il existe un chevauchement entre des segments, nous en comparons les étiquettes, sinon, nous les ignorons. Comme nous pouvons le voir dans les matrices de confusion en Figure 6.6, la classification montre des résultats satisfaisants avec une diagonale qui se démarque fortement pour les trois paires d'annotateurs. Nous observons quelques cas de désaccords, mais qui restent minimes. Le *kappa* est de 0,77 pour les annotateurs 1 et 2 et 0,88 pour les annotateurs 1 et 3 et 2 et 3. Nous avons donc un accord fort d'une valeur correcte pour une tâche sémantique entre les différents annotateurs quant à la classification des segments. Ces résultats sont néanmoins à nuancer avec la prise en compte des segments qui n'ont pas de recouvrement partiel et les cas d'assignation de l'étiquette *État du monde* qui implique qu'un token différent est annoté.

En Figure 6.7, nous pouvons observer un exemple d'ambiguïté entre les classes issu des résultats du pré-test. Le segment repéré est « PROBLEME ». Ce dernier est vague et manque de précision. De ce fait, il a été annoté avec la classe *Configuration Hors Spécification* par les annotateurs 1 et 2 et la classe *Dispositif qui ne fonctionne pas* par l'annotateur 3. Est-ce que considérer que « PROBLEME D'IMAGE » implique que la caméra ne fonctionne pas est une inférence ? La limite est ici plutôt fine.

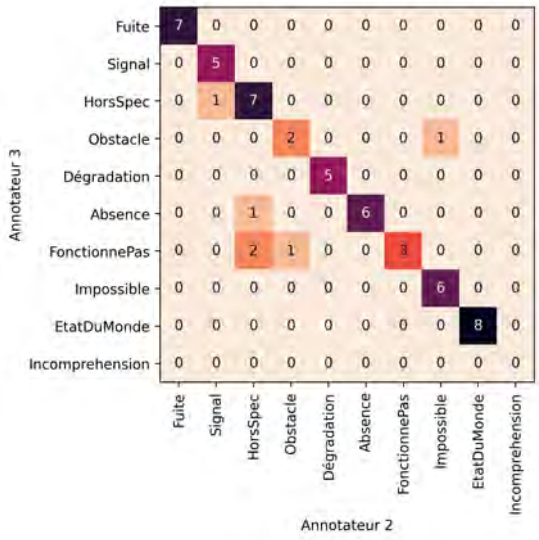
Dans l'exemple Figure 6.8, nous pouvons observer le seul cas annoté avec « Incompréhension » de tout l'échantillon, par l'annotateur 1 uniquement. Les deux autres annotateurs ont choisi l'étiquette « État du Monde ». Une discussion *a posteriori* avec l'annotateur 1 a permis de conclure que l'incompréhension ne venait pas tant du rapport lui-même, mais plutôt sur la façon d'annoter ce type de cas. De ce fait, il apparaît que les rapports de l'échantillon



A1 & A2



A1 & A3



A2 & A3

	Kappa
A1 & A2	0,77
A1 & A3	0,88
A2 & A3	0,88

Kappa de Cohen

FIGURE 6.6 – Matrices de confusion des annotations avec recouvrement partiel

sont suffisamment compréhensibles pour des annotateurs non-experts, dans la limite d'une annotation basée sur les marqueurs. L'étiquette *Incompréhension* a donc été supprimée pour l'étape d'annotation du *dataset*. Par ailleurs, l'annotateur a énoncé la fatigue liée à la tâche en raison du manque de compréhension des données. Pour effectuer cette tâche d'annotation, il a fallu 50 minutes à l'annotateur 1, l'annotateur 2 a rapporté un temps total de 3h15 mais avec un découpage en plusieurs séances qui a ralenti le processus, l'annotateur 3 a mis 1h15. Ces temps incluent la prise en main de l'outil et la lecture du guide d'annotation.

FIGURE 6.7 – Exemples d'ambiguïté entre deux classes

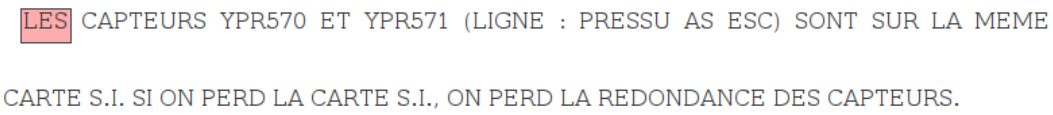


FIGURE 6.8 – Exemple d'« Incompréhension »

Cette première étape nous a permis d'adapter et de valider le guide d'annotation et le procédé choisi. Les scores obtenus sont satisfaisants, notamment en regard des difficultés posées par le corpus et la tâche (*gamma* : 0,64, *kappa* entre 0,77 et 0,88). Les cas de désaccord au niveau de la catégorisation restent marginaux. Ils découlent d'une tendance naturelle à vouloir inférer la nature de l'anomalie énoncée. Le recouvrement entre les marqueurs n'est pas exact avec certains segments qui ne sont pas annotés par tous les annotateurs (51 segments communs contre 77 segments maximum pour l'annotateur 2). En revanche, huit de ces cas concernent l'étiquette *État du monde* qui, comme nous l'avons vu, entraîne automatiquement cet écart. Sans compter ces cas-ci, il ne reste que trois marqueurs de différence entre le nombre de marqueurs communs et le nombre de segments annotés par l'annotateur 3, qui a été l'annotateur le moins prolifique (67 segments).

6.2.3 Annotation experte

Suite à cette première annotation non-experte, nous avons fait appel à un expert du domaine afin de valider les résultats obtenus. Ce dernier a procédé au même exercice, soit l'annotation du même échantillon de 50 descriptions d'anomalies.

Les résultats montrent un total de 70 segments annotés, ce qui se trouve dans l'intervalle du nombre de segments annotés par les annotateurs non-experts (entre 67 et 77 segments annotés). La Figure 6.9 montre une répartition qui paraît relativement similaire aux résultats des annotateurs non-experts. Nous pouvons cependant observer une proportion de la catégorie *Élément manquant* plus faible, au profit de la catégorie *Dispositif qui ne fonctionne pas*. Cela s'explique par l'interprétation des marqueurs « PERTE » et « NON ACQUISITION » que nous pouvons retrouver dans les rapports « PERTE DES 2 FBF CCX T2 11 sec APRES LE DECOLLAGE (H0) CF JDB JOINT » et « NON ACQUISITION YSY503. ». Pour l'expert, l'utilisation du mot *perte* signifie *la perte d'un signal* , donc qu'un appareil s'éteint, qu'un dispositif arrête de fonctionner. Le marqueur a donc été annoté avec l'étiquette *Dispositif qui ne fonctionne pas*. Concernant l'utilisation de « NON ACQUISITION » dans le second exemple, ce marqueur indique l'incapacité à obtenir les valeurs d'une mesure, ici la mesure de « YSY503 ». Ce marqueur a donc aussi été annoté comme *Dispositif qui ne fonctionne pas*.

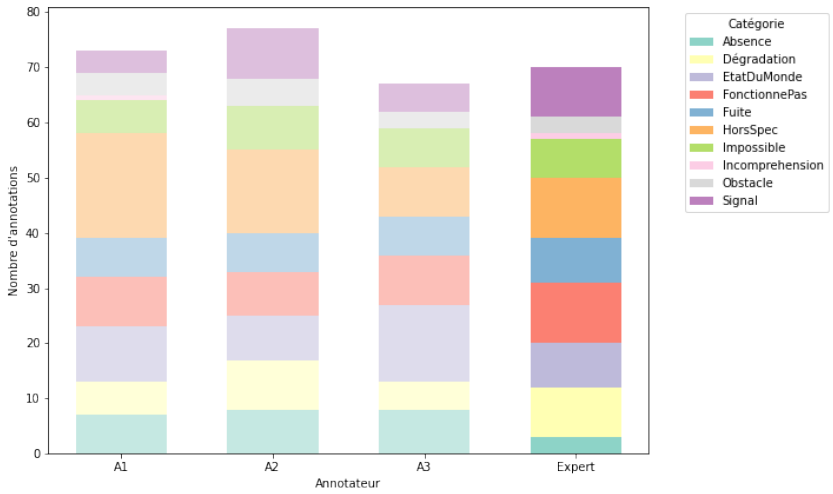
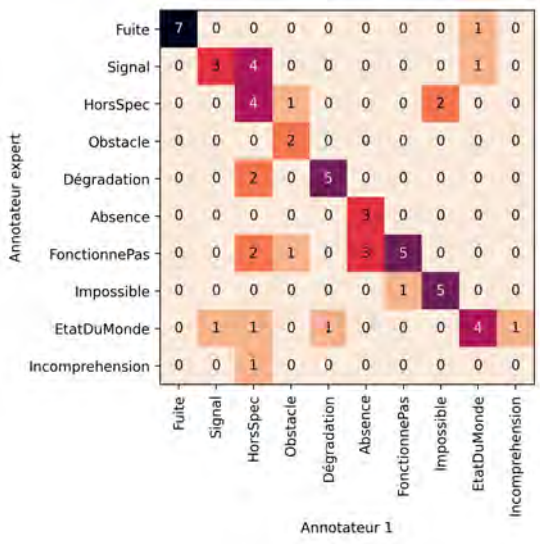


FIGURE 6.9 – Répartition de l'attribution des catégories pour l'annotateur expert

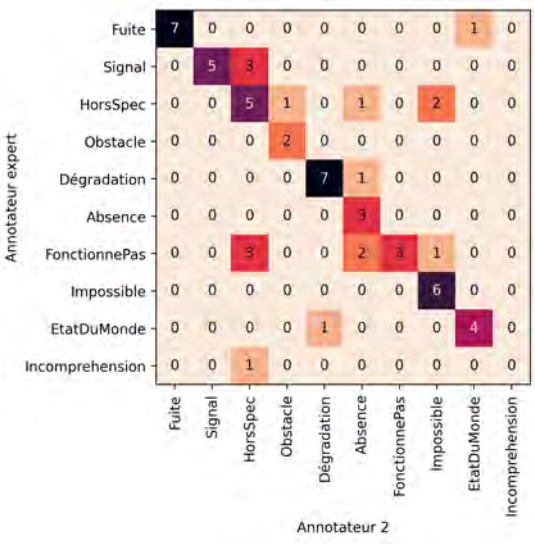
Parmi les marqueurs qui prennent une signification particulière en contexte, nous pouvons aussi noter « INVALIDE », visible dans le rapport « LE PARAMETRE DPRIE EST INVALIDE LORSQUE PMINRO OU PIEOESC EST INVALIDE. ». Selon l'expert, lorsque le terme *invalide* est utilisé, il est systématiquement question de *signal*.

Pour le calcul de l'accord inter-annotateur, nous avons à nouveau fait usage du *gamma*. Le score obtenu pour les trois annotateurs non-experts et l'annotateur expert est de **0,63** (pour rappel, il était de 0,64 pour les annotateurs non-experts). La différence est donc minime entre les deux scores.

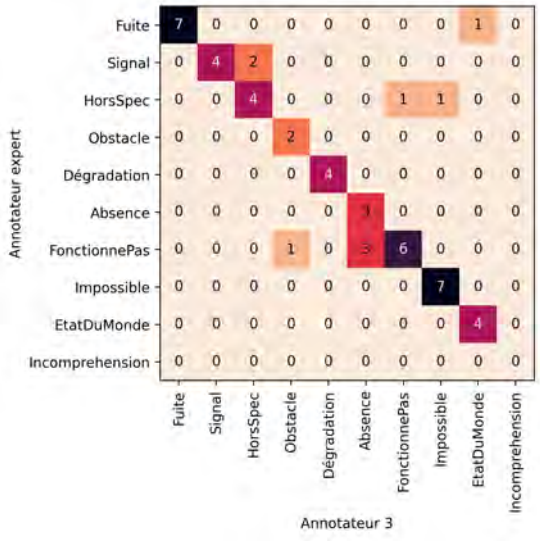
En revanche, pour la catégorisation des annotations, l'écart se creuse notamment pour l'accord entre l'annotateur 1 et l'annotateur expert. En effet, nous avons calculé le *kappa de Cohen* pour les segments avec recouvrement partiel entre les différents annotateurs et l'annotateur expert. Ces résultats sont présentés en Figure 6.10 avec leurs matrices de confusion respectives. Le *kappa* entre l'annotateur 1 et l'annotateur expert descend à 0,58 et monte à 0,8 pour l'annotateur 3, le score de l'annotateur 2 est entre les deux à 0,67.



A1 & A2



A1 & A3



A2 & A3

	Kappa
A1 & Expert	0,58
A2 & Expert	0,67
A3 & Expert	0,80

Kappa de Cohen

FIGURE 6.10 – Matrices de confusion des annotations avec l'expert

Ces résultats montrent bien que la connaissance experte offre une interprétation différente de certains termes, propres au domaine dont il est question. Les résultats ne montrent cependant pas une dégradation catastrophique au point de remettre en question l'entièreté des annotations non-expertes. Pour notre part, nous n'avons pas pu avoir accès à assez de temps de la part des experts pour effectuer l'annotation du *dataset* complet pour la tâche de classification. De ce fait, celle-ci a tout de même été effectuée par les annotateurs non-experts, en dépit des différences qu'implique le manque de connaissance du « vocabulaire scientifique général », au-delà des simples termes techniques.

	Annotateur 1	**Annotateur 2**	**Annotateur 3**	**Annotateur 4**
Nombre total de segments annotés	73	77	67	70
Par catégories				
Fuite d'un liquide ou d'un gaz	7	7	7	8
Signal qui s'est déclenché	4	9	5	9
Présence d'un obstacle	4	5	3	3
Dégradation - Usure - Saleté	6	9	5	9
Élément manquant	7	8	8	3
Configuration hors spécification	19	15	9	11
Dispositif qui ne fonctionne pas	9	8	9	11
Action difficile ou impossible	6	8	7	7
État du monde	10	8	14	8

TABLE 6.2 – Récapitulatif des résultats de l'annotation de l'échantillon

6.2.4 Annotation du *dataset*

Après ces deux étapes d'annotation d'un échantillon, nous avons lancé la campagne d'annotation du *dataset* pour la tâche de classification automatique. L'annotation manuelle a été effectuée cette fois encore par trois linguistes à qui nous avons répartis les données. Deux de ces linguistes sont les mêmes que pour le pré-test, ils possèdent donc une certaine familiarité avec les données.

Pour le *dataset*, 1000 rapports ont été sélectionnés de façon aléatoire parmi les 35 242 descriptions d'anomalies du corpus. Le *dataset* final est composé de ces 1000 rapports et des 50 annotés lors des étapes précédentes. Pour ces derniers, nous avons sélectionné les annotations de l'annotateur 1 qui obtenait un accord fort avec les annotateurs 2 et 3. Ce choix s'est fait avant d'obtenir les résultats de l'accord inter-annotateur avec l'expert. Nous obtenons donc un total de 1050 descriptions d'anomalies. La répartition des segments par catégorie

est présenté en Table 6.3. Le *dataset* contient un total de 1399 segments annotés, toutes catégories confondues. Le tableau montre un déséquilibre prononcé entre les différentes catégories. En effet, la catégorie *Configuration hors spécification* représente à elle seule près de la moitié du total avec 589 segments. Elle est constituée de plus de 3,5 fois le nombre de marqueurs par rapport à la catégorie suivante : *Élément manquant* avec 166 segments. Les catégories les moins représentées sont *Fuite d'un liquide ou d'un gaz*, *Signal qui s'est déclenché* et *Présence d'un obstacle* avec respectivement 61, 42 et 26 segments annotés. Nous pouvons également noter que ces trois catégories sont les plus hauts classées dans la typologie.

Le déséquilibre dans la répartition des données fait suite à une sélection aléatoire de l'échantillon, alors que pour le pré-test nous avions réalisé une sélection manuelle de façon à obtenir une représentation de toutes les catégories. Elle est représentative de la répartition réelle du corpus, mais aussi dû à l'acception large que nous faisons de la catégorie *Configuration hors spécification*. Elle risque aussi d'être un défi pour la tâche de classification automatique, notamment pour les catégories les moins représentées.

Type	**Nombre de segments**	**%**
Configuration hors spécification	589	42,1
Élément manquant	166	11,9
Dégradation - Usure - Saleté	152	10,9
Dispositif qui ne fonctionne pas	148	10,6
État du monde	114	8,1
Action difficile ou impossible	101	7,2
Fuite d'un liquide ou d'un gaz	61	4,4
Signal qui s'est déclenché	42	3.0
Présence d'un obstacle	26	1,9
Total	1399	100

TABLE 6.3 – Nombre de segments annotés par type

Les résultats de ces diverses étapes mettent l'accent sur plusieurs points de tension de notre étude. Tout d'abord, la question d'une annotation non-experte par rapport à une annotation experte est soulevée. Snow *et al.* (2008) ont déjà abordé cette question, et observent des résultats comparables entre une annotation experte et non-experte. En effet, ils démontrent

que pour un item, il est possible d'avoir une qualité d'annotation du même niveau qu'un annotateur expert à partir d'un certain nombre d'annotateurs non-expert (ce nombre dépend de la tâche en question). La population d'experts est dans leur cas constituée de linguistes ou de talistes, alors que les non-experts sont des individus anonymes non-spécialistes, issus d'une plateforme d'annotation en ligne. Les tâches d'annotations effectuées étaient : la reconnaissance des émotions, la similarité lexicale, la reconnaissance d'inférence textuelle, la reconnaissance d'événements temporels et la désambiguïsation lexicale. Toutefois, ces tâches et les corpus utilisés ne font pas appel à du vocabulaire et des connaissances du monde spécialisées. Dans notre cas, nous avons tout de même pu observer que, si les scores du *kappa* entre chaque annotateur non-expert et l'annotateur expert sont variables et moins élevés que souhaités dans certains cas, les résultats du *gamma* en revanche lissent ces écarts, avec un score de 0,63 pour les quatre annotateurs (expert inclus), contre 0,64 pour les trois annotateurs non-experts. L'expertise des linguistes en annotation permet peut-être de compenser certains écueils. L'annotateur expert dans notre cas avait aussi déjà une expérience en annotation de textes de spécialités. Il est donc aisé de trouver le problème et comment il s'exprime, mais quelques ambiguïtés résident encore au niveau de la typologie.

Dans le cadre de notre campagne d'annotation, nous ne pouvons cependant pas faire appel à un expert pour effectuer l'annotation du *dataset* complet pour des raisons de disponibilités. Nous ne pouvons pas non plus faire appel à des annotateurs tout-venant *via* une plateforme en ligne pour des raisons de confidentialité des données. Nous n'avons pas non plus demandé à un nombre élevé d'annotateurs linguistes d'effectuer cette annotation, afin d'avoir plusieurs annotations pour chaque item, pour des raisons de disponibilité et de confidentialité des données. De ce fait, nous avons fait effectuer l'annotation du *dataset* par trois annotateurs linguistes, non-experts du domaine.

Le *dataset* final est très déséquilibré en fonction des catégories avec *Configuration hors spécification* qui représente presque la moitié des cas du corpus et certaines catégories très faiblement représentées. Le choix d'une sélection aléatoire des données s'est fait afin d'avoir un classifieur entraîné sur un *dataset* représentatif du corpus réel. Cependant, ce déséquilibre pose un réel défi pour l'apprentissage du classifieur, et la fiabilité des résultats, notamment pour les catégories les moins représentées. Pour raison de contraintes temporelles, une seconde campagne d'annotation n'est cependant pas réalisable et l'apprentissage du modèle a été réalisé sur le *dataset* en l'état.

6.3 Apprentissage automatique : résultats et analyse

L'objectif de cette tâche d'apprentissage est d'obtenir une classification entière du corpus, en fonction de la typologie d'expression d'un problème technique. Cette classification est une première étape vers une analyse plus poussée de chacune des catégories, afin de dégager les *patterns* d'expression du problème. Nous présentons dans cette section la tâche de reconnaissance des marqueurs qui permet cette classification et les résultats obtenus.

6.3.1 Tâche et format des données

Les modèles utilisés pour l'apprentissage sont des modèles *transformers* préentraînés que nous avons *fine-tuné* sur une tâche de classification de tokens. Les modèles d'architecture *transformers* sont des réseaux de neurones caractérisés par le mécanisme d'*attention*. Ce mécanisme d'*attention* permet de traiter les données de manière non séquentielle et d'attribuer des poids aux données d'entrées (Vaswani *et al.*, 2017).

Parmi les modèles de types *transformers*, nous pouvons retrouver BERT (Bidirectional Encoder Representations from Transformers), un modèle pré-entraîné sur un vaste ensemble de textes. BERT est pré-entraîné sur une tâche de modélisation du langage masqué (*masked language model*), c'est-à-dire la prédiction de tokens masqués dans une séquence d'entrée à partir de son contexte. Il est aussi entraîné sur une tâche de prédiction de la séquentialité de deux phrases, permettant ainsi la capture de *patterns* entre les séquences (Devlin *et al.*, 2019). Au moment de la réalisation de cette tâche, BERT est le choix classique à adopter : il est (relativement) léger et a montré des résultats satisfaisants sur des tâches similaires à la nôtre. Bien que d'autres modèles plus performants auraient pu être envisagés, le choix de BERT s'avère pertinent dans une démarche d'exploration de données comme la nôtre.

Pour notre tâche, nous avons choisi de tester cinq modèles de types *transformers* basés sur BERT :

1. *bert-base-multilingual-uncased* ;
2. *camembert-base* ;
3. *camembert-large* ;
4. *camemberta* ;
5. *xlm-roberta-base*.

Le premier de cette liste, *bert-base-multilingual-uncased*, fait partie des modèles BERT initiaux. Il a été entraîné sur de l'anglais mais aussi sur d'autres langues (102 langues) dont le français, et ne prend pas en compte la casse (majuscules / minuscules). Pour chaque langue, le corpus d'entraînement est constitué de l'entièreté des pages Wikipédia dans la langue correspondante (à l'exception des pages *Utilisateurs* et *Discussions*).

Les deux modèles suivants, *CamemBERT-base* et *CamemBERT-large* (Martin *et al.*, 2020) sont basés sur l'architecture du modèle RoBERTa (Liu *et al.*, 2019), lui-même basé sur une stratégie similaire à BERT mais sans la tâche de prédiction de phrases. Ils sont entraînés exclusivement pour le français et se déclinent sous deux versions, la version *base* avec 110 millions de paramètres, et la version *large* avec 335 millions de paramètres. Le modèle camembert-base est appris sur le sous-corpus français du corpus OSCAR (Open Super-large Crawled Aggregated coRpus)[58] (Ortiz Suárez *et al.*, 2019) composé de 138 GB de données en français. Le modèle *camembert-large* est appris à partir de 135 GB de données du corpus CCNET[59] (Wenzek *et al.*, 2020).

Le modèle *CamemBERTa*[60] (Antoun *et al.*, 2023) est aussi dédié au français, mais cette fois basé sur le modèle DeBERTA V3. Il montre des scores supérieurs à ceux des modèles *CamemBERT* préalablement cités sur plusieurs tâches malgré avoir été appris sur approximativement 30% des données utilisées pour les modèles *CamemBERT*.

Enfin, le dernier modèle est *xlm-roberta-base*[61] (Conneau *et al.*, 2020). C'est un modèle multilingue qui inclut le français et qui a été entraîné sur des données issues des corpus *Common Crawl*[62]. Il est particulièrement adapté à du *fine-tuning* pour des tâches de classification de séquences, de tokens ou de questions-réponses.

Pour effectuer la tâche, nous procédons à un *fine-tuning*. Le *fine-tuning* consiste à entraîner un modèle (déjà pré-entraîné) sur des données et/ou une tâche spécifique. En effet, le préentrainement d'un modèle est très coûteux en termes de quantité de données nécessaires mais aussi de temps de traitement. De plus, les tâches génériques de ce type de modèles permettent au système d'apprendre des comportements langagiers spécifiques, les représentations des tokens et aussi des structures du langage. Le *fine-tuning* permet de prendre appui sur ces résultats tout en rendant possible une adaptation à des données spécifiques qui diffèrent des données d'entraînement (comme dans notre cas) et de se concentrer sur une tâche spécifique. Il peut prendre deux modalités. Pour la première, le modèle sert d'extracteur de caractéristiques et on applique un classifieur en sortie et c'est ce dernier qui est entraîné sur les données. Dans la seconde, la phase d'entraînement vient modifier les couches du modèle et le classifieur. La deuxième méthode est plus coûteuse mais plus appropriée lorsque nous sommes en présence d'un domaine cible différent de celui des données du préentraînement, ce qui est notre cas. Nous donnons donc au modèle pré-entraîné un échantillon de nos données annotées afin qu'il se familiarise avec ses spécificités, et nous l'entraînons sur une

58. https://oscar-project.org/
59. https://paperswithcode.com/dataset/ccnet
60. https://gitlab.inria.fr/almanach/CamemBERTa
61. https://github.com/facebookresearch/fairseq/tree/main/examples/xlmr
62. https://commoncrawl.org/overview

tâche spécifique d'étiquetage de séquences.

Afin de mettre en place ce *fine-tuning*, la question du format des données s'est posée. En effet, l'annotation manuelle a été effectuée sans restriction sur les frontières des segments, laissant même la possibilité d'annoter à l'intérieur d'une chaîne de caractères, ne sélectionnant ainsi qu'une partie de la chaîne. Cela s'est fait notamment en raison des absences d'espaces pouvant être présentes dans les textes. Cette annotation a donc été effectuée sur un texte non-traité, sans *tokenisation* préalable du *dataset*. Or, les modèles que nous utilisons effectuent leur propre *segmentation* en tokens qui prennent la forme de *subwords*, soit des unités définies statistiquement lors de l'apprentissage sans respect pour les unités langagières. Par exemple, pour *bert-base-multilingual-uncased*, le mot « desserrage » de notre corpus est découpé en trois sous-unités : « des ##ser ##rage ». Le symbole « ## » indique que le *subword* fait partie du même mot. S'ils sont absents, cela signifie le début d'un nouveau mot. La raison de ce découpage provient du fait que le modèle n'a connaissance que d'un nombre fini de tokens, de l'ordre de quelques dizaines de milliers, qui doivent couvrir toutes les séquences de caractères possibles. Le découpage (en 1, 2, 3... subwords) est basé sur les distributions de caractères dans les séquences.

Tokens	**BIO**
des	B-Degradation
##ser	I-Degradation
##rage	I-Degradation
de	O
2	O
vis	O
sur	O
le	O
levier	O
de	O
de	O
##connexion	O
e	O
##pso	O

TABLE 6.4 – Tokenisation et étiquetage

Il nous a donc fallu faire correspondre nos annotations à la version *tokenisée* par le modèle du *dataset*. Nous avons donc fait usage de la convention **BIO** (*Begin In Out*) pour délimiter

les segments annotés. Le premier *subword* annoté d'une séquence se trouve attribué le **B**, et les suivants le **I**. Tous les autres *subwords* du rapport, ne faisant pas partie d'une annotation, se trouvent attribuer le **O**. Les annotations ayant été effectuées au niveau du caractère, il est possible que la frontière d'une annotation ne correspondent pas à la frontière d'un *subword*. Le cas échéant, nous sélectionnons le *subword* entier qui inclut la frontière de l'annotation en question. Nous avons aussi converti tout le *dataset* en minuscules afin d'éviter les problèmes de casse (pour *bert-base-multilingual-uncased*, cette conversion se fait automatiquement). Nous pouvons voir dans la Table 6.4 ce procédé pour le rapport « DESSERRAGE DE 2 VIS SUR LE LEVIER DE DECONNEXION EPSO » dans lequel « DESSERRAGE » est un segment du type *Dégradation-Usure-Saleté*.

Le *dataset* complet (1050 descriptions d'anomalies) a ensuite été découpé en deux, un corpus d'entrainement et un corpus de test[63] au ratio 80%/20%. Les sous-parties n'ont pas été produits de façon aléatoire mais de manière stratifiée afin de représenter la proportion globale des catégories. Le ratio s'est appliqué non pas sur le nombre de documents, mais sur le nombre de segments. Pour chaque catégorie, nous avons fait en sorte d'avoir le nombre de segments nécessaire pour représenter 20% du nombre total de segments dans le corpus de test (exception faite de *Configuration hors spécification* qui a servi de marge de manœuvre pour avoir une représentation exacte des autres catégories). Nous obtenons à la fin un test set qui contient 288 segments. Le détail de la répartition des segments par catégorie est dans la Table 6.5.

Catégorie	**Nombre d'items dans le train set**	**Nombre d'items dans le test set**
Fuite d'un liquide ou d'un gaz	54	11
Signal qui s'est déclenché	41	11
Présence d'un obstacle	27	5
Dégradation - Usure - Saleté	156	28
Élément manquant	161	37
Configuration hors spécification	612	119
Dispositif qui ne fonctionne pas	144	33
Action difficile ou impossible	95	21
État du monde	111	23
Total	1401	288

TABLE 6.5 – Répartitions des segments par types dans le train et test set

63. Étant donné la petite taille du *dataset* et notamment la faible représentation de certaines catégories, nous avons décidé de faire l'impasse sur le corpus de *dev*.

6.3.2 Présentation des résultats

Maintenant que nous avons établi le procédé d'annotation manuelle et le format nécessaire à la tâche, nous pouvons observer les résultats obtenus. Étant donné le caractère particulier de la catégorie *État du monde* (l'absence de marqueurs spécifiques), nous avons choisi de tester la tâche avec et sans cette étiquette. Nous présentons en premier les résultats avec l'étiquette et dans un second temps, sans l'étiquette.

6.3.2.1 Résultats de l'apprentissage

Nous avons présenté dans la section précédente les différents modèles que nous avons testés pour effectuer l'apprentissage du modèle. Pour ce faire, nous avons lancé chaque modèle cinq fois sur la même tâche, afin de circonvenir au non-déterminisme des algorithmes d'apprentissage. Les paramètres utilisés sont les suivants :

— taille du *batch*[64] = 16 ;
— *learning-rate* = 10^{-4} ;
— nombre d'*epochs* = 20.

Les résultats en Table 6.6 montrent que *bert-base-multilingual-uncased* et *camembert-large* sont les deux modèles qui obtiennent les meilleurs scores, avec 0,68 de F-mesure pour *bert-base-multilingual-uncased*, et 0,67 pour *camembert-large* (les scores correspondent à la moyenne des 5 runs et sont calculés au niveau du token avec une variation de 0,01 à 0,03 points entre les différents runs). Le modèle *camembert-large* agrège plus de paramètres que les autres, et nous pourrions attendre une meilleure performance de sa part. Cependant, nous faisons ici l'hypothèse que, puisque les modèles CamemBERT sont entraînés pour le français, l'absence d'accents et de casse a pu rendre la reconnaissance des tokens moins facile pour le modèle.

Modèle	Précision	Rappel	F-mesure
bert-base-multilingual-uncased	**0,68**	**0,68**	**0,68**
bert-base-multilingual-cased	0,65	0,63	0,64
camembert-base	0,64	0,68	0,66
camembert-large	0,65	0,69	0,67
camemberta-base	0,65	0,68	0,66
xlm-roberta-base	0,64	0,66	0,65

TABLE 6.6 – Résultats de l'apprentissage automatique par modèle (macro-moyenne)

64. Pour *camemBERT-large*, la taille du *batch* est de 8 en raison de limites de mémoire de la plateforme

Afin d'observer les résultats, nous souhaitons mesurer la catégorisation au niveau du segment entier (et non des subwords), pour obtenir une estimation de l'efficacité du système. La question des scores pour cette tâche a donc posé des questions similaires à celle du calcul de l'accord inter-annotateur, c'est-à-dire la question des frontières des segments. En effet, nous avons pu voir lors de l'annotation manuelle que les annotateurs pouvaient annoter le même segment mais avec des frontières qui diffèrent (cf. Figure 6.4). Nous considérons de même qu'il n'y a pas de frontières exactes systématiques au niveau des marqueurs. De plus, les données d'entraînement contiennent des annotations issues de trois annotateurs différents qui ont donc des logiques de sélection des frontières qui peuvent différer. Par exemple, pour le marqueur « MESSAGE D'ALARME », un annotateur peut annoter systématiquement le segment entier, alors que le second annote uniquement le mot « ALARME ». Le modèle est donc confronté à deux logiques d'annotation qui peuvent différer et démontrent d'une inconsistance au niveau des frontières des segments. Par ailleurs, l'objectif de cette tâche est de pouvoir permettre une analyse fine des différentes catégories, et non la reconnaissance parfaite du marqueur. Ainsi, afin d'obtenir des scores qui reflètent cette spécificité, nous avons fait usage des métriques proposées par *nervaluate* (Batista, 2018). Ces métriques sont adaptées aux tâches de reconnaissance d'entités nommées, et par extension à la tâche d'étiquetage de séquences. Elles mesurent l'efficacité d'un système au niveau de l'entité (tous les éléments étiquetés avec B ou I), plutôt qu'au niveau du token.

Plusieurs mesures sont proposées par *nervaluate*, parmi lesquelles se trouve *entity-type*. Cette dernière calcule les scores selon la présence d'un chevauchement entre l'entité étiquetée par le modèle et le *gold*. Tout recouvrement est une correspondance mais la catégorie doit être correcte. Dans l'exemple (factice) Table 6.7, la métrique *entity-type* compte la sortie du modèle comme correcte malgré le décalage des frontières. La Table 6.8 montre les résultats obtenus pour le modèle *bert-base-multilingual-uncased*. Selon cette métrique, le score de F-mesure atteint 0,74 de macro-moyenne.

Token	**Gold**	**Output**
un	O	B-Signal
message	B-Signal	I-Signal
d'	I-Signal	I-Signal
erreur	I-Signal	I-Signal

TABLE 6.7 – Exemple de segment avec recouvrement partiel

Comme pressenti, le type *Fuite d'un liquide ou d'un gaz* obtient les meilleurs résultats avec un score parfait (F-mesure de 1). Cela s'explique par le peu de variabilité au sein des

marqueurs caractéristiques de cette catégorie. À l'inverse, le type *État du monde* obtient les moins bons résultats, à hauteur de 0,43 de F-mesure. Ce résultat s'avère tout de même plus que correct pour une catégorie dont les marqueurs n'ont aucun trait sémantique ou lexical commun (nous rappelons ici que les annotateurs avaient pour consigne d'annoter le premier mot de la FA). Une analyse des résultats montre que le modèle a saisi le fait de sélectionner le premier mot dans le rapport de cette catégorie. Ainsi, s'il le repère correctement, le segment à annoter est systématiquement correct malgré l'absence de traits sémantiques à repérer. Faisant suite à *État du monde*, la catégorie *Présence d'un obstacle* obtient une F-mesure à hauteur de 0,60, cette catégorie étant la moins représentée au sein du corpus avec 26 segments annotés seulement, dont 5 dans le corpus de test.

Score *entity-type*				
Catégorie	**Précision**	**Rappel**	**F-mesure**	**Items**
Fuite d'un liquide ou d'un gaz	1	1	1	11
Signal qui s'est déclenché	1	0,73	0,84	11
Présence d'un obstacle	0,60	0,60	0,60	5
Dégradation - Usure - Saleté	0,85	0,78	0,80	28
Élément manquant	0,76	0,86	0,81	37
Configuration hors spécification	0,66	0,79	0,72	119
Dispositif qui ne fonctionne pas	0,72	0,76	0,74	33
Action difficile ou impossible	0,75	0,86	0,80	21
État du monde	0,43	0,43	0,43	23
Total (macro-moyenne)	0,71	0,77	0,74	288

TABLE 6.8 – Résultats par étiquettes pour *bert-base-multilingual-uncased*

La matrice de confusion de la Figure 6.11 montre les résultats obtenus par catégorie pour les segments du test set avec recouvrement non-nul. Nous pouvons voir que la diagonale se distingue, sauf pour la catégorie *Présence d'un obstacle* qui reste très claire avec seulement 3 occurrences. Étant donné qu'il n'y a que 5 segments de cette classe dans le test set, ce résultat n'est pas significatif. Nous pouvons voir que dans les cas avec un recouvrement partiel, les désaccords sur la catégorie à attribuer sont relativement faibles. La majorité se retrouvent sur l'attribution erronée de l'étiquette *Configuration hors spécification* qui, nous le rappelons, est fortement majoritaire dans le corpus et créé donc un déséquilibre dans les données d'entraînement qui se répercute dans les résultats.

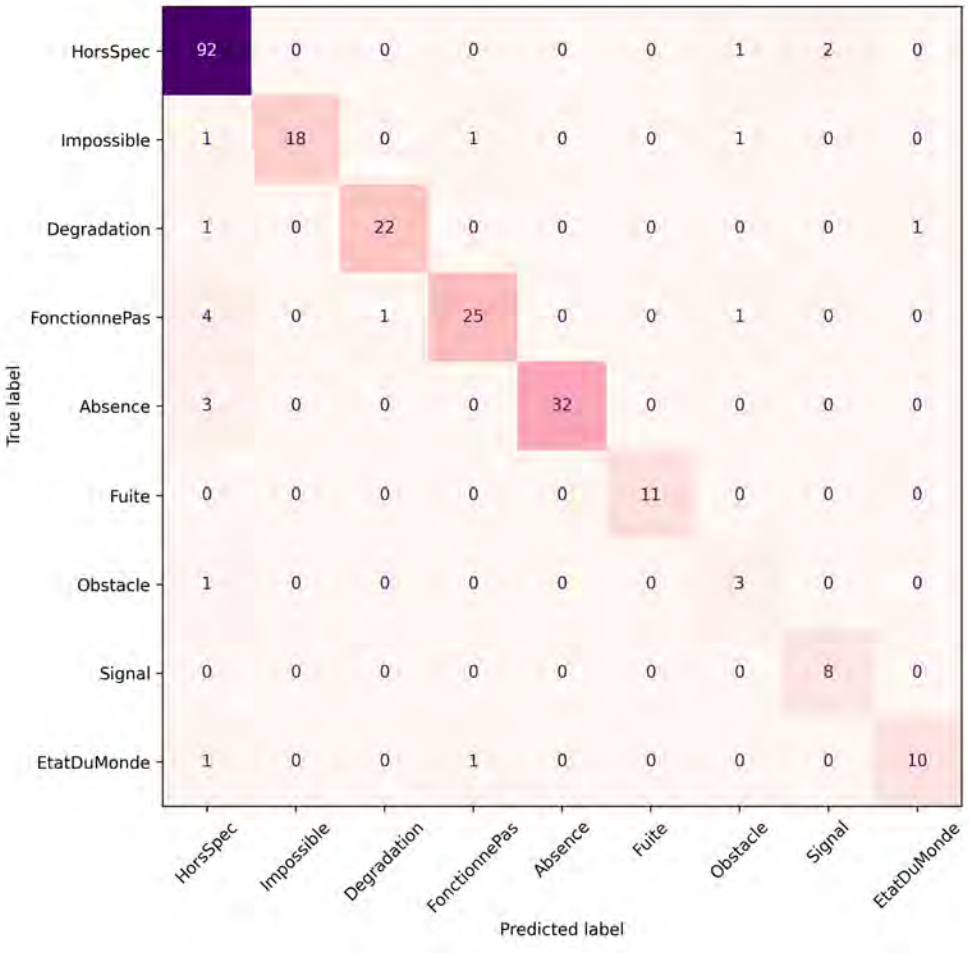


FIGURE 6.11 – Matrice de confusions des résultats par types de l'annotation

Nous avons effectué une deuxième version de cette tâche en supprimant cette fois les segments *État du monde*. Les rapports contenant un segment avec cette étiquette sont maintenus au sein du corpus, mais cette fois sans aucun segment annoté (tous les *subwords* se voient attribuer l'étiquette O). Nous avons cette fois-ci uniquement lancé les modèles *bert-base-multilingual-uncased* et *camembert-large*, soit les deux modèles ayant obtenu les meilleurs scores lors de la classification avec les neuf étiquettes.

Les scores obtenus au niveau du token sont à hauteur de 0,67 de F-mesure pour les deux modèles, soit presque les mêmes que ceux obtenus avec la catégorie. Nous n'observons donc pas d'amélioration, voir une légère diminution pour le modèle *bert-base-multilingual-uncased*. Le score de la métrique *entity-type* est de 0,72 pour le modèle appris sur *bert-base-multilingual-uncased*, soit légèrement inférieur à nouveau que les résultats avec l'étiquette.

Nous pouvons observer les résultats par étiquette dans la table 6.9. Ils montrent à la fois une baisse du rappel et/ou de la précision en fonction des catégories sans tendance

spécifique. La catégorie *Signal qui s'est déclenché* montre cependant une baisse importante pour les deux métriques, alors que les écarts pour les autres catégories sont moins marqués. La catégorie *Présence d'un obstacle* ne subit aucune modification dans les scores. Le rappel de l'étiquette *Élément manquant* est le seul qui montre une légère augmentation.

Score *entity-type*				
Catégorie	**Précision**	**Rappel**	**F-mesure**	**Items**
Fuite d'un liquide ou d'un gaz	1	0,92 (-0,08)	0,96 (-0,04)	11
Signal qui s'est déclenché	0,75 (-0,25)	0,55 (-0,18)	0,63 (-0,21)	11
Présence d'un obstacle	0,60	0,60	0,60	5
Dégradation - Usure - Saleté	0,82 (-0,02)	0,76 (-0,02)	0,79 (-0,01)	28
Élément manquant	0,71 (+0,05)	0,92 (+0,06)	0,80 (+0,01)	37
Configuration hors spécification	0,65 (+0,01)	0,69 (-0,10)	0,67 (-0,05)	119
Dispositif qui ne fonctionne pas	0,68 (+0,04)	0,76	0,71 (+0,03)	33
Action difficile ou impossible	0,69 (-0,06)	0,86	0,76 (-0,04)	21
Total (macro-moyenne)	0,70 (+0,01)	0,75 (-0,02)	0,73 (-0,01)	288

TABLE 6.9 – Résultats du système *bert-base-multilingual-uncased* par type sans la catégorie *État du monde*

Pour la suite du travail, nous avons donc choisi de faire usage du modèle appris à partir du modèle *bert-base-multilingual-uncased* avec la catégorie *État du Monde*. Celui-ci a obtenu 0,74 de F-mesure pour la métrique *entity-type* est avec des scores au-dessus de 0,6 pour toutes les catégories sauf *État du Monde*.

6.3.2.2 Un aperçu des résultats

L'application du modèle appris sur *bert-base-multilingual-uncased* sur le corpus entier nous a permis d'obtenir une classification de toutes les FA. Cette classification s'est faite en fonction de la catégorie d'expression du problème de la description de l'anomalie, à partir des marqueurs repérés en leur sein. Dans la Table 6.10, nous pouvons observer la répartition des segments, et en conséquence des FA, par catégorie. La répartition est similaire à celle retrouvée lors de l'annotation manuelle du corpus. Le total atteint est de 41 396 segments annotés, pour 35 242 rapports. Pour rappel, plusieurs segments peuvent être annotés dans un rapport, un même rapport peut donc être classé dans plusieurs catégories s'il agrège des segments de classes différentes. La catégorie *Configuration hors spécification* représente la catégorie la plus fréquente (42,5% des segments) et les trois catégories les moins représentées sont toujours *Fuite d'un liquide ou d'un gaz*, *Signal qui s'est déclenché* et *Présence d'un obstacle*.

Catégorie	Effectifs	En %
Configuration hors spécification	19 716	42,5
Élément manquant	6315	13,6
Dégradation - Usure - Saleté	6168	13,3
État du monde	4134	8,9
Dispositif qui ne fonctionne pas	4072	8,8
Action difficile ou impossible	3169	6,8
Fuite	1997	4,3
Signal qui s'est déclenché	873	1,9
Présence d'un obstacle	499	1,1
Total	46 396	100

TABLE 6.10 – Répartition des segments annotés automatiquement par catégorie

En revanche, certaines erreurs dans l'application du schéma d'annotation par le modèle ont été repérées. En effet, nous avons pu observer des cas où le modèle est inconsistant au niveau des types lors de l'annotation d'un même segment. Dans l'exemple Table 6.11, nous pouvons voir un cas dans lequel le marqueur, divisé en deux tokens, a été annoté avec deux types différents. D'autres cas similaires ont été repérés avec un seul marqueur qui porte deux catégories différentes en fonction des tokens. Ces incohérences peuvent apparaître entre une étiquette B et une étiquette I, mais aussi entre deux étiquettes I.

Rapport	LA MESURE M2TC531 (TEMPERATURE GAZ DE PRESSURISATION) EST BRUITEE : ENTRE -1 ET +4 CODES.	
Marqueur	BRUITEE	
Tokenisation	bru	##itee
Annotation	B-Degradation	I-HorsSpec

TABLE 6.11 – Exemple d'incohérence dans l'annotation

Dans la Table 6.12, nous pouvons retrouver le taux de segments présentant une incohérence pour chaque étiquette. Ces taux ont été calculés d'après le corpus complet, soit les 35 242 rapports après application du modèle. Pour la classe *Signal qui s'est déclenché*, ce taux atteint 18,10% des segments. Il est élevé, notamment si on prend en compte le faible nombre de segments annotés pour cette catégorie (973 segments). Le taux le plus faible est retrouvé dans la catégorie *Fuite d'un liquide ou d'un gaz*, pour laquelle les marqueurs montrent peu de diversité (*fuite, fuyard, écoulement*).

Type	**% de segments incohérents**
Signal qui s'est déclenché	18,10
Dispositif qui ne fonctionne pas	10,56
Action difficile ou impossible	9,91
Dégradation - Usure - Saleté	9,29
Élément manquant	7,33
Présence d'un obstacle	4,81
Configuration hors spécification	3,33
État du monde	2,54
Fuite d'un liquide ou d'un gaz	0,75

TABLE 6.12 – Taux d'incohérences par étiquette

Dans l'évaluation des résultats précédemment présentée, la mesure *entity-type* prend en charge ces cas de la façon suivante :

— si le segment annoté et le *gold* ont la même étiquette pour le premier *subword* (avec le label *B-*) mais que les *subwords* suivant (avec le label *I-*) diffèrent, le score est compté bon aux deux-tiers ;
— si le label *B-* diffère entre le *gold* et le segments annoté par le modèle, mais que les labels *I-* sont les mêmes, le score est à O.

Il semblerait donc que l'outil attribue un poids plus conséquent à l'étiquette du premier *subword* du segment.

Une analyse manuelle des segments a démontré que la raison principale semble venir de la catégorie *Configuration hors spécification*. En effet, la majorité des associations avec une incohérence impliquent cette catégorie. Au sein du tableau 6.13, nous pouvons observer que parmi les cinq paires B-I et I-I qui présentent une incohérence, il n'y en a qu'une qui n'implique pas la catégorie *Configuration hors spécification*. Le déséquilibre des données serait donc à l'origine de ces anomalies, plus qu'une difficulté à séparer deux catégories.

L'application du modèle sur le corpus complet nous a permis d'observer les cas avec plus d'un marqueur. Étant donné les incohérences que nous venons de présenter, nous avons choisi de prendre en compte uniquement les étiquettes « B- ». Nous relevons donc, pour les 35 242 descriptions d'anomalies du corpus, un total de 9 623 rapports avec plus d'un marqueur annoté. Nous retrouvons 4341 rapports avec plusieurs segments de la même catégorie. Concernant les rapports avec au moins deux catégories différentes, nous en retrouvons 5282. Le plus grand nombre de segments par rapport est de 9 avec 3 rapports. Nous retrouvons un rapport avec 9 segments *Dégradation-Usure-Saleté*, un rapport avec 2 segments *Dégradation-*

Paires B-X / I-Y les plus fréquentes	**Nombre d'occurrences**
B-Degradation, **I-HorsSpec**	271
B-HorsSpec, I-FonctionnePas	222
B-Absence, **I-HorsSpec**	145
B-Impossible, **I-HorsSpec**	119
B-FonctionnePas, **I-HorsSpec**	95
Paires I-X / I-Y les plus fréquentes	
I-FonctionnePas, **I-HorsSpec**	185
I-Degradation, **I-HorsSpec**	132
I-HorsSpec, I-Absence	119
I-Impossible, **I-HorsSpec**	56
I-FonctionnePas, I-Absence	37

TABLE 6.13 – Paires incohérentes les plus fréquentes

Usure-Saleté, 5 segments *Dispositif qui ne fonctionne pas* et 2 segments *Configuration hors spécification* et enfin, un rapport avec 7 segments *Dispositif qui ne fonctionne pas* et 2 segments *Configuration hors spécification*.

Catégorie	**Nombre de rapports avec plusieurs segments de la même catégorie**	**% par rapport au nombre de rapports par catégorie**
Fuite d'un liquide ou d'un gaz	167	9,32
Signal qui s'est déclenché	44	5,34
Présence d'un obstacle	12	2,46
Dégradation - Usure - Saleté	713	13,51
Élément manquant	615	11,02
Configuration hors spécification	3597	23,75
Dispositif qui ne fonctionne pas	272	7,26
Action difficile ou impossible	207	7,03
État du monde	33	0,80
Total	5660	

TABLE 6.14 – Nombre de rapports avec plusieurs segments de la même catégorie

Dans la Table 6.14, nous pouvons observer le nombre de rapports avec plusieurs segments de la même catégorie. Nous pouvons observer naturellement un nombre très élevé pour la catégorie *Configuration hors spécification* (3597) qui représentent 23,75% des rapports de

cette catégorie, soit quasiment un quart. *Dégradation-Usure-Saleté* et *Élément manquant* sont les suivants avec respectivement 13,51% et 11,02% des rapports dans lesquels on trouve un segment de cette classe. Nous avons pu voir dans le corpus de nombreux cas de rapports avec plusieurs dégradations listées. Les catégories avec les moins de cas avec plusieurs segments du même type sont *Présence d'un obstacle* et *Signal qui s'est déclenché* avec 2,46% et 5,34%. Enfin, lors de l'annotation manuelle, seul un segment est annoté dans un rapport pour la catégorie *État du monde*. Le modèle semble cependant déroger à cette règle avec 33 rapports avec plusieurs segments de cette catégorie.

La matrice de la Figure 6.12 montre que *Configuration hors spécification* est le type le plus « sociable » et montre une forte proximité avec les catégories *Élément manquant* et *Dégradation-Usure-Saleté*. Ces dernières montrent aussi une cooccurrence importante entre elles (à hauteur de 306 rapports communs) mais aussi avec les catégories *Action difficile ou impossible* et *Dispositif qui ne fonctionne pas*. Les catégories *Présence d'un obstacle* et *Fuite d'un liquide ou d'un gaz* sont celles qui partagent le moins de rapports avec les autres.

6.3.2.3 Causes et solutions par types

En appliquant le modèle appris sur le corpus entier, nous obtenons une classification du rapport et, de ce fait, une classification des champs « Cause(s) » et « Action(s) réalisée(s) » du corpus par rapport à la classe de la description. Nous pouvons donc observer si la classe d'expression de la description de l'anomalie à une influence sur ces champs.

Nous avons pu voir que dans les champs « Cause(s) » et « Action(s) réalisée(s) », nous retrouvons de nombreux rapports pour lesquels le contenu n'est pas informatif. Cela peut aller de la simple mention de « * » ou « NEANT », ou bien de la mention « FA ANNULEE » ou « RAS ». Dans la Figure 6.13, nous pouvons observer le taux de ce type de renseignement par catégorie pour le champ « Cause(s) » (l'ordre des catégories et la même que celui de la typologie). Il est intéressant d'observer que les classes les plus basses dans la typologie, comme *Action difficile ou impossible* et *État du monde*, soit celles qui fournissent le moins d'information sur le problème survenu, sont celles qui ont le taux le plus bas de causes non informatives. Le manque de description de l'anomalie requerrait donc le besoin de complétion dans le champ « Cause(s) ». Nous retrouvons aussi la catégorie *Signal qui s'est déclenché* qui est elle placée haut dans la typologie, en raison de la quantité d'information qu'elle fournit. Cependant, elle se trouve à un niveau de description différent des autres classes puisque les énoncés ne rapportent pas le problème en lui-même, mais le signalement d'un problème.

À l'inverse, les catégories *Fuite d'un liquide ou d'un gaz* et *Présence d'un obstacle*, qui offrent des descriptions fournies, ont des champs « Cause(s) « moins informatives.

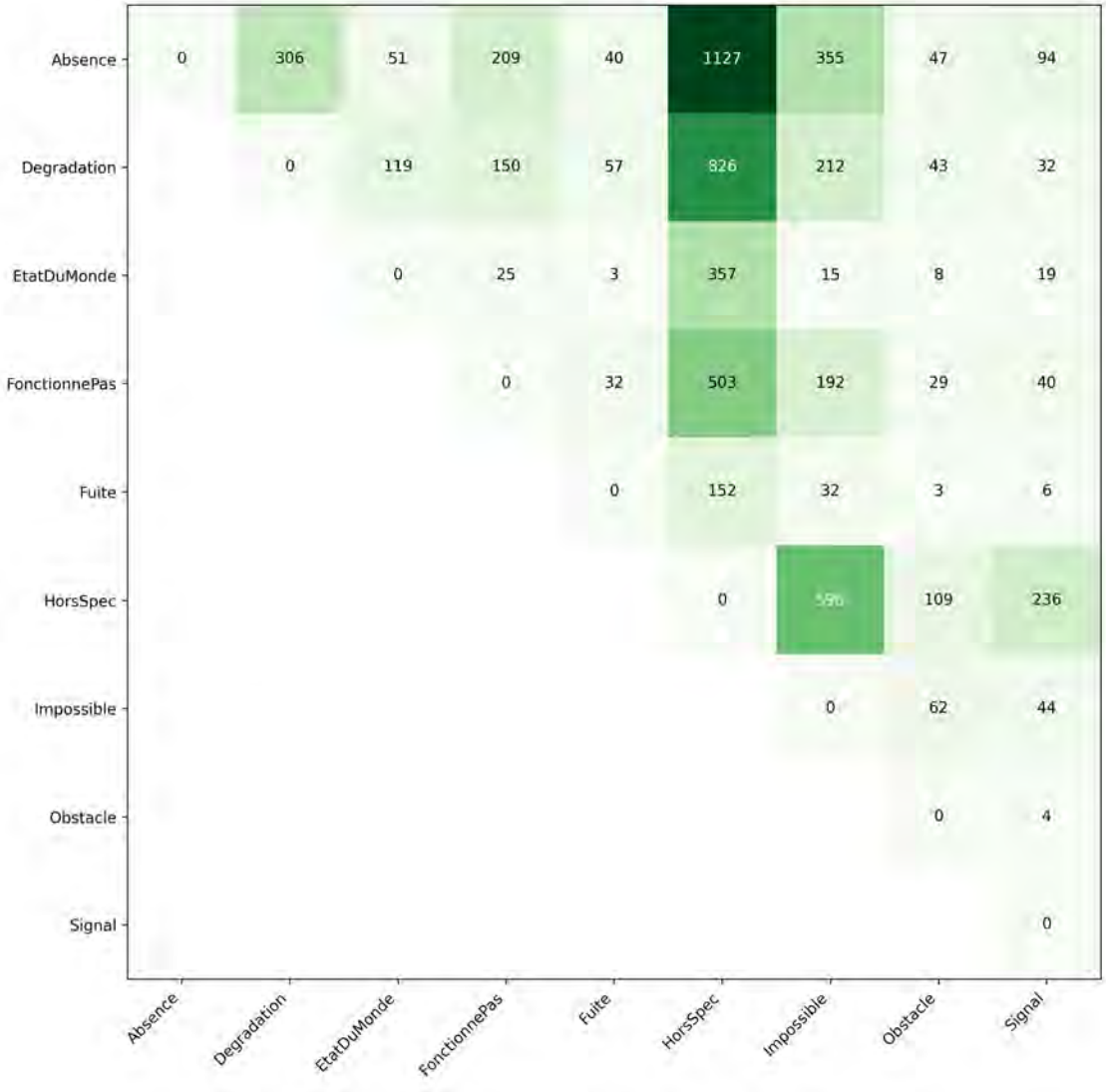


FIGURE 6.12 – Matrice de cooccurrence des segments de catégorie différents dans un rapport

Nous avons effectué la même opération pour le champ « Action(s) réalisée(s) « . Celui-ci est bien plus fréquemment renseigné. Cependant, nous pouvons observer une tendance inverse se dessiner avec un taux moins fort pour les catégories en haut de la Figure 6.14 (exception faite de *Signal qui s'est déclenché* à nouveau). Dans ce champ, il est souvent fait la mention « VIE EN L'ETAT « pour indiquer que l'anomalie n'a pas besoin d'être redressée. Nous avons donc voulu isoler ces cas. Nous retrouvons les mêmes tendances indiquant et pouvons notamment noter que les catégories ayant le moins besoin de résolutions sont *État du monde*, *Configuration hors spécification*, *Signal qui s'est déclenché* et *Élément manquant*.

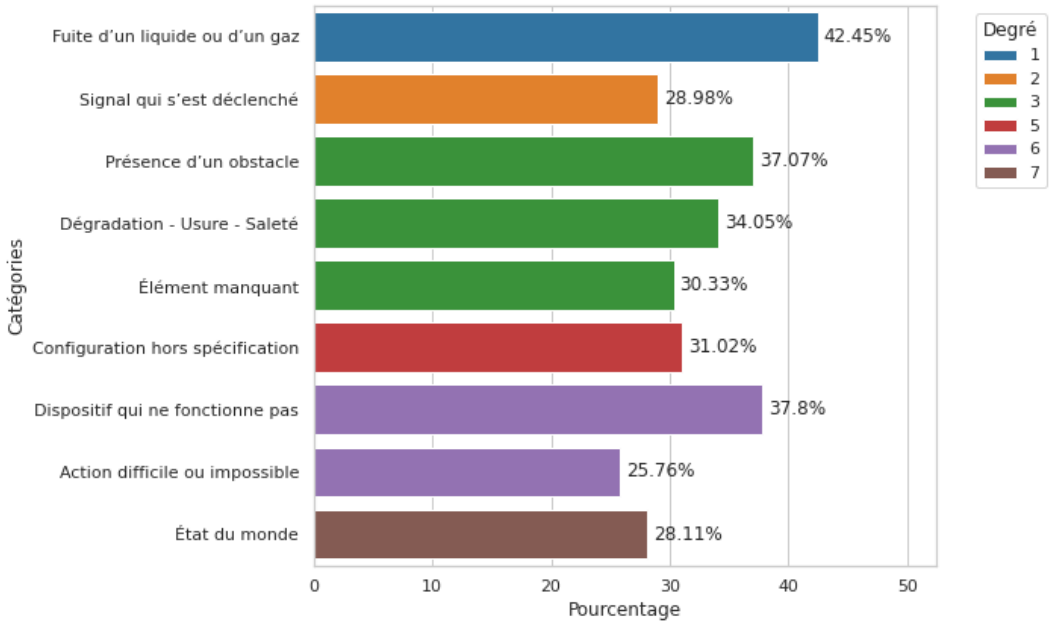


FIGURE 6.13 – Causes non-informatives par catégories et par degré

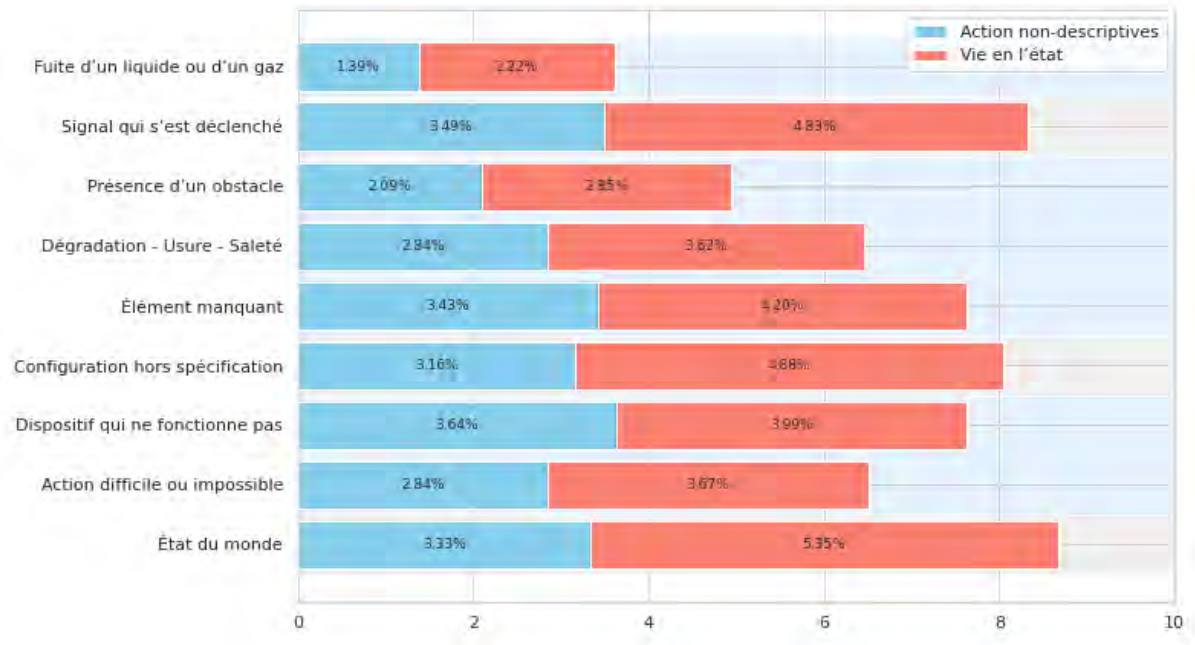


FIGURE 6.14 – Actions non-informatives et « Vie en l'état » par catégories et par degré

6.4 Conclusion

En conclusion, nous pouvons dire que les résultats de cet étiquetage automatique sont relativement satisfaisants. Les scores atteints avec la métrique *entity-type* (0,74 de F-mesure) sont loin de l'état de l'art sur des tâches de classification de séquence avec des *transformers*, mais sont satisfaisants si nous prenons en compte les particularités du corpus. En effet, malgré les marques de spécialités, le bruit et le déséquilibre dans la répartition des annotations, le modèle arrive à repérer suffisamment de régularités pour identifier et catégoriser les marqueurs.

En revanche, les incohérences repérées au niveau du schéma d'annotation viennent nuancer ces résultats. En effet, le déséquilibre des catégories dans le corpus d'entraînement semble apporter une difficulté au classifieur. En conséquence de ces anomalies, certains rapports se retrouvent à tort classé dans plusieurs catégories. D'autre part, l'extraction en vue d'obtenir des listes de marqueurs propres à chaque catégorie est aussi mise à mal, avec des marqueurs sous-divisés dans des catégories différentes.

Du point de vue des objectifs de la thèse, cette classification s'est faite en vue de préparer l'analyse approfondie de chacun des types. Les résultats ne sont pas suffisamment satisfaisants pour une automatisation totale de la chaîne de traitement que nous pourrions envisager afin de passer directement du texte à une formalisation à la TRIZ. En revanche, elle permet tout de même d'extraire des listes de marqueurs (fournit dans l'annexe B) qui, après correction des anomalies, peuvent servir à une classification des rapports sans nécessairement faire appel au modèle appris.

Concernant les suites de notre travail, nous avons choisi de focaliser notre étude sur deux des catégories de la typologie, *Fuite d'un liquide ou d'un gaz* et *Présence d'un obstacle*, respectivement aux chapitres 8 et 9. Dans le cas de la première, les résultats obtenus à partir du modèle atteignent un score parfait en F-mesure, les résultats de la classification ne posent donc pas de problèmes particuliers. Pour la seconde, nous concentrons notre étude sur un échantillon des données. De ce fait, nous avons pu sélectionner cet échantillon en s'assurant de la viabilité des données quant à leur correspondance avec la catégorie.

Dans le chapitre suivant (chapitre 7), nous poursuivons tout de même le procédé de classification automatique vu dans ce chapitre. En effet, nous explorons l'utilisation d'un LLM afin de normaliser les données, c'est-à-dire corriger les phénomènes de bruit qui l'habite. Nous évaluons ensuite l'impact de cette correction sur la tâche de classification des marqueurs.

Chapitre 7

Normalisation des données et impact sur la tâche d'étiquetage

Sommaire

7.1 Introduction

Dans ce chapitre, nous explorons la possibilité d'utiliser un LLM (*Large Language Model*) afin de normaliser les données avec lesquelles nous travaillons. En effet, nous avons pu voir lors du chapitre 4 que le corpus d'étude est bruité, et que les outils de correction automatiques usuels sont inadéquats de deux façons : d'une part à corriger toute l'étendue des phénomènes, et de l'autre à effectuer une correction sans impacter les éléments techniques du texte. L'absence de ressources supplémentaires spécialisées (comme des dictionnaires, thésaurus ou autre base de connaissances) empêche par ailleurs d'isoler le vocabulaire technique et de cibler les corrections sur le bruit (tel que nous l'avons déjà qualifié, c'est-à-dire les majuscules, l'absence d'accents...) sans déborder sur le vocabulaire technique. Ainsi, l'utilisation d'un LLM nous a paru appropriée, puisqu'elle permet, *a priori*, de cibler le processus de normalisation sur les seuls phénomènes de bruit.

Dans un premier temps, nous présentons ici le procédé de normalisation que nous avons mis en place. Nous testons différents modèles et différents prompts afin d'observer les performances d'un LLM sur cette tâche. Nous faisons une évaluation intrinsèque en comparant les résultats avec un *gold standard* manuellement corrigé.

Dans un second temps, nous reproduisons la tâche de classification des déclencheurs vue au chapitre 6, mais en faisant cette fois usage des textes normalisés. Cela nous permet de constater l'impact de la normalisation sur la tâche de classification. Nous verrons que l'impact est finalement minime, et même négatif sur les résultats obtenus.

7.2 Est-ce qu'un LLM peut normaliser des données techniques ?

Le nettoyage de données spécialisées techniques a déjà fait l'objet de différents travaux. Hodkiewicz et Ho (2016) tentent une approche à partir de règles afin de nettoyer des textes de maintenance, là où Bikaun *et al.* (2024b) mettent en place une normalisation lexicale en amont (toujours sur des textes de maintenance). L'utilisation de LLM a aussi été explorée, mais pour la correction de scories dans des textes océrisés et produit un taux d'erreur satisfaisant (Thomas *et al.*, 2024; Zhang *et al.*, 2024). Bolding *et al.* (2023) démontrent que l'utilisation de LLM pour la correction de textes bruités n'empêche pas de préserver l'intégrité sémantique du texte, et Wang *et al.* (2024) viennent confirmer ces résultats tout apportant une certaine nuance. Les performances du modèle dépendraient du type de bruit et du modèle utilisé. Dans cette lignée, nous reprenons l'utilisation d'un LLM génératif et certaines méthodes d'évaluation (le *Character Error Rate* utilisé pour les textes océrisés) afin d'observer les performances sur notre corpus.

7.2.1 Méthode employée

Afin de réaliser cette étude, nous avons procédé à une analyse comparative de trois LLMs sur la même tâche. Ces modèles sont généralistes, multilingues, *instruction-tuned*. Ce dernier signifie que le modèle a été entraîné pour répondre à des instructions formulées en langage naturel. Les modèles que nous avons choisis sont aussi disponibles en *open-weight*, c'est-à-dire que nous avons accès aux paramètres du réseau de neurones, et nous n'avons effectué aucune opération de *fine-tuning*. Nos expérimentations ne portent que sur le *prompt*. En regard des contraintes de confidentialité des données, nous avons utilisé les « petites » versions de ces modèles (soit entre 7 et 8 milliards de paramètres). Celles-ci peuvent être lancées localement sans nécessiter l'usage d'une infrastructure conséquente. Les modèles

en question sont les suivants :

— Meta-Llama-3.1-8B-Instruct (Grattafiori *et al.*, 2024) ;
— Meta-Llama-3-8B-Instruct (Grattafiori *et al.*, 2024) ;
— Mistral-7B-Instruct-v0.3 (Jiang *et al.*, 2023).

Pour interagir avec ces modèles, nous avons fait usage de l'outil *Ollama*[65].

Nous avons aussi testé quatre versions du même *prompt*. Dans chacune des versions, nous ajoutons des informations qui viennent préciser la demande. Cela nous permet de comparer l'efficacité du modèle en fonction du degré d'information dont il dispose. Dans la Table 7.1, nous pouvons observer le *prompt* en question. Ainsi, dans les différents *prompts*, nous retrouvons les informations suivantes :

— Prompts 1 à 4 : le rôle que doit prendre le LLM, la tâche à réaliser, et une brève description de cette tâche ;
— Prompts 2 à 4 : l'ajout du but de la tâche (le texte sera traité automatiquement par la suite) ;
— Prompts 3 et 4 : l'ajout d'une liste des phénomènes à corriger ;
— Prompt 4 : deux exemples (*few-shot prompting*).

Les *prompts* ont été rédigés en anglais, malgré des données en français, suivant l'étude de Jin *et al.* (2024). Des études ultérieures ont cependant montré qu'il peut être plus efficace d'utiliser la langue des données, en fonction de la langue, de la tâche et du modèle utilisé.

La température a été fixée à zéro afin de limiter le non-déterminisme des modèles, ce qui empêche le modèle de faire preuve de « créativité ». Lors du calcul du token suivant, c'est systématiquement le plus probable qui est choisi. Ainsi, il n'est pas nécessaire d'effectuer plusieurs lancements pour cette tâche. Alizadeh *et al.* (2024) recommandent cette pratique pour de l'annotation de texte par LLM.

Chaque rapport est traité individuellement. Le modèle est prompté à chaque fois sur une description d'anomalie complète, pour éviter les généralisations, résumés ou autres comportements indésirables.

7.2.2 Résultats obtenus

Afin d'évaluer les résultats obtenus, nous effectuons une première évaluation intrinsèque à l'aide du *Character Error Rate* (CER). Cette métrique permet de calculer, caractère par caractère, les différences entre deux versions d'un même texte. Elle est couramment utilisée pour évaluer une sortie d'OCR (Optical Character Recognition) ou d'ASR (Automatic Speech Recognition). Chaque différence (casse, accents, espaces...) est prise en compte dans le calcul

65. https://ollama.com/

Inclus dans la version du prompt	**Texte**
1,2,3,4	You are a trained linguist working with maintenance operators. Your task is to correct sentences written in French by these operators. These texts describe problems occurring during the maintenance of a rocket. You are correcting these texts because they contain a lot of noise. You must write a standardized version of these texts without modifying, reformulating, or changing any words. Do not alter the vocabulary.
2,3,4	You need to clean these text because they will be automatically processed afterward.
3,4	Here is a list of the different phenomenons to correct you may encounter : - missing spaces and punctuation - mispelled words - the whole text in uppercase - missing accents. Even if you encounter an unfamiliar word, keep it as it is. When displaying your answer, write only the corrected version of the sentence without adding line breaks, additional information, explanations, or notes.
4	Here are two examples. The text "CORROSION LEGERE SUR OVM50005CORROSION PLUS IMPORTANTE SUR OVM5006 (VANNE ALIM PISCINE)" becomes "Corrosion légère sur OVM50005. Corrosion plus importante du OVM5006 (vanne ALIM piscine)." The text "POULIE DE RENVOI SUR CAISSSON LBS LH2 GRIPPEE SUR SON AXE." becomes "Poulie de renvoi sur caisson LBS LH2 grippée sur son axe.".
1,2,3,4	Here is the text to rectify : [text inserted]

TABLE 7.1 – Prompt

du score. Ici, les deux versions que nous comparons sont la sortie du modèle et un *gold standard* constitué des rapports corrigés manuellement.

Pour effectuer l'évaluation, nous avons procédé à une correction manuelle de 50 rapports. Cette correction vise à normaliser les phénomènes de bruits et ne touche donc pas aux éléments techniques du texte, sauf pour en corriger la casse et les accents éventuels. Les rapports ont été choisis de façon aléatoire parmi le *dataset* de 1050 rapports utilisé pour la tâche de classification. Le seul critère lors de cette sélection était d'obtenir dans l'échantillon une représentation de tous les phénomènes de bruits que nous avons jusque-là rencontrés, listés dans la version 3 du prompt visible dans la Table 7.1. Quelques cas ont nécessité de prendre une décision quant aux règles appliquées comme garder ou non les abréviations comme « hors spec. » par exemple, que nous avons choisi de garder tels quels. La question s'est aussi posée de certains signes comme « on » ou « off » à garder ou non en majuscules. Ces questions influencent les résultats obtenus car la métrique utilisée pour l'évaluation se

situe au niveau du caractère et que les modèles peuvent avoir une logique de correction qui diffère de nos règles.

Nous avons pu calculer, pour chaque rapport, la distance entre la version produite par la combinaison modèle / prompt. La moyenne des résultats pour les 50 rapports est donnée dans la Table 7.2. Le CER représente le taux de caractères corrects dans la version corrigée, donc entre 0 et 1. Plus le score est bas, plus les deux textes sont proches. À l'inverse, plus le score est haut, plus les textes diffèrent. Deux textes complètement différents obtiendraient donc un score de 1. Nous pouvons observer que le score entre le *gold* et les rapports originaux atteint 0,46, ce qui montre une différence élevée entre les deux. Les résultats obtenus pour les rapports normalisés sont tous inférieurs, avec 0,37 maximum obtenu pour la combinaison de Llama3 et le prompt 2. En revanche, ce score reste élevé. Les résultats obtenus pour les prompts 1 et 2 sont tous élevés et montrent que ces prompts ont peu d'efficacité dans une tentative de normalisation, quel que soit le modèle utilisé. En revanche, l'ajout de la liste des phénomènes à normaliser entraîne une nette amélioration des résultats puisque les prompts 3 et 4 obtiennent des scores entre 0,08 (soit le meilleur score obtenu avec Llama3 / prompt 4) et 0,12 (Llama 3.1 / prompt 3). Nous pouvons aussi noter que l'ajout des exemples dans le prompt 4 permet une légère amélioration des résultats pour les modèles Llama 3.1 et Mistral.

Dataset	**Character Error Rate**
Rapports originaux	0.46
Llama 3 prompt 1	0.26
Llama 3 prompt 2	0.37
Llama 3 prompt 3	**0.10**
Llama 3 prompt 4	**0.10**
Llama 3.1 prompt 1	0.30
Llama 3.1 prompt 2	0.32
Llama 3.1 prompt 3	0.12
Llama 3.1 prompt 4	**0.08**
Mistral prompt 1	0.35
Mistral prompt 2	0.33
Mistral prompt 3	**0.10**
Mistral prompt 4	**0.09**

TABLE 7.2 – CER pour la normalisation des rapports

Dans la Table 7.3, nous pouvons observer un exemple de rapport normalisé avec Llama 3 pour le rapport « SYNTHESE HSY062 A OFF ATTENDU A ON ». Dans cet exemple, nous

pouvons voir que les prompts 1 et 2 produisent des hallucinations avec en sortie les phrases « Synthèse HSY062 : À l'occasion de l'arrêt attendu à bord. » et « Synthèse HSY062 : Analyse d'anomalie - À l'occasion de l'offre attendue à ce moment-là. ». Nous pouvons noter dans la sortie du prompt 2 la présence du terme « anomalie », malgré le fait que ce terme soit exempt du *prompt*. Nous mentionnons uniquement le terme « problème ». Nous pouvons toutefois observer que le modèle corrige effectivement les phénomènes listés, c'est-à-dire ici les majuscules et les accents, sans modifier le vocabulaire technique (« HSY062 »).

Les *prompts* 3 et 4 obtiennent cependant des résultats beaucoup plus satisfaisants. Les seules différences avec le *gold* proviennent de l'utilisation de majuscules pour « ON » et « OFF », et dans le cas du *prompt* 3, l'ajout du « : » et d'une minuscule dans le code « HSy062 ».

Dataset	**Sortie**	**CER**
Rapport original	SYNTHESE HSY062 A OFF ATTENDU A ON.	0.35
Rapport gold	Synthèse HSY062 à OFF attendue à ON.	
Llama3 prompt 1	Synthèse HSY062 : À l'occasion de l'arrêt attendu à bord.	0.47
Llama3 prompt 2	Synthèse HSY062 : Analyse d'anomalie - À l'occasion de l'offre attendue à ce moment-là.	0.64
Llama3 prompt 3	Synthèse HSy062 : À off attendu à on.	0.20
Llama3 prompt 4	Synthèse HSY062 à off attendu à on.	0.13

TABLE 7.3 – Exemples de correction

L'évaluation de cette tâche, à travers le calcul du CER, montre des résultats satisfaisants quant à la normalisation par l'utilisation de LLM, pour peu que le *prompt* soit bien défini. La précision des phénomènes que nous souhaitons corriger et les exemples apportent une amélioration notable avec environ 0,20 points de différence. Cela laisse suggérer que dans un cas comme le nôtre, c'est-à-dire lorsque les outils habituels de nettoyage sont inappropriés en raison de la spécificité des données et que des ressources additionnelles relatives à ces données ne sont pas disponibles, l'utilisation de LLM est une option efficace. L'utilisation de *spellChecker* sur les mêmes données a produit un CER de 0,21. Par ailleurs, ces modèles de « petite taille » n'ont en aucun cas été *fine-tunés* sur les données et ont pu être utilisés tels

quels. Cette tâche ne requiert donc pas d'infrastructure conséquente, ni de modifications techniques au niveau du modèle.

Le CER pose tout de fois certaines limites dans l'évaluation de la correction. En effet, le score n'accorde pas de poids différent en fonction du type d'erreur. Ne pas appliquer la bonne casse compte autant que de supprimer un mot s'ils impactent le même nombre de caractères. Pourtant, ces fautes ne pèsent pas le même poids dans le sens de l'énoncé. Elle ne permet pas non plus de compter différentes variantes comme correctes, comme le fait de mettre ou non des majuscules à « on » et « off » dans l'exemple 7.3.

Dans la seconde partie de ce chapitre, nous explorons donc l'utilisation de ces *datasets* pour la tâche de repérage des déclencheurs présentée dans le chapitre 6. Cela nous permet d'envisager cette normalisation sous deux aspects, à la fois pour obtenir une classification plus efficace du corpus, mais aussi comme évaluation de la normalisation de façon extrinsèque. Nous répondons à la question de la pertinence d'effectuer une normalisation des données pour la tâche d'étiquetage de séquence dans ce type de corpus.

7.3 Son impact sur une tâche d'étiquetage

Dans cette section, nous revenons sur la tâche d'étiquetage des marqueurs, mais cette fois à partir des *datasets* normalisés. Nous avons utilisé les mêmes 1050 rapports qui ont été annotés manuellement afin de repérer et typer les marqueurs lexicaux. Parmi les douze combinaisons de modèles/prompts utilisés pour la normalisation, nous avons sélectionné les cinq les plus performantes (en gras dans la Table 7.2). Parmi les modèles utilisés pour l'étiquetage des marqueurs, nous avons sélectionné *bert-base-multilingual-uncased* et *camembert-large*, qui sont les deux modèles ayant obtenu les meilleurs scores sur les données non corrigées.

7.3.1 Préparation des données

Une des difficultés qui s'est posée est que les *datasets* normalisés peuvent produire des décalages par rapport aux caractères présents dans le rapport original. Les textes sont modifiés, mais nous considérons que leur interprétation reste la même et, par conséquent, les annotations manuelles aussi. Les modifications sont censées avoir visé uniquement le bruit de surface et non pas le sens véhiculé par l'énoncé. De ce fait, les corrections sont censées être superficielles. En revanche, elles peuvent impliquer un décalage dans le nombre de caractères du rapport, ce décalage pouvant varier en fonction des positions. En effet, certains caractères peuvent être sujets à une suppression ou un ajout. Dans la figure 7.1, nous pouvons observer un exemple de ce cas. En rouge, nous indiquons les endroits où des

caractères ont été supprimés dans le rapport original. En bleu, nous pouvons observer les endroits ou des caractères ont été ajoutés dans le texte normalisé. Par exemple, « outillage » devient « équipement ». Nous observons donc un ajout de 1 caractère dans cette zone du rapport. Au total, on observe une différence de 2 caractères au total pour ce rapport, avec 9 caractères supprimés et 11 caractères ajoutés. La superposition des annotations avec les *subwords* issus de la tokenization des modèles se fait en fonction des index de ces deux éléments. De ce fait, la projection des annotations manuelles directement sur les nouveaux *datasets* est impossible, au risque de créer des annotations qui ne couvrent pas réellement le marqueur. En effet, la projection que nous effectuons pour faire correspondre les *subwords*, une fois les rapports tokenizés par un modèle, et les annotations manuelles est basée sur les *offsets*. De ce fait, l'ajout ou la suppression de caractères dans la version normalisée du rapport peut créer un décalage avec les *offsets* des annotations. Dans cet exemple, le marqueur « IMPOSSIBLE DE REALISER » deviendrait « possible de réaliser l », inversant totalement sa signification. La projection se doit donc, au minimum, de prendre en compte les décalages au niveau des zones dans lesquelles un changement a eu lieu, pour en définitive obtenir des annotations qui correspondent au texte initialement sélectionné.

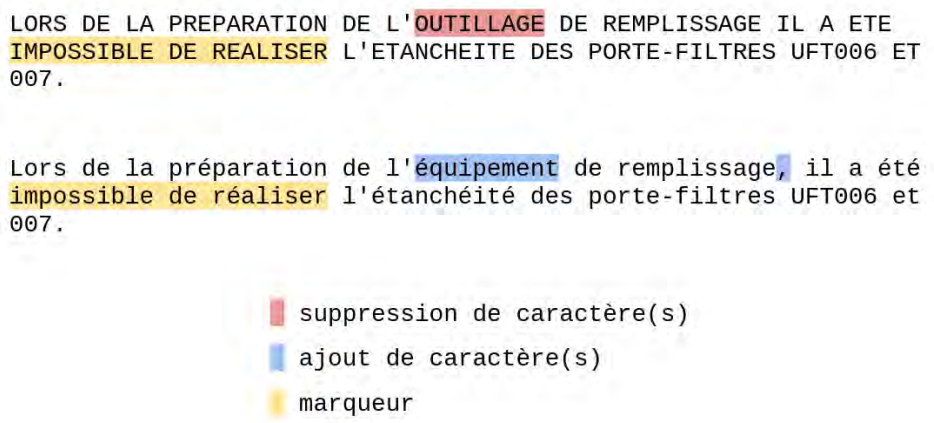


FIGURE 7.1 – Exemple de décalage entre un rapport et sa version normalisée

Afin de résoudre ce problème, nous avons propagé les annotations en faisant usage de l'utilitaire GNU wdiff[66]. Celui-ci permet d'isoler les séquences de mots ayant subi une modification dans la chaîne de caractère. Dans l'exemple suivant, nous pouvons voir en premier lieu la description originale (passée en minuscule), en second lieu une version redressée et en dernier, la sortie de wdiff :

1. sur les 4 piles les 3 temoins de temperature 25 degre c sont rouges.les temoins 37 degre c sont intacts.les conditions climatiques ne sont toujours pas rappelees

66. https ://www.gnu.org/software/wdiff/

sur la caisse.

2. sur les 4 piles, les 3 thermomètres de température à 25°C sont rouges. Les thermomètres 37°C sont intacts. Les conditions climatiques ne sont toujours pas rappelées sur la caisse. [67]

3. sur les 4 [-piles-] {+piles,+} les 3 [-temoins-] {+thermomètres+} de [-temperature 25 degre c-] {+température à 25°C+} sont [-rouges.les temoins 37 degre c-] {+rouges. Les thermomètres 37°C+} sont [-intacts.les-] {+intacts. Les+} conditions climatiques ne sont toujours pas [-rappelees-] {+rappelées+} sur la caisse.

L'outil récupère toutes les séquences ayant été modifiées. Les séquences du rapport original ayant subi une modification, qui ont donc été supprimées, sont signifiées par « [--] », et les séquences du rapport normalisé qui viennent en remplacement par {+ +}. La sélection est plus ou moins longue puisqu'elle s'arrête au niveau des espaces, une fois qu'une chaîne de caractère non modifiée se présente. Ainsi, pour `[-rouges.les temoins 37 degre c-]`, le segment fait cette longueur puisque l'absence d'espace après le « . » fait considérer « `rouges.les` » comme un seul mot qui subi une modification avec l'ajout de l'espace en question. Le mot suivant « témoins » est aussi modifié ainsi que la séquence qui suit. Le terme « sont » est lui intact et ne fait donc pas partie de la sélection.

Pour effectuer les correspondances, nous calculons l'alignement entre les offsets de la chaîne originale et ceux de la chaîne d'arrivée. Pour les segments inchangés, nous calculons le décalage, et pour les segments modifiés nous faisons une homothétie. Une fois *wdiff* appliqué, nous récupérons les index de début et de fin de la séquence originale, ainsi que le nombre de caractères de décalage, et calculons un coefficient de décalage pour la séquence. Ainsi, pour une séquence donnée, nous effectuons une extrapolation des correspondances internes en se basant sur les positions des extrémités. Cela nous permet de faire correspondre chaque *offset* de départ et d'arrivée des annotations [68].

Ainsi, nous obtenons en sortie un *dataset* normalisé avec les mêmes annotations que celles effectuées sur le *dataset* des rapports originaux, obtenues grâce à la correspondance des *offsets* de départ et d'arrivée entre la version originale et la version normalisée du rapport. Une fois ces éléments obtenus, nous pouvons donc effectuer la même opération de *fine-tuning* des modèles *transformers* afin de les entraîner à la tâche de reconnaissance des marqueurs.

67. Cet exemple n'est pas issu des *datasets* normalisés ayant obtenu les meilleurs scores.
68. Nous fournissons le pseudo-code pour cet algorithme en annexe C

7.3.2 Présentation des résultats

La tâche d'apprentissage opérée est la même que celle présentée dans le chapitre 6. Les mêmes rapports constituent les corpus d'entraînement et de tests, mais dans des versions normalisées du *dataset*. Afin d'obtenir une *baseline* pour nos résultats, nous avons aussi appliqué l'outil *spellChecker* sur le *dataset* original. Nous avons déjà mentionné cet outil au chapitre 4 et nous avons pu observer que son application sur nos données produit un résultat peu satisfaisant quant à la correction du corpus et tend à ajouter du bruit supplémentaire. Ainsi, l'objectif en utilisant cet outil est de montrer les résultats obtenus lors de l'utilisation d'un corpus encore plus bruité et avec une sémantique dégradée due à une mauvaise correction (ou une hypercorrection) de certains termes.

Pour évaluer les résultats, nous avons repris la métrique *entity-type* déjà présentée au chapitre précédent, qui permet de calculer les scores dans les cas où un chevauchement non strict existe entre le *gold* et la sortie du modèle. Les résultats sont présentés dans la Table 7.4 avec en gras les meilleurs score de F-mesure obtenus par version du *dataset*/modèle.

Classifieur	**bert-base-multilingual-uncased**			**camembert-large**		
Dataset	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
Rapports originaux	0.72	0.77	**0.74**	0.71	0.79	**0.74**
Llama 3 prompt 3 (liste)	0.60	0.67	0.64	0.66	0.78	**0.71**
Llama 3 prompt 4 (exemples)	0.63	0.73	**0.68**	0.58	0.74	0.65
Llama 3.1 prompt 4 (exemples)	0.58	0.66	0.62	0.64	0.77	**0.70**
Mistral prompt 3 (liste)	0.57	0.66	0.62	0.63	0.73	**0.68**
Mistral prompt 4 (exemples)	0.62	0.67	0.64	0.62	0.74	**0.68**
spellChecker	0.54	0.57	0.55	0.66	0.73	**0.69**

TABLE 7.4 – Scores de la tâche d'étiquetage de séquences des différentes versions corrigées au niveau des entités (métrique *entity-type*)

Les résultats obtenus montrent que l'utilisation des *datasets* normalisés ne permet pas d'obtenir de meilleurs résultats pour cette tâche. Les résultats par catégories ne montrent pas de différence notable. En effet, le score obtenu pour le *datasets* composés des rapports originaux, sans modifications, atteint 0,74 de F-mesure pour les deux modèles que sont *bert-base-multilingual-uncased* et *camembert-large*. Pour les *datasets* normalisés, les meilleurs scores sont entre 0,68 et 0,71. Nous observons donc une légère baisse des résultats. De même, le *dataset* « corrigé »par *spellChecker* atteint un score de 0,69 avec le modèle *camembert-large*,

et uniquement 0,55 avec *bert-base-multilingual-uncased*.

Cette tendance se confirme aussi avec les *datasets* normalisés. Le modèles *camembert-large* obtient de meilleurs résultats que *bert-base-multilingual-uncase*. L'hypothèse la plus probable quant à cette différence est que les modèles CamemBERTs ont été appris spécifiquement pour le français. Ils sont donc plus aptes à gérer la casse correctement et surtout la présence d'accents. Le modèle *bert-base-multilingual-uncase* ne prend quant à lui pas en compte la casse et convertit tout en minuscules. Concernant les accents, la sortie du *tokenizer* de ce modèle efface aussi les accents. Ainsi, leur présence n'a pas d'impact sur le modèle et justifie probablement en partie le fait qu'il obtienne le meilleur score sur le corpus des rapports originaux.

Concernant la baisse des résultats par rapport au corpus des rapports originaux, plusieurs hypothèses peuvent être émises :

1. la normalisation entraîne une perte d'informations sémantiques ;
2. la tâche n'est pas sensible à la normalisation.

Pour le premier point, nous avons observé les cas pour lesquels la catégorie du marqueur diffère pour une même FA au sein du corpus de test. Dans la Figure 7.2, nous pouvons voir un exemple en ce sens. Le terme « défectueux » a été remplacé par « défaut ». Or, les deux n'appartiennent pas à la même catégorie. Le premier, « défectueux » renvoie à la catégorie *Dispositif qui ne fonctionne pas*, alors que le second renvoie à la catégorie *Configuration hors spécification*. En cela, le modèle étiquette correctement le marqueur « défaut », mais celui-ci ne devrait pas être présent et n'existe qu'à cause de la normalisation. Si ce type de modification peut expliquer une baisse des résultats, ils ne sont cependant pas présents massivement dans le corpus de test d'après nos observations. Ils n'expliquent par ailleurs pas pourquoi le *dataset* produit avec *spellChecker* obtient des résultats comparables aux autres *datasets* pour le modèle *camembert-large*, malgré l'inaptitude de l'outil à corriger le texte correctement.

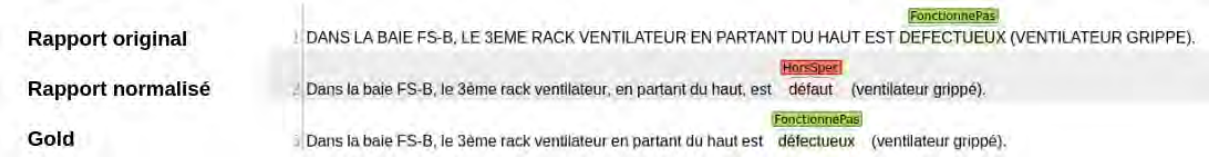


FIGURE 7.2 – Exemple de normalisation qui affecte l'étiquetage

La deuxième hypothèse est que la normalisation n'a pas d'impact à cause des spécificités de la tâche. En effet, les marqueurs annotés ne contiennent pour la majorité pas de vocabulaire technique spécifique. Le phénomène qui les impacte le plus est l'absence d'accents.

Les marqueurs font majoritairement partie du vocabulaire commun (*fuite, ne fonctionne pas, cassé...*), et nous pouvons supposer que cela suffit pour que le modèle puisse les appréhender malgré les scories que nous pouvons y trouver. Pour le modèle *bert-base-multilingual-uncased*, les accents ne figurent même plus une fois passé l'étape de *tokenisation*. Par ailleurs, la tâche concerne l'étiquetage des marqueurs. Si un même marqueur est présent à plusieurs reprises, le modèle est censé pouvoir reproduire cet étiquetage à partir de ce peu d'exemples. Cette hypothèse paraît plus cohérente avec les résultats obtenus par le *dataset* produit par *spellChecker*.

En conclusion, nous dressons un bilan mitigé de cette étude. Les scores de la normalisation montrent des résultats satisfaisants. Les meilleurs scores atteints par le CER sont bas et l'analyse qualitative est correcte. Par ailleurs, cette évaluation est stricte et les scores sont peut-être à nuancer puisque certaines modifications, bien que non parfaitement fidèles au *gold*, restent du domaine de l'acceptabilité. L'utilisation de LLM pour normaliser un texte technique semble donc avoir fait ses preuves.

En revanche, l'utilisation des *datasets* normalisés ne nous a pas permis d'obtenir une classification du corpus plus efficace. En effet, les résultats de l'étiquetage automatique des marqueurs sont moins élevés qu'avec l'utilisation du *dataset* bruité. Plusieurs hypothèses peuvent être émises à ce propos. De façon à explorer plus avant cet effet, nous considérons que la réalisation d'autres tâches pourrait entraîner des résultats différents. Des tâches comme l'étiquetage morpho-syntaxique, ou l'analyse syntaxique en dépendances reposent davantage sur ces informations et un résultat plus favorable pourrait en ressortir.

Quatrième partie

Analyse fine du contenu linguistique pour une modélisation : méthode, traitements et limites

Chapitre 8

Analyse inductive du phénomène de la fuite

Sommaire

Dans ce chapitre, nous souhaitons focaliser notre étude sur la catégorie *Fuite d'un liquide ou d'un gaz*. L'objet de ce travail est d'effectuer une analyse fine de l'expression du problème dans les rapports appartenant à cette catégorie. Le choix de cette catégorie s'est fait, dans un premier temps, par l'exclusion de certaines classes peu adaptées à un traitement par TRIZ. En effet, des classes comme *Action difficile ou impossible* ou bien *Dispositif qui ne fonctionne pas* présentent un degré d'informations exprimées limité. Certaines classes comme *Signal qui s'est déclenché* et *Configuration hors spécification* sont trop abstraites dans l'expression du problème en question. Dans un second temps, la catégorie *Fuite d'un liquide ou d'un gaz* possède l'avantage d'avoir des marqueurs très facilement identifiables en raison de leur

faible variabilité et propose, comme nous pourrons le voir, une certaine régularité dans sa structure d'expression. Nous aurons l'occasion d'explorer une autre catégorie de la typologie dans le chapitre suivant.

Cette classe pose une première problématique d'approche, puisque, comme nous le verrons dans la première section de ce chapitre, le *frame* proposé par *Framenet* présente des limites et ne peut s'appliquer tel quel à nos données et nos besoins. Ainsi, nous effectuons une analyse du contenu de nos rapports en combinant trois approches, qui nous ont permis de faire émerger un *frame* de la fuite dans un environnement technique. Nous finissons par l'étude de l'adéquation entre ce *frame* et le *vépole* cible de cette catégorie.

8.1 Adéquation du *frame* issu de *Framenet*

Les rapports d'anomalies qui font le constat de la fuite possèdent très peu de marqueurs [69], à savoir *fuite(s), fuyard(es)* et *écoulement* (ce dernier étant peu représenté, à hauteur de 1,8%, dans les données de cette catégorie, nous focalisons notre étude sur les deux premiers). En revanche, nous pouvons trouver une certaine variété dans leur expression et dans les informations qu'ils véhiculent. Dans la Figure 8.1, nous pouvons justement observer cette diversité. Chaque exemple constitue une description complète extraite des données étiquetées automatiquement avec la classe *Fuite d'un liquide ou d'un gaz*. Le premier exemple montre un rapport avec pour seule information, en plus du marqueur, une localisation introduite par la préposition « sur ». Dans le deuxième exemple, nous pouvons trouver une description légèrement plus précise avec la mention d'une fuite « interne » (c'est-à-dire que la fuite a lieu à l'intérieur du composant) et d'une localisation qui semble plus précise mais dont les composants sont peu compréhensibles pour un non-expert. Les deux derniers exemples montrent des cas pour lesquels la description est riche en information avec des énoncés longs et qui impliquent différents composants du système, des notions de quantité (« 26bar » et « legere »), des outils de détection (« detecteur electronique et mille bulles »), mais aussi, dans le dernier exemple des détails sur le contexte lors de la fuite « apres montee en pressurisation... ». Cette diversité dans l'expression de la fuite nous amène à questionner la présence des informations que nous souhaitons extraire. Afin de nous guider dans ce repérage, nous avons souhaité pouvoir associer un schéma standard pour regrouper ensemble les informations de même nature, ce qui nous a conduit aux *frames* de *FrameNet*.

69. La liste des marqueurs extraient de l'étiquetage automatique peut être retrouvée en Annexe B

FUITE SUR SM HVC0199.

FUITE INTERNE DU ROBOTEUR SUR DEVERSEUR YDV560.

GF8 CEG BAF. FUITE R22 SUR SOUDURE ENTRE BOUTEILLE LIQUIDE ET SOUPAPE DE SECURITE (RECHERCHE DE FUITE AVEC DETECTEUR ELECTRONIQUE ET MILLE BULLES.

APRES MONTEE EN PRESSURISATION A 26bar LORS DU CTRL AU SNOOP, ON OBSERVE UNE LEGERE FUITE AU NIVEAU DU RACCORDEMENT DU FLEXIBLE HFX310 (FDV3) SUR LE PANNEAU DE PRESSURISATION.

FIGURE 8.1 – Exemples de descriptions d'anomalies de la catégorie *Fuite d'un liquide ou d'un gaz*

Le *frame* correspondant à celui de la catégorie *Fuite d'un liquide ou d'un gaz* que nous avons sélectionné est : « *Fluidic motion*[70] ». Dans ce *frame*, nous retrouvons le déclencheur « leak » (tr. fuire) parmi les unités lexicales qui lui sont associées. La définition qui lui est donnée dans *FrameNet* est la suivante : « Dans ce cadre, un Fluide se meut d'une Source vers une Destination en suivant un Trajet ou dans une Zone (« *In this frame a Fluid moves from a Source to a Goal along a Path or within an Area.* »). Pour l'unité lexicale « leak », la définition donnée est : « [un fluide] entre ou sort à travers un trou ou une fissure d'une certaine manière (« *pass in or out through a hole or crack in such a way* »). Parmi les autres unités lexicales de ce *frame*, nous retrouvons le verbe « couler » (*flow*) ou bien encore « se répandre » (*spill*).

Le schéma d'annotation avec la liste des éléments du *frame* est en Figure 8.2. Comme nous pouvons le voir, ce cadre implique quinze *frame elements*, dont cinq sont des éléments *core*, c'est-à-dire des éléments noyaux, propres au *frame*. Ce *frame* n'est cependant pas entièrement adapté au cas de fuites telles qu'elles sont exposées dans notre corpus, notamment en rapport au domaine spécifique dans lequel ces textes sont rédigés.

Dans la Figure 8.3, nous pouvons observer d'une part des exemples annotés issus de *FrameNet* pour l'unité lexicale *leak*, et de l'autre des exemples issus des rapports d'anomalie pour lesquels nous avons appliqué le même schéma d'annotation.

Dans les exemples issus de *FrameNet*, nous pouvons voir qu'ils dépeignent des scènes de la vie courante. Dans les deux exemples, nous retrouvons les *frame elements Fluid* (en violet) et *Source* (en vert), qui sont deux éléments noyaux (*core*) du cadre. Pour l'élément *Source*, nous pouvons voir que plusieurs niveaux de spécification quant à l'emplacement de la source sont agrégés dans une seule annotation. Dans le second exemple se trouve aussi le *frame element Path* (en bleu vif) qui précise la trajectoire du mouvement du fluide.

Dans les exemples issus des rapports, nous avons tenté d'appliquer le schéma d'annotation proposé par *FrameNet* pour le *frame Fluidic_Motion*. Parmi les éléments noyaux, nous

70. https://framenet.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Fluidic_motion&banner=

Frame Element	Core Type
Area	Core
Configuration	Extra-Thematic
Depictive	Extra-Thematic
Distance	Peripheral
Duration	Peripheral
Explanation	Extra-Thematic
Fluid	Core
Goal	Core
Manner	Peripheral
Path	Core
Place	Peripheral
Result	Extra-Thematic
Source	Core
Speed	Peripheral
Time	Peripheral

FIGURE 8.2 – Schéma d'annotation de *Fluidic_motion*

Exemples issus de FrameNet :

. Firemen from Thornaby and Stockton swilled away an amount of petrol LEAKING from a car in Yarm Street , Stockton .

Water is LEAKING through our ceiling from the waste outlet of the shower tray in the bathroom above .

Exemples issus des rapports :

GF8 CEG BAF. FUITE R22 SUR SOUDURE ENTRE BOUTEILLE LIQUIDE ET SOUPAPE DE SECURITE (RECHERCHE DE FUITE AVEC DETECTEUR ELECTRONIQUE ET MILLE BULLES.

APRES MONTEE EN PRESSURISATION A 26bar LORS DU CTRL AU SNOOP, ON OBSERVE UNE LEGERE FUITE AU NIVEAU DU RACCORDEMENT DU FLEXIBLE HFX310 (FDV3) SUR LE PANNEAU DE PRESSURISATION.

FUITE DE LIQUIDE DE REFROIDISSEMENT (LEGER SUINTEMENT) CONSTATEE AU NIVEAU DE LA PURGE D'AIR DU CIRCUIT DE REFROIDISSEMENT DE GE1 DU GST2.

FIGURE 8.3 – Exemples de *fuites* annotés avec *FrameNet*

retrouvons le *Goal* (en bleu clair dans les trois exemples) et le *Fluid* (en violet dans l'exemple 3). Pour l'élément *Goal*, de même que pour l'élément *Source* dans les exemples précédents, nous retrouvons une grande séquence annotée avec différentes précisions quant à l'emplacement regroupées dans un seul bloc. Par ailleurs, lors de l'annotation de ces exemples, nous avons choisi d'utiliser l'élément *Goal*. Cependant, cette annotation est à remettre en question puisque les rapports manquent de clarté, du moins pour un non-expert. En effet, dans les exemples issus de *Framenet*, la préposition *from* évoque avec certitude la notion

de provenance, impliquant que l'élément qui suit (« a car » ou « the waste outlet ») est bien la source de la fuite. En revanche, pour les rapports, les prépositions employées sont *sur* et *au niveau de*. Ces deux prépositions apportent une ambiguïté quant à la qualification des informations de localisations énoncées. La différenciation entre l'origine et le point d'arrivée de la fuite est moins claire. De plus, le procédé de création de ces rapports (que nous avons vu dans le chapitre 4) influence probablement aussi cette description. À ce stade de la FA, l'opérateur a pour but de faire une description de l'anomalie. Il n'a pas encore effectué de diagnostic et, de ce fait, n'a soit pas identifié la source effective du phénomène, soit ne le mentionne que lors de la description de la cause de l'anomalie, c'est-à-dire dans un champ distinct. Nous pouvons donc avoir ici, non pas une ambiguïté, mais plutôt un choix conscient de rester à un niveau d'indécision. Enfin, nous remarquons que certaines informations, ayant pourtant trait à la description de la fuite ne sont pas couverts par le schéma d'annotation, comme toute la première partie du deuxième exemple issu des rapports.

La définition du *frame Fluidic Motion* se concentre principalement sur l'idée de *mouvement* du fluide. Elle hérite d'ailleurs du *frame* « *Motion* ». Dans le cas des rapports d'anomalies, la focalisation est plutôt sur une description factuelle instantanée et statique de la situation. De ce fait, le schéma d'annotation proposé par *FrameNet* ne semble pas couvrir l'entièreté des éléments que nous pouvons retrouver dans les rapports.

Ces différences et ambiguïtés nous ont amenés à reconsidérer le schéma d'analyse de *FrameNet*, et son adéquation à la fois au français, mais surtout à l'environnement technique dans lequel se placent les rapports d'anomalies. En effet, nous avons pu voir avec notamment les travaux de L'Homme *et al.* (2020) que des modifications sur un *frame* peuvent être nécessaires afin de refléter les caractéristiques de la situation en rapport avec le domaine dont il est question. Ainsi, dans la section suivante, nous proposons l'étude des énoncés ayant trait à la fuite dans un environnement technique afin d'obtenir un schéma d'annotation adapté par notre situation. Ce schéma vise à refléter le contenu effectif des éléments qui expriment la fuite, mais dans un environnement spécifique qui est celui d'anomalies rapportées lors d'opérations de maintenance. Nous abordons plus largement certains aspects méthodologiques autour de l'étude fine de textes techniques pour lequel le linguiste n'a qu'une compréhension partielle de l'énoncé.

8.2 Trois approches

Pour mener cette étude, nous avons procédé à trois approches de linguistique outillée en parallèle pour aborder le contenu des données sous différents angles. Ces approches visent à

repérer des *patterns* récurrents dans l'expression de la fuite à travers des structures similaires. Pour ce faire, nous avons procédé à une opération d'effacement de texte, que nous définissons dans la section dédiée, afin de repérer les éléments récurrents autour des marqueurs de la fuite. Une seconde approche a consisté en l'extraction de structures syntaxiques récurrentes au sein du corpus. Enfin, nous avons pu effectuer un entretien avec un expert pour éclaircir certains éléments du texte.

8.2.1 Effacement de texte

Afin de repérer des *patterns* au sein du texte, nous avons mis en place une approche d'analyse du texte par effacement. Nous parlons d'effacement de texte en référence au procédé de *text bleaching*, que nous pourrions traduire par « javellisation du texte » introduit par les travaux de Van der Goot *et al.* (2018). Ce procédé est défini comme « transformant des chaînes lexicales en caractéristiques plus abstraites [...] qui présente l'avantage d'estomper les éléments afin de faire ressortir de patrons plus superficiels toujours basés sur le texte original, sans introduire de traitement additionnel tel que l'étiquetage POS ». Nous nous sommes inspirés de ce procédé afin de mettre en place deux étapes d'effacement des mots pleins et repérer les structures de surface récurrentes autour de la fuite.

8.2.1.1 Effacement manuel

Avant d'envisager un effacement automatique du texte, nous avons mis en place un premier essai d'effacement manuel dans l'objectif de déterminer les éléments minimaux autour de la fuite. Nous avons utilisé un ensemble d'exemples afin de déterminer quels sont les éléments et les structures récurrentes à partir d'une « simplification » des énoncés.

Dans la Table 8.1, nous pouvons voir les cas extraits manuellement des données. Nous avons sélectionné des descriptions d'anomalies complètes où apparaissent les marqueurs de la fuite, en excluant les cas avec plusieurs phrases ou des marqueurs d'autres catégories. Dans la colonne 1 se trouve un ensemble d'exemples qui correspondent à ces modes d'expression. Dans la colonne 2, nous présentons différents schémas issus de ces exemples. Nous pouvons constater une certaine phraséologie dans l'expression de la fuite avec des structures particulières. En revanche, ces structures sont multiples, nous n'observons *a priori* pas de régularités qui laissent transparaître une liste figée d'éléments qui devraient être présents. Nous avons donc effectué une première analyse des éléments récurrents afin d'en faire ressortir des patrons. Nous ne cherchons pas à identifier des schémas complets du rapport, mais plutôt à repérer les éléments minimaux autour de l'expression de la fuite en elle-même. Ce repérage se base sur notre propre analyse basée sur l'expertise de ces textes

que nous avons pu acquérir au cours de cette thèse. Un entretien avec un expert, que nous détaillons en section 8.2.3 nous permettra de confirmer la nature des éléments repérés.

Ainsi, après observation des exemples, nous avons choisi de sélectionner deux éléments récurrents dans les schémas :

— le fluide qui fuit (F) ;

— la source ou la destination, que ce soit un composant ou une localisation (L).

Dans le premier exemple de la table, nous pouvons donc voir que nous passons d'un énoncé plutôt long pour arriver à un schéma très réduit. En premier lieu, nous souhaitons observer les éléments qui caractérisent la situation de fuite. Nous ignorons de ce fait toute la séquence « CONDUISANT...24 HEURES » puisque celle-ci fait référence à la conséquence de la fuite. Dans un second temps, nous pouvons convenir que la séquence introduite par « SUR » fait référence à la localisation de la fuite. Cependant, celle-ci possède plusieurs constituants que nous ne pouvons, sans savoir expert, identifier et donc différencier. Parmi les deux éléments que nous considérons dans cette analyse, soit le fluide ou la localisation, seul celui de la localisation apparait. Nous avons procédé au même type d'opération pour les exemples suivants, nous amenant à nous focaliser sur les prépositions pour segmenter les différentes séquences.

Ainsi, dans les deux derniers exemples, nous pouvons voir deux cas avec le marqueur « fuyarde ». Le premier implique une localisation qui « est fuyarde ». Dans le second, nous pouvons cependant distinguer deux localisations séparées par la préposition « de ». Cette distinction est cette fois permise car les deux éléments « vanne » et « piscine » sont compréhensibles même pour un non-expert.

Nous pouvons émettre quelques observations d'après ces effacements manuels. Tout d'abord, nous observons la présence de plusieurs prépositions (ou locutions prépositionnelles) telles que *sur, au niveau de, à, de*, afin d'indiquer l'emplacement de la fuite, que ce soit le composant ou la localisation géographique. La connaissance limitée des composants et des localisations en question ne nous permet cependant pas de déduire si ces prépositions sont interchangeables ou si un usage particulier en est fait en fonction du complément qui le suit. Plusieurs de ces prépositions peuvent aussi être employées au sein d'un même rapport, afin de spécifier l'emplacement de la fuite à différents niveaux. Cependant, l'exemple 6 montre une utilisation différente de la préposition « à » dans « A SON OUVERTURE ». Nous aurons l'occasion de voir dans la section suivante que ces mêmes prépositions peuvent effectivement introduire des éléments autres que la localisation.

Nous avons aussi pu repérer les patrons les plus minimalistes de l'expression de la fuite qui sont : Fuite L / L fuyard. La manière d'exprimer une fuite le plus brièvement possible dans un rapport est donc de mentionner *fuite* ou *fuyard* ainsi qu'un élément de localisation.

Index	Exemples	Schémas
1	FUITE SUR LE PROCESS DE MAINTIEN EN FERMETURE VIRH GAM CONDUISANT A REGONFLER LA BOUTEILLE TOUTES LES 24 HEURES	Fuite sur L
2	FUITE D'AIR SUR RACCORD ALIMENTATION SERVO MOTEUR	Fuite de F sur L
3	FUITE AU NIVEAU DE LA BAGUE DE HVM0085 SUR LA PARTIE TUYAUTERIE	Fuite au niveau de L sur L
4	FUITE D'HUILE AU NIVEAU DE LA POMPE A EAU	Fuite de F au niveau de L
5	FUITES AU NIVEAU DE LA CEINTURE SUPERIEURE RIVETEE DE LA CASE	Fuites au niveau de L
6	FUITE AU PE DE OVM0016 A SON OUVERTURE	Fuite au L
7	HDT538 FUYARD DETARE.	L fuyard
8	RACCORD GROMELLE HFX001 FUYARD NIVEAU SERTISSAGE	L fuyard niveau L
9	VANNE, ANALYSE INTERFACE 370, LEGEREMENT PLIEE OCCASIONNANT UNE FUITE.	État de L occasionne une fuite
10	VANNE DE PURGE ENTRE HVP036 ET HVP037 FUITE EN LIGNE.	L fuite
11	FUITE EN LIGNE DE HVM9125	Fuite de L
12	FUITE LIGNE GAM SOL EVEC	Fuite L
13	FUITE DU JOINT GONFLABLE DU VOLET DE PONT 300KN AU BIL COTE YN.	Fuite de L au L
14	FUITE DANS LE CIRCUIT AIR DU SYSTEME DE LEVAGE DU VOLET DE PONT.	Fuite dans L du L
15	APRES RACCORDEMENT LIGNE AZ SUR PAC LH2, LA LIAISON PONTET/PAC EST FUYARDE	L est fuyarde
16	LA VANNE GUILLOTINE DE LA PISCINE LH2 EPC EST FUYARDE	L de L est fuyarde

TABLE 8.1 – Patrons minimaux de la fuite après analyse manuelle

Ce procédé semble en mesure de faire émerger différents patrons minimaux d'expression de la fuite. Par ailleurs, nous avons pu voir que la construction implique fortement des prépositions afin de spécifier la localisation de la fuite, localisation qui est l'élément le plus prépondérant de ces patrons. Dans un second temps, dans la section suivante, nous effectuons un procédé similaire, mais avec une automatisation de la formation des patrons pour confirmer et faire émerger des tendances quant à ces modes d'expressions.

8.2.1.2 Effacement automatique

La méthode que nous utilisons pour l'effacement automatique du texte est similaire à celle effectuée par Tanguy et Rebeyrolle (2020) sur des titres de publications scientifiques,

elle-même inspirée de méthode d'extraction de motifs de (Quiniou *et al.*, 2012). Elle est fondée sur un repérage de séquences partielles reposant sur l'ordre des éléments dans le texte, plutôt que sur des répétitions à l'identique. Ce que nous cherchons ici à mettre en place est de pouvoir délimiter les segments. Nous effaçons ensuite des éléments de surface afin de repérer les séquences récurrentes.

Pour le traitement du corpus, nous avons utilisé tous les rapports considérés comme relevant de la catégorie fuite, c'est-à-dire comportant un marqueur appartenant à la catégorie *Fuite d'un liquide ou d'un gaz*, lors de l'application du modèle, soit 1827 descriptions d'anomalie. Nous rappelons que les résultats de l'étiquetage automatique a donné des scores de 1 pour cette classe, en rappel et en précision lors de l'évaluation sur le test set. Nous travaillons donc, *a priori* sur tous les cas disponibles dans le corpus. Les rapports ont été traités avec *Stanza*, c'est-à-dire étiquetés en parties du discours, lemmatisés et analysés syntaxiquement en dépendances. Nous avons pu voir que les performances de cet outil sur nos données sont limitées. Cependant, notre objectif dans cette première approche n'est pas d'obtenir une analyse parfaite. Nous souhaitons uniquement récupérer le lemme et la catégorie morpho-syntaxique en nous concentrant sur les éléments les plus fréquents au sein des rapports, pour ensuite analyser les *patterns* retrouvés. Nous procédons donc par étapes successives jusqu'à arriver à un schéma très minimal. La méthode vise cependant à récupérer les schémas intermédiaires récurrents. Nous avons donc extrait les segments autour des lemmes « fuite » ou « fuyard », c'est-à-dire les marqueurs étiquetés par le modèle d'étiquetage automatique, en découpant sur les éléments de ponctuation avant et après. Au sein de ces séquences, les mots d'une fréquence inférieure à 10 occurrences dans l'ensemble des rapports avec une fuite sont remplacés par « V » et « A » si ce sont des verbes ou des adjectifs et sinon par « X ». Pour les mots dont la fréquence est supérieure à 10, nous gardons les lemmes que nous remplaçons au fur et à mesure des opérations d'effacement. Les prépositions sont systématiquement conservées au fur et à mesure de l'opération mais les déterminants sont effacés.

Le procédé employé vise à un effacement itératif des énoncés. Dans l'exemple Table 8.2, nous pouvons observer un rapport ayant subi le processus d'effacement. Une fois la séquence segmentée (sur le point en l'occurrence), les règles de remplacement et d'effacement entrent en jeu. Les effacements se font de façon itérative et visent à regrouper les syntagmes nominaux. Les règles instaurées entraînent la fusion des groupes nominaux en un seul élément X et les éléments peu informatifs comme les déterminants et les nombres sont éliminés. Les règles sont pour la majeure partie reprises de Tanguy et Rebeyrolle (2020).

Les regroupements se font d'abord sur les séquences incluant un nom entouré d'adjectifs. Dans un second temps, les noms isolés sont remplacés par X. Les séquences avec plusieurs

Rapport original	FUITE D'HUILE AVEC COULURE SUR LE BRAS DE CROIX AXE Z. FLEXIBLE HP GAM/SVE AU NIVEAU DU PASSE CLOISON (VOIR PHOTOS).
Séquence lemmatisée extraite	fuite de huile avec coulure sur le bras de croix axe z.
1	fuite de huile avec X sur le bras de X axe X
2	fuite de huile avec X sur le bras de X
3	fuite de huile avec X sur X axe X
4	fuite de huile avec X sur X
5	fuite de X avec X sur le bras de X axe X
6	fuite de X avec X sur le bras de X
7	fuite de X avec X sur X axe X
8	fuite de X avec X sur X

TABLE 8.2 – Effacement progressif du texte

X consécutifs sont regroupés en un seul. Les structures de type X suivit de *de, le* ou *un* et d'un nouveau X sont ensuite agrégées en un seul X. Les déterminants simples devant un X sont supprimés. Et enfin, les structures avec un chiffre précédé ou suivi d'un X sont aussi agrégées en un seul X.

Nous pouvons voir par exemple que la séquence « axe X » en numéro 1 est aplatie en numéro 2 puisque « X axe X » est agrégée sous un seul « X ». Cependant, en *pattern 3*, il fait sa réapparition et cette fois, c'est « le bras de X » qui est agrégé en un seul « X ». Cette réapparition se fait car, lorsqu'on est en présence d'un mot d'une fréquence supérieure à 10 occurrences, nous créons deux versions du pattern. Une version ou le mot est effacé, et une seconde où il est conservé. À partir de la ligne 5, c'est « huile » qui est remplacé par X et le reste de la séquence redevient entière, sans aplatissement mais avec les remplacements initiaux des termes inférieurs à 10. Ils sont cependant peu à peu effacés jusqu'au *pattern* le plus petit pour cette phrase soit *fuite de X avec X sur X*.

Ce procédé nous a permis d'identifier un certain nombre de structures récurrentes autour des lemmes *fuite* et *fuyard*, des prépositions qui les entourent et des termes récurrents. Dans la Table 8.3, nous retrouvons la liste des schémas extraits par le biais de l'effacement automatique classés par ordre de fréquence. Nous avons inclus les 30 schémas les plus fréquents. Tous les *patterns* issus du procédé d'effacement sont conservés, et non uniquement les *patterns* finaux (les plus petits). Nous pouvons observer que le schéma « fuite sur » est le plus fréquent parmi les rapports. Nous retrouvons aussi de nombreuses fois « fuite de » et « fuite au niveau de » (dans le tableau, ces derniers sont aussi couverts par les séquences

Fréquence	**Schéma extrait**
272	fuite sur X
85	fuite X
64	fuite à X
58	X fuyard
45	fuite A sur X
43	fuite de X
39	X fuyard en X
39	fuite de X sur X
37	fuite à le niveau de X
32	X fuyard en ligne
25	fuite
24	fuite en X
21	X être fuyard
21	fuite A de X
19	fuite en X sur X
18	fuite sur X à X
18	fuite externe sur X
18	fuite de X à X
17	fuite en ligne sur X
17	fuite A à X
16	fuite A
13	fuite interne de X
12	fuite sur bride X
12	fuite en ligne de X
12	fuite de huile sur X
11	X fuite sur X
11	X fuite à X
11	fuite sur X et X
11	fuite sur raccord X

TABLE 8.3 – Schémas minimaux de la fuite extraits avec effacement de texte

« fuite à ») . Nous pouvons aussi noter la présence du mot « huile » à plusieurs reprises. Ce fluide semble être celui qui fait le plus souvent l'objet d'une fuite. Nous notons aussi la présence de l'expression « en ligne » qui n'apparaît dans le corpus que pour des cas de fuite.

Un entretien avec l'expert, sur lequel nous revenons par la suite, a permis de l'expliciter, indiquant qu'il s'agit alors d'une fuite interne. Les *patterns* montrent aussi la combinaison de plusieurs prépositions « fuite de X à X », « fuite de X sur X »... Nous retrouvons ici les combinaisons de prépositions que nous avons déjà repérées lors de l'effacement manuel.

Les schémas que nous avons ainsi retrouvés nous permettent d'observer les différentes structures dans l'expression de la fuite. Nous souhaitons approfondir ces schémas du point de vue de l'utilisation des prépositions qui sont apparues comme nombreuses et pour dont pour l'instant l'utilisation paraît indifférenciée. Par ailleurs, les positions des éléments ne sont pas nécessairement significatives et nous souhaitons donc observer les relations autour des prépositions en faisant abstraction de la position des éléments.

8.2.2 Extraction de patrons syntaxiques

Une des approches que nous avons menées a visé l'extraction de patrons syntaxiques à partir du texte. Cette extraction est à nuancer parce que, comme nous avons pu le voir, les outils tels que *Stanza* (que nous avons utilisé à nouveau dans le cas présent), ne sont pas adaptés au type de textes avec lequel nous travaillons. Nous n'avons donc pas fait cette extraction dans l'objectif de récupérer des éléments syntaxiquement systématiquement corrects, mais plutôt afin de repérer, une fois encore, les éléments qui s'articulent autour de la fuite. Cette fois, plutôt que de se concentrer sur des patrons, nous souhaitons au contraire observer quels sont les éléments concrets que nous retrouvons au sein de ces *patterns*.

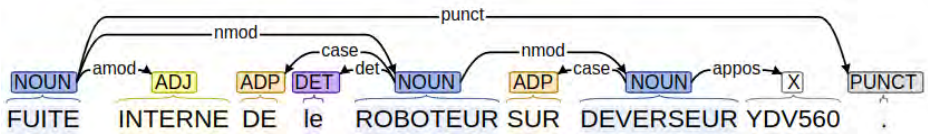


FIGURE 8.4 – Exemple de dépendance syntaxique à partir de la tête

Dans cette approche, nous souhaitons récupérer les éléments de nos énoncés qui s'articulent autour de « fuite » en tête syntaxique. Dans la Figure 8.4, nous pouvons retrouver un exemple de rapport annoté en dépendance syntaxique. Nous pouvons repérer deux structures avec la fuite en tête avec une préposition (ADP) et un nom (NOUN) qui y est rattaché : « fuite de le roboteur » et « fuite sur deverseur ». À cela s'ajoute l'adjectif « interne ».

Dans la Table 8.4, nous pouvons observer les extractions que nous avons faites des différentes catégories grammaticales reliées à « fuite » en tête syntaxique. Nous y retrouvons la liste des 10 prépositions les plus fréquentes et des 20 noms, adjectifs et syntagmes préposi-

tionnels[71] les plus fréquents. Parmi les prépositions, nous retrouvons « sur » comme plus fréquente avec 775 occurrences suivie de « de » et « à » avec respectivement 418 et 296 cas. Dans la liste des noms, nous retrouvons « niveau » en première position qui appartient à la locution prépositionnelle « au niveau de ». Elle est suivie de « ligne » avec 151 occurrences que nous avons pu observer dans l'expression « fuite en ligne » pour indiquer une fuite interne. Dans la liste des syntagmes prépositionnels, nous retrouvons d'ailleurs « en ligne » avec 70 occurrences. Les premiers fluides arrivent en quatrième et cinquième position avec « eau » et « huile » et respectivement 86 et 79 occurrences. Nous retrouvons aussi plusieurs éléments qui peuvent être rattachés à des localisations (« vanne », « bride », « servomoteur »...). Pour les adjectifs, les plus fréquents servent à indiquer une notion de quantité « importante » (71) et « legere » (14), ou si la fuite est « interne » (61) ou « externe » (46).

Prep	**Fréq**	**Noms**	**Fréq**	**Adj**	**Freq**	**Prep et nom**	**Freq**
sur	775	niveau	199	importante	71	à niveau	189
de	418	ligne	151	interne	61	d' huile	76
à	296	raccord	103	externe	46	d' eau	74
en	150	eau	86	legere	14	sur raccord	70
a	49	huile	78	audible	14	en ligne	70
entre	39	vanne	66	hydraulique	13	sur bride	38
dans	38	fuite	64	petite	7	sur ligne	33
pour	22	bride	45	constatee	6	d' air	27
par	16	servomoteur	33	grosse	5	à snoop	21
hors	10	cote	31	externes	5	sur flexible	19
		air	30	permanente	5	sur vanne	16
		joint	26	detectee	4	sur tuyauterie	14
		ncm3	26	superieure	4	sur verin	13
		flexible	25	importantes	4	sur circuit	13
		membrane	25	persistante	3	sur servomoteur	13
		bulles	24	visible	3	en position	12
		spec	23	localisees	3	sur joint	12
		snoop	22	probable	3	sur garniture	12

TABLE 8.4 – Éléments les plus fréquents autours de la fuite selon l'analyse syntaxique

71. Nous pouvons repérer quelques erreurs dans cette liste, comme « constatee » ou « detectee » dans la liste des adjectifs, en raison des difficultés à analyser un texte spécialisé et bruité éprouvées par Stanza. Pour rappel, nous avions estimé sont taux d'erreur pour l'étiquetage POS à 20% sur nos données.

Concernant les syntagmes prépositionnels les plus fréquents extraient, nous avons par la suite pu aller voir en corpus comment ils se manifestent. Dans la Table 8.5, nous pouvons voir une sélection de rapports pour l'occurrence « fuite sur raccord ». Nous pouvons observer ainsi quels sont les autres éléments présents autour de ces patrons. Dans les exemples 1 et 3, nous identifions donc un moyen de détection de la fuite « detectee au snoop » et des localisations supplémentaires (étant donné que le « raccord » relie deux éléments du système, des précisions sur ces derniers sont présents). Le moyen de détection « snoop » fait également son apparition dans l'exemple 3 de la table. Dans l'exemple 2, nous pouvons aussi observer l'apparition d'une mesure avec « 10^{-2} » et dans l'exemple 4, nous retrouvons la mention du fluide qui fuit (« huile »). Ces deux cas, « au snoop » et « à 10^{-2} » nous montrent aussi que les prépositions dépassent le simple cadre des localisations pour introduire d'autres éléments du problème. Dans les exemples « FUITE **AU** RACCORD MANOMETRE HPI 047 (STOCKAGE 600 b). ISOLEMENT DU MANOMETRE PAR HVM 0226. » et « HSS502 : EVENT TPCH - FUITE **AU NIVEAU DU** RACCORD ENTRE LA TAPE ET LA SOUPAPE (RACCORD AU PLUS PRES DE LA SOUPAPE). », nous constatons l'utilisation cette fois des prépositions « au » et « au niveau de » de façon interchangeable pour faire référence à une fuite au niveau d'un raccord. Cela nous montre aussi l'absence de différenciation par les prépositions pour introduire une localisation.

	fuite sur raccord
1	**FUITE** (DETECTEE AU SNOOP) **SUR** LE **RACCORD** ENTRE LE FLEXIBLE EPSRH ET LA PCP.
2	**FUITE** A 10^{-2} **SUR RACCORD** OMBILICAL DE PURGE H2 ESC.
3	**FUITE SUR RACCORD** DES LIAISONS ECCN1 (PCPO) ET VIRH (PCPO) DETECTEES AU SNOOP.
4	**FUITE** D'HUILE **SUR** UN **RACCORD** DU VERIN DE LEVAGE C DURANT L'OPERATION DE LEVAGE.

TABLE 8.5 – Exemples de « Fuite sur raccord «

Concernant les prépositions, nous avons pu voir avec *fuite sur* que celles-ci peuvent introduire une variété d'éléments différents. À travers les exemples 1, 2 et 4 de la Table 8.5, nous avons aussi pu voir la présence d'autres prépositions directement dépendantes du marqueur mais qui introduisent des éléments qui diffèrent de la localisation. Nous avons ainsi pu constater la présence d'un moyen de détection (le *Snoop*), d'une mention d'un débit (10^{-2}) et nous retrouvons le fluide (*huile*). Nous avons aussi vu que les différentes prépositions peuvent aussi indiquer des même localisations. Ce procédé nous a donc permis d'identifier certains éléments récurrents autour des *patterns* préalablement identifiés avec l'effacement de texte. Nous avons par exemple vu émerger le moyen de détection et le débit du fluide comme éléments récurrents, en plus du fluide et des différents niveaux de localisation déjà

repérés, mais qui viennent se confirmer par cette étape. Il nous a aussi permis de constater une interchangeabilité apparente dans l'utilisation des prépositions.

8.2.3 Entretien avec l'expert

Enfin, la dernière étape de cette exploration de corpus a été l'entretien avec un expert. En effet, afin d'éclairer les informations obtenues lors des deux premières étapes, il a été nécessaire de questionner un expert sur le sens d'un certain nombre de mots apparaissant dans les rapports. Pour cet entretien, nous avons donc préparé une série d'éléments à lui présenter. Tous d'abord, nous avons souhaité obtenir sa définition de certains éléments techniques fréquents que nous avons pu repérer au cours de nos analyses. Par ailleurs, nous avons sélectionné des exemples de sorties d'analyses syntaxiques afin de pouvoir présenter des structures récurrentes mais dont l'emploi, notamment des prépositions, pouvait différer.

Tout d'abord, une première étape a simplement consisté en la définition des termes récurrents du sous-corpus, que ce soit des termes techniques ou du vocabulaire scientifique général comme :

— *table* : table de lancement ;
— *bulles* ou *mille bulles* : moyen de détection d'une fuite ;
— *LH2* : hydrogène liquide...

De même, cela a permis de définir certaines structures comme « en ligne », qui indique que la fuite est interne, ou bien « au Snoop » qui est la marque d'un produit qui permet de détecter une fuite ou encore « bave de crapaud » qui fait référence à de petites bulles que forme le Snoop et qui est indicateur d'une petite fuite (ce dernier prend à la fois le rôle de moyen de détection, mais aussi d'indicateur de quantité dans l'énoncé).

D'autre part, cela a permis de passer en revue certains des résultats de l'extraction syntaxique vue en section 8.2.2, et notamment autour de la question des prépositions. L'objectif ici était de déterminer, par le biais d'exemples soumis à l'expert, si les différentes prépositions sont propres à certains types de compléments, s'il existe une raison à l'emploi de « sur », « au », « au niveau de »... L'hypothèse étant que certaines prépositions pouvaient impliquer un niveau de localisation qui diffère. Le résultat de cet entretien a cependant révélé que ce choix est indifférencié et non justifié. D'une part, il n'existe pas d'indication particulière à ce propos, d'autre part, ses explicitations et observations sur les exemples ont montré un usage interchangeable des différentes prépositions. C'est par la connaissance des éléments du système que se font les distinctions.

8.2.4 Synthèse de l'analyse

Ces trois approches nous ont permis de faire plusieurs observations quant aux régularités dans l'expression de la fuite dans un environnement technique. Tout d'abord, nous avons pu voir une grande variété dans les schémas utilisés. Il n'existe pas un seul *pattern* utilisé de manière systématique. Cependant, nous observons une régularité certaine dans la nature des éléments récurrents et dans les schémas : *fuite de NOM*, *fuite ADJ*, *fuite à NOM*, *fuite au niveau de NOM*... Ces schémas sont aussi combinables. De par notre expérience du corpus, nous avons également pu identifier certains éléments. En revanche, d'autres restent incompréhensibles et l'interchangeabilité des prépositions empêche de les identifier. La présence d'un expert reste donc indispensable.

Nous proposons ici une liste des différents éléments récurrents que nous avons repérés à partir de la façon dont ils s'expriment.

— Les marqueurs de la fuite sont au nombre de trois : *fuite(s)*, *fuyard(es)* et *écoulement*. Ces marqueurs correspondent aux unités lexicales qui font office de déclencheurs de la situation. Les cas d'écoulement sont rares, avec seulement 35 occurrences sur les 1997 relevés par l'étiquetage automatique (soit 1,8%), c'est pourquoi nous nous focalisons essentiellement sur les deux autres. Les schémas extraits lors de l'effacement automatique ont montré que *fuite(s)* est majoritairement suivi d'une préposition (« à, de, au niveau de, sur, en »), avec éventuellement un adjectif avant cette préposition. Pour *fuyard(es)*, ce marqueur est exprimé de deux façons : soit « élément *fuyard* », soit « un élément *est fuyard* ». Nous retrouvons aussi quelque exemples de l'emploi du verbe conjugué sous la forme *fui(en)t* avec 21 occurrences.

  (1) LA CANALISATION D'EVACUATION DES EAUX DE PLUIE **FUIT** AU NIVEAU DU COUDE SOUS L'ETABLIE COTE FENETRE EN SALLE PROPRE SUR ELA2.

— Nous retrouvons aussi le *fluide*, présent dans le *frame Fluidic motion*. Ce dernier n'est pas toujours exprimé. Il suit généralement le marqueur directement et est souvent introduit par une préposition mais parfois sans. Nous avons vu des exemples avec *FUITE D'HUILE* mais pouvons aussi noter :

  (2) FUITE D'**AZOTE** (LEGERE) SUR AVM9106 (TABLE 2)

  Dans certains cas, il est possible qu'un fluide soit mentionné comme spécifiant la localisation ou le débit, mais il n'est pas systématiquement synonyme du fluide qui fuit. Dans l'exemple 3, nous pouvons ainsi remarquer que la localisation mentionne le fluide « eau » mais le fluide qui fuit est du *freon*. Dans l'exemple 4, il est mention de *13,5 BAR AZOTE*, mais pas d'une fuite d'azote. Dans ce type de cas, c'est l'expert qui peut trancher si par exemple, il a connaissance du fait que seul de l'azote peut

fuire au niveau du « AFX0605 » ou s'il sait qu'en le mentionnant sans autre fluide dans le rapport, cela fait office de désambiguïsation.

(3) FUITE **FREON** SUR GROUPE DE PRODUCTION EAU GLACEE N°4. (CONDENSEUR)

(4) FUITE SUR AFX0605 AU NIVEAU RACCORDEMENT (13.5 BAR **AZOTE** VENTILATION CASE)

— Parmi les éléments repérés se trouvent aussi des informations relatives à l'expression de la *quantité* de fluide qui fuit. Pour autant, la mention du fluide et de la quantité ne sont pas systématiquement exprimés ensemble. Cette dernière peut être seule. Nous observons deux façons d'exprimer cette quantité, soit *via* un qualificatif, soit une mention précise d'unité de mesure. Nous avons pu voir le « 13.5 BAR » dans l'exemple 4, ou *LEGERE* dans l'exemple 5. En de rares occasions, nous retrouvons différentes sortes de mentions, comme pour l'exemple 6 où c'est un taux de diminution qui est indiqué.

(5) **LEGERE** FUITE DE VENTILATION PAR LES BORNES DE MASSE RMM EN Z ET ZN.

(6) FUITE VCO **- 0.3 %** O2 LU SUR ANALYS

— Parmi les éléments qui ont émergé, nous pouvons aussi parler des moyens de détection employés pour repérer ou quantifier la fuite. Celui-ci n'est pas présent dans le *frame Fluidic motion*. Nous pouvons en citer quelques-uns comme le *Snoop*, qui semble être une antonomase du nom de la marque du produit, le *mille bulles*, la méthode *drager*... Nous trouvons aussi le cas de l'expression imagée *bave de crapaud* qui représente les bulles formées par le *Snoop*. Celle-ci est à la fois un moyen de détection et un indicateur de quantité (elle indique une petite fuite). Elle peut être exprimée conjointement avec le *Snoop* (exemple 7) ou seule (exemple 8). Ces moyens de détections peuvent être introduits par le participe passé « détecté » rendant ainsi clair leur nature (comme dans l'exemple 3 de la Table 8.5). Nous les trouvons aussi souvent directement rattachés à « fuite » ou entre parenthèses, comme dans l'exemple 7. L'exemple 8 montre un cas intéressant où il est précisé une « OBSERVATION VISUEL(LE) » d'une fuite. Elle pose la question d'une potentielle annotation de cette observation comme un moyen de détection alors que celle-ci est implicite dans de nombreux rapports.

(7) FUITE AU **SNOOP (BAVE DE CRAPAUD)** SUR EAPH ET SUR BMH1.

(8) PROCESS ESC :LORS DE L'ETANCHEITE A 4 bars SUR XFX0602 ET XFX1601 IL Y A UNE FUITE STYLE "**BAVE DE CRAPEAU**" SUR LES DEUX CAPOTS JONHSTON.- XFX0602 : LIGNE 31- XFX1601 : LIGNE 33

(9) OBSERVATION VISUEL DE FUITE D'ELECTROLYTE AUX ALENTOURS DU BAC DE BATTERIES.

— Nous retrouvons aussi certains éléments pour indiquer si la fuite est *interne ou externe*. Pour les fuites internes, l'expression « EN LIGNE » est fréquemment retrouvée avec 138 occurrences, 90 pour *INTERNE* et 75 pour *EXTERNE* (par défaut, selon l'expert, une fuite est externe). Nous pouvons voir un cas d'usage dans l'exemple 10, mais retrouvons aussi les expressions *fuire en ligne, fuite en ligne* ou *fuyarde en ligne* dans le corpus. Nous avons aussi pu repérer quelques cas de fuite « INTERNE/EXTERNE ».

(10) LES 2 VANNES DE VIDANGE FUIENT **EN LIGNE**.

(11) FUITE **INTERNE/EXTERNE** SUR HVM0377 (DISTRIBUTION 243 CANA K).

— Parmi les éléments, nous avons pu noter la présence des localisations, parfois seules présentes en compagnie du marqueur comme dans l'exemple 12. Il paraît être l'élément fondamental de la description d'une fuite dans une FA. Les éléments de localisation peuvent être nombreux (comme le laisse entendre l'accumulation des prépositions successives que nous avons vu dans les patrons), variés (ce que nous avons pu voir dans la Table 8.4) et montrent différents niveaux de précision. Dans l'exemple 13, nous retrouvons donc le « BAF », qui est un bâtiment, et deux précisions, soit la « GARNITURE » qui est un élément du « COMPRESSEUR N°6 ». Cela montre l'importance de la localisation de la fuite pour sa résolution. Nous nous éloignons ici des localisations telles qu'elles sont exprimées dans le *frame* de *FrameNet* qui se focalise sur le mouvement du fluide et donc sa source et son arrivée. Ici, il existe un flou langagier qui ne permet pas de distinguer ces deux éléments mais traduit aussi un certain rapport à la fuite. La focalisation n'est pas sur la description de son mouvement, mais plutôt sur la transmission de sa localisation, le plus précisément possible, afin de résoudre le problème, qui est l'objectif premier d'un REX.

(12) **RACCORD NVM057 / NPP005** FUYARD.

(13) **CEG BAF** : FUITE SUR **GARNITURE** DU **COMPRESSEUR N°6**.

— Enfin, le dernier élément que nous pouvons citer est un ensemble de circonstants qui viennent préciser la situation autour de la fuite. Dans les exemples 14 et 15, nous retrouvons deux cas avec des informations temporelles pour préciser les situations dans lesquelles apparaissent les fuites « QUAND LE SKID EST EN PRESSION » et « LORS DE L'ESSAI DE LA POMPE N°3 ». Dans l'exemple 16, nous trouvons cette fois une précision sur les conséquences potentielles la fuite en raison de son environnement « NOTA : PRESENCE DE NOMBREUX EQUIPEMENTS ELECTRIQUES DANS LE FAUX-PLANCHER ».

(14) LEGERE FUITE D'HUILE AU NIVEAU DE L'ECROU INFERIEUR DE L'INDICATEUR DE NIVEAU DU SKID GAM/GAT **QUAND LE SKID EST EN PRESSION** (> 2bar)

(15) **LORS DE L'ESSAI DE LA POMPE N°3**, IL Y A UNE FUITE AU P.E DE CETTE POMPE.POMPES DE VIDANGE CARNEAU EN POST-LANCEMENT.

(16) PRESENCE D'EAU SUR LE SOL DU LOCAL CLIM 4.1 TABLE 1 (FUITE D'EAU PROVENANT DU PLAFOND DU LOCAL). **NOTA : PRESENCE DE NOMBREUX EQUIPEMENTS ELECTRIQUES DANS LE FAUX-PLANCHER**.

Nous avons ici présenté les différents éléments qui participent à l'expression de la fuite [72]. Ces éléments sont variés, nombreux, mais présentent tout de même des structures récurrentes grâce auxquelles nous avons pu les identifier. Ils possèdent aussi la particularité d'être, sauf le marqueur et la localisation, facultatifs. Cela démontre le flou qui perdure dans l'expression de la fuite, malgré le besoin d'en exprimer sa localisation de façon aussi précise que possible, sachant que le phénomène est par nature élusif. Nous percevons aussi la dissociation avec le frame de *FrameNet Fluidic motion* qui se concentre sur la description d'une situation qui diffère de la nôtre, notamment en raison de ses enjeux.

8.3 Un *frame* de la fuite dans un environnement technique

Maintenant que nous avons présenté les différents moyens d'étudier les énoncés de la fuite et les éléments repérés, nous présentons le *frame de la fuite dans un environnement technique* qui en découle. Dans un premier temps, nous présentons les différents éléments du *frame*, puis l'annotation effectuée par un expert qui nous a permis de mettre à l'épreuve la grille d'annotation.

8.3.1 Présentation du schéma d'annotation

Les différentes approches vues précédemment nous ont permis d'établir un schéma d'annotation pour la fuite dans un environnement technique.

72. Nous pouvons tout de même noter quelques cas ou le marqueur *fuite* a été repéré mais n'est pas la focalisation du rapport tout en ayant une importance certaine dans le problème comme « EN FIN D'OPERATION APRES AVOIR RINCE LE SOL POUR EVACUER LES TRACES DE SOUDE DEVANT LA COLONNE DE LAVAGE N, L'OPERATEUR A EU UNE PROJECTION D'EAU DU RIA QUI ETAIT **FUYARD**. IL S'EST FROTTE LE VISAGE AVEC LA MAIN SUR LAQUELLE IL Y AVAIT DES TRACES DE SOUDE CE QUI A ENTRAINE UNE PENETRATION DE SOUDE DANS L'OEIL ».

Dans la Table 8.6, nous présentons le schéma d'annotation en question. Celui-ci est composé de huit *frame elements* distincts (ils sont classés de façon aléatoire dans le tableau). Ceci est la grille finale qui est issue à la fois des analyses présentées ci-avant et des deux étapes d'annotations donc nous discutons ci-après.

1. Le premier élément qui compose ce schéma d'annotation est le ***Fluide***, soit le liquide ou le gaz qui fuit. Cet élément est déjà présent dans le schéma d'annotation de *FrameNet* et nous avons pu voir qu'il est récurrent au sein des rapports d'anomalies.
2. Le second élément est le ***Dispositif de détection***. Il correspond au dispositif qui a été utilisé pour détecter la fuite ou le type de fuite (en termes de quantité notamment) dont il est question. Nous avons pu voir quelques expressions renvoyant à cette catégorie telles que « au snoop », « bave de crapaud » ou encore « mille bulles ».
3. Le troisième élément est la ***Quantité***, pour référer à un indicateur quant à l'importance de la fuite comme « légère », « conséquente »... Cet élément se retrouve dans le schéma de *FrameNet* sous l'étiquette plus générale de *Configuration*.
4. En quatrième élément se trouve le ***Débit***. Il fait référence à l'utilisation d'unités de mesure précises pour quantifier la fuite. Nous avons choisi de différencier cet élément de *Quantité* car il démontre un niveau de description plus précis, une indication chiffrée et mesurée de l'événement.
5. ***Condition d'apparition*** est le cinquième élément de ce schéma. Ce dernier est assez rarement présent et n'a de ce fait pas initialement fait partie du schéma d'annotation. Cependant, l'annotation de la fuite par un expert, que nous décrivons par la suite, à fait émerger cette catégorie. Elle concerne les cas où une fuite apparaît lorsqu'une condition est remplie. Elle est donc relier aux éléments circonstanciels de la fuite, mais la focalisation est sur le fait que la présence de la fuite est dépendante de cette circonstance.

Enfin, les trois derniers éléments font tous référence à l'emplacement de la fuite. Nous avons choisi de faire fi de la différenciation *Source / Goal* effectuée dans *FrameNet* en raison des écueils que nous avons repérés lors des différentes étapes d'analyses du corpus. En effet, la différenciation entre ces deux éléments est moins aisée à cause du manque de clarté des prépositions utilisées en français. De plus, la visée de ces rapports est de faire une description statique, instantanée de la situation et n'implique pas forcément une différenciation ou un diagnostic permettant d'identifier ces deux éléments. L'accent est mis sur la description de **où la fuite a lieu**, et non du mouvement du fluide d'une source jusqu'à sa destination. Nous avons donc effectué un découpage des éléments de localisation en fonction de leur précision :

Index	Participants	Définitions
1	Fluide	Liquide ou gaz qui fuit
2	Dispositif de détection	Dispositif utilisé pour détecter la fuite
3	Quantité	Renvoie à la quantité de fluide (ex : fuite importante, légère...)
4	Débit	Débit hors standard (conséquence ou signal de la fuite)
5	Condition d'apparition	Condition nécessaire à l'apparition de la fuite
6	Localisation précise	Élément précis du processus fluide (ex : "joint", "vanne")
7	Localisation moyenne	Circuit ou système avec une composante hydraulique (ex :"partie eau froide du compresseur")
8	Localisation géographique	Localisation géographique large (ex : "Bâtiment d'assemblage")

TABLE 8.6 – Schéma d'annotation de la fuite dans un environnement technique

7. La ***Localisation précise*** renvoie à un composant de la chaîne du fluide, c'est-à-dire le ou les systèmes hydrauliques dans lesquels le fluide est censé circuler. Ce composant est généralement de petite taille et porteur de la fuite. C'est l'endroit où la fuite est percevable. Les termes que nous retrouvons le plus souvent sont « joint », « vanne », « raccord » ou des référents spécifiques comme « XVP003 », « OVM 5002 »...
8. L'élément ***Localisation moyenne*** renvoie à une partie du circuit hydraulique sur lequel se trouve la localisation précise. Par exemple, nous pouvons trouver « stockage LO2 », « ligne de remplissage LOX 31 », ou encore « réseau eau de ville ».
9. Enfin, le dernier élément ***Localisation géographique*** renvoi à une indication plus large sur l'emplacement de type « Bâtiment d'assemblage » ou bien « table[73] 2 ».

Pour illustrer ce schéma, nous reprenons dans la Figure 8.5 les exemples de la Figure 8.3, mais cette fois en comparaison avec le schéma d'annotation que nous proposons. Nous observons que le schéma d'annotation pour la fuite dans un environnement technique permet d'obtenir une plus grande couverture de l'énoncé et un niveau de détail supérieur, notamment en regard de la localisation de la fuite, dans les exemples 2 et 3, et des moyens de détection dans les exemples 1 et 2.

8.3.2 Annotation par un expert

Afin de mettre à l'épreuve la grille, nous avons fait réaliser une annotation par un expert. Pour ce faire, nous avons effectué deux processus d'annotation.

73. Pour « table de lancement »

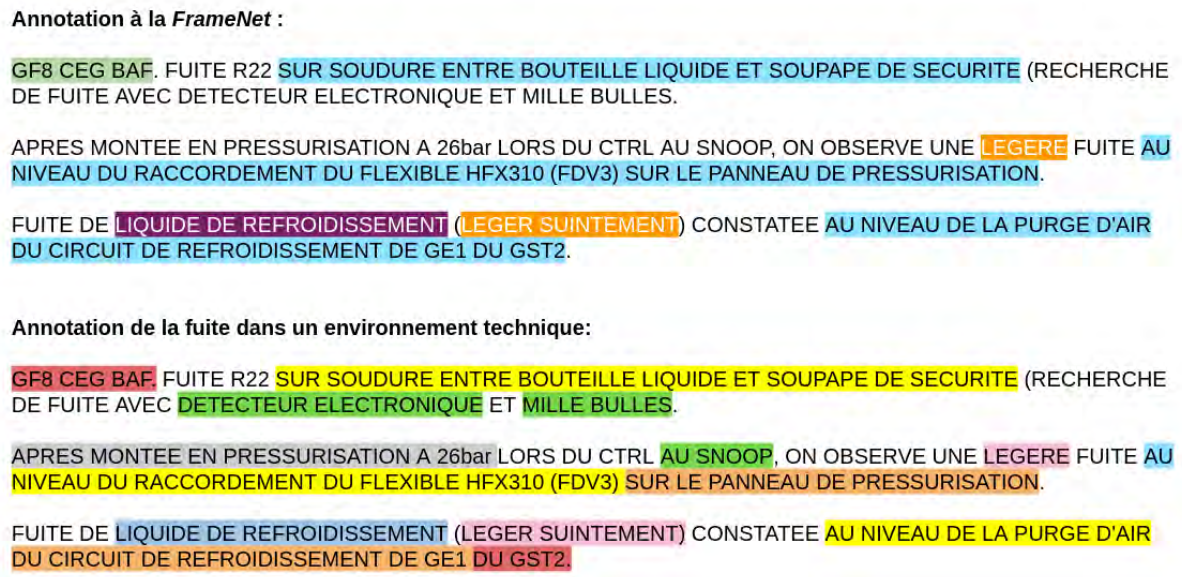


FIGURE 8.5 – Comparaison des annotations de la *fuite* - Les codes couleurs sont issus du schéma d'annotation de *Fluidic Motion* pour l'annotation à la *FrameNet* et de la grille de la Table 8.6 pour notre version

Les phases d'annotation ont été effectuées sur deux échantillons différents de 50 rapports chacun. Ces rapports ont été sélectionnés de façon aléatoire parmi ceux catégorisés dans la classe *Fuite d'un liquide ou d'un gaz* par le modèle d'étiquetage automatique. De même que lors de l'annotation des marqueurs, l'annotateur a sélectionné librement les segments et les a catégorisés (toujours avec *Glozz* mais avec une grille d'annotation différente). Nous pouvons aussi noter que l'annotateur avait la possibilité d'utiliser une même étiquette à plusieurs reprises au sein d'un rapport.

Lors de la première phase d'annotation, nous avons choisi de ne pas prendre en compte les deux éléments qui sont : la condition d'apparition de la fuite et la mention du caractère interne ou externe de la fuite. La première donne une information quant à la formation de la fuite qui est conditionnée à la rencontre de certains critères. Dans un premier temps, nous avons estimé que cette information ne relevait pas directement de l'expression de la fuite en tant que telle, mais qu'elle relevait davantage du contexte ou des circonstances.

La seconde prend la forme des expressions « interne » ou « en ligne » pour signifier que la fuite est interne. Cette information est considérée comme importante pour l'expert, car elle donne une information quant à la gravité de la fuite, une fuite interne est généralement moins grave qu'une fuite externe. Nous avons considéré que nous sortons ici de l'expression de la fuite à proprement parler. Cela concerne plutôt une opinion d'un expert, quant à l'interprétation des informations transmises en raison de ses connaissances de spécialiste.

Cependant, suite à la première phase d'annotation et après avoir conversé avec l'expert qui a annoté les données, nous avons choisi de les inclure lors de la seconde phase d'annotation afin de mesurer leur utilisation seuls et avec les autres éléments. Nous avons donc ajouté trois étiquettes pour représenter ces informations supplémentaires : *Condition d'apparition, Interne* et *Externe*. Si la présence de la première s'est confirmée dans plusieurs rapports, les deux autres n'ont été que très peu employées. L'étiquette *Interne* a été utilisée à une seule reprise et l'étiquette *Externe* n'a jamais été employée. Nous n'avons d'ailleurs jamais observé de mention du mot « externe » pour caractériser une fuite dans le corpus. Cette information n'a donc pas été retenue dans la grille d'annotation définitive.

Étiquette	**Phase 1**	**Phase 2**	**Total**
Localisation moyenne	34	63	97
Localisation précise	40	34	74
Débit	12	10	22
Dispositif de détection	13	4	17
Quantité	6	11	17
Fluide	11	5	16
Localisation géographique large	8	8	16
Condition		8	8
Interne		1	1
Externe		1	1
Total	124	144	269

TABLE 8.7 – Décompte des étiquettes pour l'annotation de la fuite

Dans la Table 8.7, nous pouvons observer le décompte des différents segments pour les deux phases et le total des deux. Nous rappelons ici qu'une étiquette peut apparaître à plusieurs reprises dans un rapport. Nous obtenons respectivement 124 et 144 annotations pour les phases 1 et 2, soit 268 étiquettes au total. Lors de la première phase, les étiquettes *Condition d'apparition, Interne* et *Externe* n'étaient pas incluses dans la grille et sont donc grisées dans la Table. Nous pouvons observer que les étiquettes *Localisation précise* et *Localisation moyenne* sont significativement plus utilisées que les autres dans les rapports. En effet, comme nous l'avons évoqué précédemment, l'emplacement de la fuite est l'élément qui semble le plus important à mentionner dans ces rapports. D'ailleurs, lors des deux annotations, chaque rapport inclut au minimum une étiquette de localisation. Dans les résultats de l'effacement manuel, nous avions déjà pu repérer que l'énoncé minimal de ces rapports est Fuite + L (Localisation) ou L + Fuyard. Nous observons ensuite une baisse importante des

Résidus	Loc. moyenne	Détection	Débit	Fluide	Loc. large	Quantité	Condition
Loc. précise	-0,21	-0,63	1,83	-0,96	-0,55	0,77	-3,54
Loc. moyenne		0,53	4,65	-0,37	-1,09	-0,17	-4,41
Détection			0,19	-1,90	0,24	-0,32	-1,93
Débit				-1,15	-1,52	2,39	-2,20
Fluide					0,83	-1,05	-1,34
Loc. large						-0,62	-1,88
Quantité							-1,93
Condition							

TABLE 8.8 – Résidus de Pearson

effectifs avec entre 8 et 22 annotations pour le reste des étiquettes, à l'exception de *Interne* et *Externe* qui n'y figurent qu'une seule fois chacun.

Avec l'obtention de ces annotations, nous avons voulu observer les cooccurrences des différentes étiquettes, et s'il existe une dépendance entre deux types d'éléments. Pour ce faire, nous avons construit une matrice de cooccurrence et calculé les *résidus de Pearson*. Cette métrique permet d'observer la différence entre les valeurs attendues sous hypothèse d'indépendance et les valeurs observées. Avant d'effectuer ce calcul, nous avons toutefois supprimé les annotations *Interne* et *Externe* des données, car les étiquettes n'ont pas été retenues dans la grille d'annotation finale. Nous avons aussi ajouté trois occurrences de l'étiquette *Condition d'apparition* parmi les annotations des 50 rapports de la première phase à la suite de l'ajout de cette classe dans la grille d'annotation. Nous obtenons un total de 270 annotations.

Les résultats du calcul des *résidus de Pearson* sont dans la Table 8.8. Les résultats sont considérés comme significatifs au seuil de $p<0,05$ s'ils sont supérieurs à 2 ou inférieurs à -2 en valeur absolue. Parmi les scores obtenus, seuls deux sont donc supérieurs à 2. En effet, nous observons une forte corrélation entre les étiquettes *Localisation moyenne* et *Débit*, et à nouveau entre *Débit* et *Quantité*. Pour cette dernière association, il est aisément concevable que lorsqu'une information sur la quantité est indiquée, une précision quant au débit observé soit ajoutée, comme dans l'exemple suivant : `MICRO-FUITE SUR EAPO (environ 10-3)`. Dans les scores inférieurs à -2, nous trouvons aussi trois cas pour la catégorie *Condition d'apparition d'une fuite* en combinaison avec *Localisation précise*, *Localisation moyenne* et

Débit. Cependant, nous avons vu que cette étiquette est la moins fréquente (toute la colonne est en nombre négatif) et ainsi, étant donné la forte fréquence des étiquettes *Localisation moyenne et précise*, voire même *Débit*, l'écart par rapport à l'attendu se creuse.

Nous avons donc pu élaborer un modèle de la fuite dans un environnement technique et le mettre en application *via* l'annotation d'un échantillon du corpus. Ce processus d'annotation en deux étapes, a permis de faire émerger des éléments pertinents qui ont échappé aux procédés de repérage des *patterns*. Toutefois, cette annotation a été effectuée sur 100 rapports seulement. Et, si ce chiffre se révèle suffisant pour valider la grille d'annotation proposée, il ne permet pas de mettre en place une annotation automatique du corpus et la disponibilité des experts ne permet pas d'effectuer une annotation plus extensive.

8.4 Vers un *vépole* de la fuite ?

Maintenant que nous avons pu repérer les différents éléments qui constituent l'expression de la fuite dans un environnement technique, la question de l'adéquation avec la méthode TRIZ peut être envisagée. Lors de la présentation de TRIZ dans le chapitre 2, nous avons pu explorer les différents outils qu'elle propose. Parmi ces outils, nous avons mis l'accent sur la manière dont peut être formalisé un problème technique dans cette méthode, c'est-à-dire le *vépole*. Pour rappel, la version minimale du *vépole* est composé de trois éléments : l'outil, le champ et le produit [74]. Le premier élément, l'outil, émet une action, le champ, qui influe sur le produit qui subit donc un changement d'état.

Dans le chapitre 2, nous avons utilisé l'exemple du pneu crevé pour illustrer cette modélisation. Nous retrouvons cet exemple dans la Figure 8.6. Le problème du pneu crevé rejoint le problème de fuite une fois ramené à une définition concrète du problème. Nous sommes en présence d'un objet qui ne retient pas correctement un fluide, le pneu ne retient pas correctement l'air. Une fois transposé au domaine théorique et abstrait comme le propose TRIZ, cela signifie qu'un solide (le pneu) ne retient pas correctement (l'action en question est retenir) un gaz (l'air).

Pour ramener ces éléments à nos données, il faut pouvoir dresser un parallèle entre les éléments de la fuite et les éléments qui constituent le *vépole* dans le domaine des réalités industrielles. Comme nous avons pu le voir avec l'exemple du pneu crevé, le *vépole* type de la fuite est donc un outil qui ne retient pas suffisamment / correctement un produit. Parmi les étiquettes qui constituent notre *frame*, le produit, soit l'élément

74. Ces éléments peuvent prendre diverses appellations en fonction des sources.

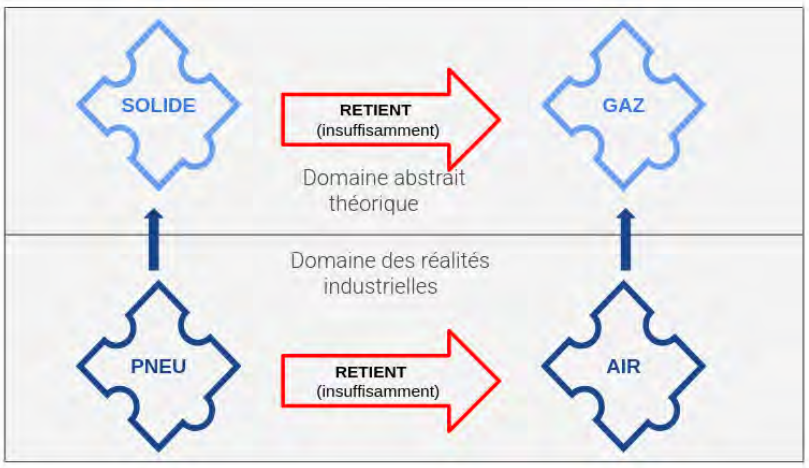


FIGURE 8.6 – *Vépole* du pneu crevé

qui n'est pas correctement retenu, est naturellement le fluide. Concernant l'outil, celui-ci correspond à l'emplacement de la fuite, soit l'élément *Localisation précise*. Si ces deux éléments sont présents, il est donc possible de reconstituer le *vépole*, au moins au niveau réalité de la situation. Ainsi, nous pouvons reprendre le dernier exemple de la Figure 8.5 : `FUITE DE LIQUIDE DE REFROIDISSEMENT (LEGER SUINTEMENT) CONSTATEE AU NIVEAU DE LA PURGE D'AIR DU CIRCUIT DE REFROIDISSEMENT DE GE1 DU GST2`. Le *vépole* en résultant est donc :

> PURGE D'AIR - RETIENT (INSUFFISAMMENT) - LIQUIDE DE REFROIDISSEMENT.

L'étape suivante est de transposer ces éléments au niveau d'abstraction où se situe la méthode TRIZ (qui permet d'obtenir des solutions). Pour l'élément produit, faisant référence au composant qui retient le fluide, c'est systématiquement un solide. En revanche, pour le fluide en question, nous ne disposons pas systématiquement d'indication *via* l'annotation ou autre pour déterminer automatiquement *a priori* si c'est un liquide ou un gaz. Pour ce faire, l'usage d'un thésaurus ou d'un dictionnaire pourrait compléter ce procédé.

Une fois le *vépole* dans le domaine abstrait obtenu, il est désormais possible de faire usage de la ressource I4KN, que nous avons présentée dans le chapitre 2, afin de parcourir les solutions pour l'action « Retenir un liquide » ou « Retenir un gaz ». Dans la Figure 8.7), nous pouvons observer les pages dI4KN qui correspondent à ces deux actions. Elles proposent chacune une liste de solutions qui sont elles-mêmes instanciées dans des pages individuelles. Pour l'action « Retain a liquid », ces solutions sont elles-mêmes classées en plusieurs catégories.

Ainsi, nous pouvons observer en Figure 8.8, la chaîne de traitement complète pour la fuite dans un environnement technique. Celle-ci permet de reconstituer le *vépole* type de la fuite et d'avoir accès à un ensemble de solutions afin de résoudre le problème posé par la situation par le biais de la ressource I4KN.

8.5 Conclusion

L'étude de la catégorie *Fuite d'un liquide ou d'un gaz* nous a menés à reconsidérer l'adéquation du *frame* préalablement sélectionné dans *FrameNet*, à savoir *Fluidic motion*. Bien que ce *frame* inclue le déclencheur *leak* dans ses unités lexicales, il nous a semblé ne pas correspondre à la situation de fuite telle qu'elle est exprimée dans notre corpus.

De ce fait, nous proposons une méthode pour l'étude de l'expression de la fuite dans un environnement technique fondée sur trois approches complémentaires. Cette méthode s'est focalisée sur l'émergence de *patterns* récurrents au sein du sous-corpus de la fuite, d'une part par une méthode d'effacement manuel et automatique et de l'autre par une analyse syntaxique en dépendance. Cela nous a permis d'observer les structures employées pour exprimer la fuite. D'autre part, un entretien avec un expert est venu compléter ces deux premières approches afin d'éclaircir les éléments de lexique et confirmer ce qui est exprimé.

À partir de ces trois approches, nous avons donc pu faire émerger un *frame de la fuite dans un environnement technique* qui implique sept *frame elements*. Deux phases d'annotation par un expert ont été mises en place afin de valider ce *frame*. Il en est ressorti l'apparition d'un huitième *frame element* jusque-là passé inaperçu et moins fréquent dans les données.

De ces analyses et de ces annotations est ressorti le schéma minimal de l'expression de la fuite dans notre corpus, soit le marqueur ainsi qu'un élément de localisation. En effet, la localisation est omniprésente dans chacun des rapports et parfois seule présente. Cela renvoie à la vocation du REX et des FA : faire un rapport pour apporter une solution au problème. Dans le cas présent, l'importance est donc mise sur le fait de pouvoir, à partir du rapport, retrouver l'emplacement de cette fuite pour en effectuer le diagnostic, amenant ainsi à en donner la localisation, parfois de façon très précise jusqu'à la vanne concernée. En observant les énoncés du champ « Action(s) réalisée(s) » pour cette catégorie, nous pouvons aussi voir que la solution mise en place le plus couramment est le remplacement du composant. Si celle-ci est la solution la plus privilégiée, cela justifie d'autant plus le besoin de localiser la fuite à la lecture d'un rapport, et donc de véhiculer cette information de manière précise et immédiate. Les éléments de localisation sont donc nombreux mais restent flous avec un recours à des localisations imprécises comme le montre l'utilisation des prépositions « au

niveau de », « sur », « à »... Il est parfois difficile d'identifier le lieu précisément car la trace de la fuite peut être indirecte, par exemple une baisse de pression.

La grille d'annotation que nous proposons reste pourtant à ce stade de notre étude non automatisable, puisqu'elle requiert obligatoirement une annotation experte. Pour l'annotation des marqueurs, nous avons pu mettre en place une annotation avec des non-experts du domaine puisque celle-ci s'appuyait sur des expressions du langage commun, même si elles peuvent parfois être trompeuses quand on ignore la phraséologie. En revanche, nous avons pu voir dans ce chapitre, et dans cette thèse plus généralement, que les éléments du système dont il est question ne sont pas accessibles sans expertise du domaine. Nous avons aussi pu voir dans ce chapitre que les structures peuvent être utilisées de façon indifférenciée et ne peuvent donc pas servir de pilier pour une compréhension de la nature des éléments en question. Cependant, si une annotation experte d'un *dataset* suffisamment conséquent est effectuée, une tâche de reconnaissance de séquences telles que nous l'avons effectuée pour les marqueurs pourrait être mise en place pour les rapports de la fuite. Pour cette tâche, nous faisons l'hypothèse que le redressement de données que nous avons effectué au chapitre 7 aurait plus de poids dans les performances du système puisque les séquences annotées seraient cette fois beaucoup plus porteuses de bruit que les simples marqueurs.

Cette grille s'éloigne cependant partiellement de l'objectif de modélisation en *vépole* puisqu'elle couvre plus d'éléments que ceux nécessaires au *vépole type* de la fuite. Cependant, il nous est apparu nécessaire de définir les éléments qui constituent la fuite dans leur entièreté afin d'obtenir une description formelle de cette expression plutôt que de se focaliser sur certains éléments précis. En effet, nous pensons qu'en partant des données directement dans une approche inductive, cela nous a permis d'éviter un biais dans la caractérisation des éléments du *frame*. De plus, cela nous a permis de repérer les différents niveaux de localisation exprimés plutôt que de s'arrêter à la simple recherche d'une localisation non définie et spécifiée pour la modélisation. Par ailleurs, en dehors du *vépole* type, il est possible d'enrichir ce même *vépole* avec des informations complémentaires. Nous n'avons pas présenté ce type de cas jusqu'alors puisque cela dépasse le cadre de notre expertise de TRIZ et la modélisation que nous nous sommes fixée comme objectif. Cependant, l'ajout d'information peut permettre de préciser le dysfonctionnement dans le *vépole* et en conséquence spécifier le champ des solutions qui peuvent y être apportées.

Une dernière limite entre la grille et la modélisation en *vépole* existe puisqu'à ce stade, il manque encore une étape pour passer du fluide annoté à la caractérisation de ce même fluide en *gaz* ou en *liquide* qui serait nécessaire pour accéder aux solutions proposées par TRIZ et I4KN. Nous pensons cependant qu'une liste de lexiques pourrait permettre de résoudre ce manque. En faisant coïncider les éléments *Fluide* extraient des données et des listes

préétablies (et potentiellement enrichies de ces mêmes extractions), cette caractérisation devient faisable et permet ainsi d'avoir tous les éléments pour obtenir les solutions au problème.

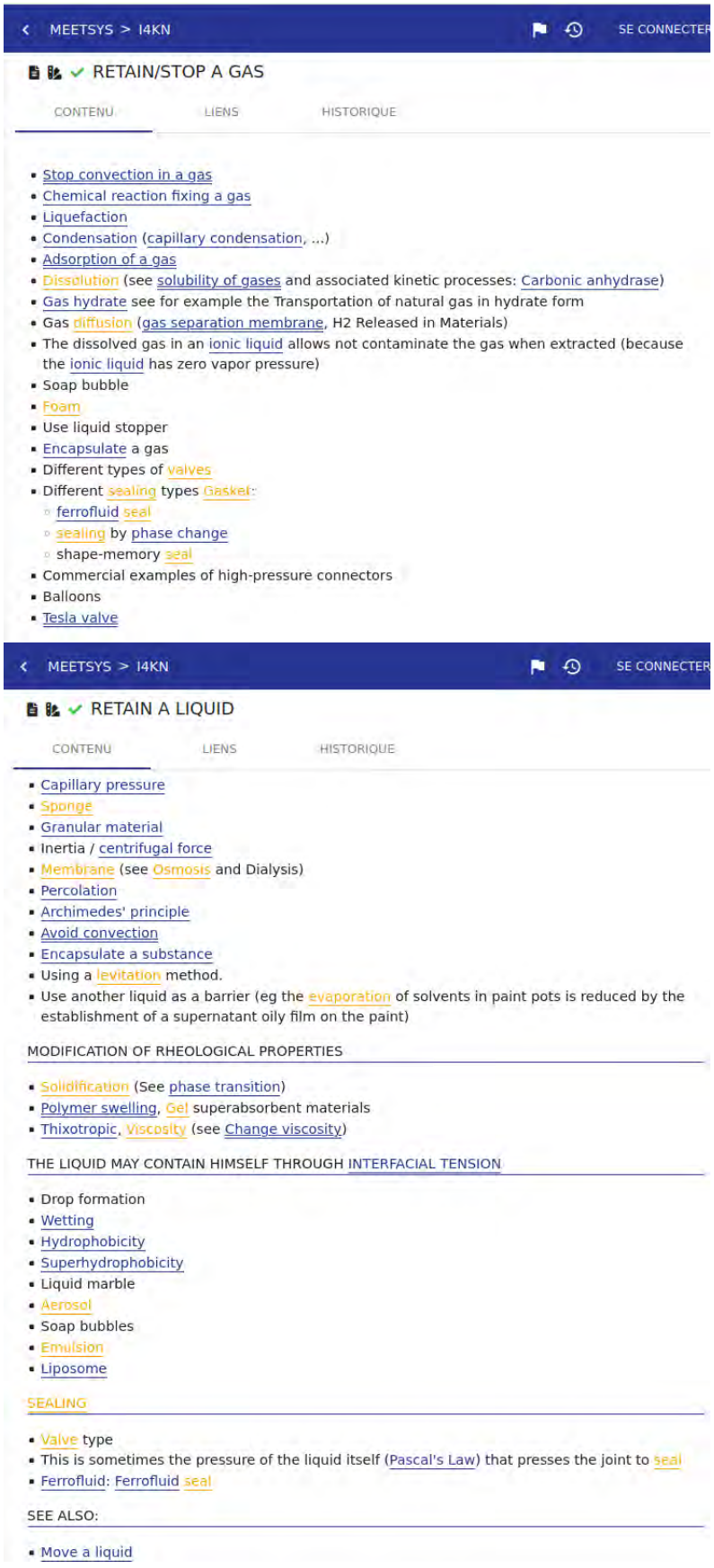


FIGURE 8.7 – Retenir un liquide ou un gaz

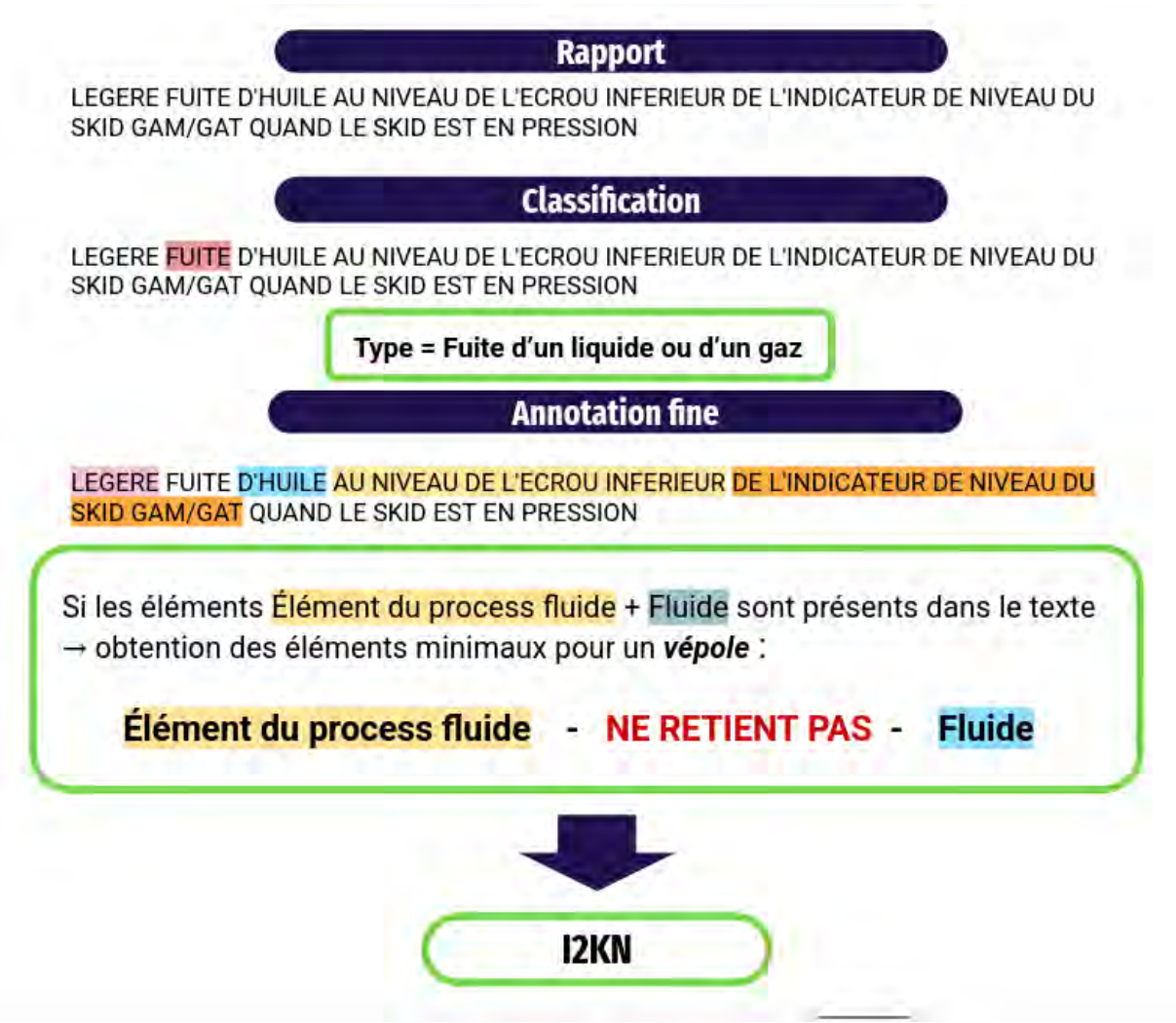


FIGURE 8.8 – Chaîne de traitement de la fuite dans un environnement technique

Chapitre 9

Automatiser l'annotation d'un *frame* : la classe Obstacle

Sommaire

Dans ce chapitre, nous proposons l'étude d'une autre classe de la typologie qui est : *Présence d'un obstacle*. De même que dans le chapitre 8 sur la fuite, nous souhaitons réaliser une analyse fine du contenu de l'énoncé afin d'observer comment la notion d'obstacle est exprimée dans un environnement technique ou industriel. Ce travail s'effectue aussi dans l'objectif de parvenir à une modélisation de ce type de problème *à la TRIZ*, ouvrant ainsi des possibilités pour la résolution de ce type de problèmes techniques. La justification du choix de cette classe pour une étude approfondie trouve des similitudes avec celui de la catégorie *Fuite d'un liquide ou d'un gaz*. En effet, nous avons exclu certaines catégories peu adaptées à la modélisation, pour les raisons que nous décrivons dans le chapitre précédent. Par ailleurs, de même que pour la fuite, nous notons une régularité dans les structures d'expressions de l'obstacle qui nous ont incité à choisir cette classe, mais aussi une facilité à identifier

les différents éléments présents. Ce dernier point implique aussi que, contrairement à la fuite, nous n'avons pas besoin de faire un travail exploratoire poussé afin de déterminer les différents éléments qui composent l'énoncé. De ce fait, dans ce chapitre, nous proposons une approche différente qui vise plutôt à une automatisation du procédé d'étiquetage des éléments présents dans les énoncés de la catégorie *Présence d'un obstacle*.

Dans un premier temps, nous explorons deux *frames* de *FrameNet* et nous mesurons leur adéquation avec la classe *Présence d'un obstacle*. Ensuite, nous passons en revue la méthode d'étiquetage que nous avons choisie, qui est fondée sur l'utilisation d'un LLM. Enfin, nous présentons les propositions et les limites d'une modélisation en *vépole*.

9.1 Adaptation du *frame*

Les énoncés de la catégorie *Présence d'un obstacle* correspondent aux situations pour lesquelles une action ne peut pas être accomplie en raison d'un élément bloquant. Elle est proche de la catégorie *Action difficile ou impossible* mais apporte une précision quant à la raison de la difficulté ou impossibilité. Elle recouvre des marqueurs tels que « blocage », « obstruction » ou « gêne » (lorsqu'il est question d'une gêne physique dans l'accomplissement d'une action) et nous pouvons en voir trois exemples dans la Figure 9.1. Pour la catégorie *Présence d'un obstacle*, deux *frames* nous ont paru recouvrir les énoncés de cette situation : *Hindering* et *Preventing or letting*. Dans cette section, nous présentons d'abord *Hindering* puis *Preventing or letting* et comment chacun peut correspondre à cette classe.

LE SUPPORT VISSE DE LA TRESSE DE MASSE ENTRE LES AXES Z ET YN BLOQUE L'INTEGRATION D'UNE VIS DE FIXATION MFD (EQUERRE RMM).

PRESENCE D'UN COPEAU VISIBLE EN SORTIE DUMP COOLING ENTRE TPO ET VCH OBTURANT PARTIELLEMENT UN DES ORIFICICES DE SORTIE DUMP.

CABLE BERB CLIENT DANS CADRE MCT EMPECHE LA MISE EN PLACE DE LA POE.

FIGURE 9.1 – Exemples de rapports de la catégorie *Présence d'un obstacle* avec annotation du marqueur

9.1.1 Le cadre *Hindering*

Le premier *frame* issu de *FrameNet* qui nous a paru en accord avec notre classe est *Hindering*. Il dispose de trois éléments noyaux qui sont *Protagonist*, *Action* et *Hindrance*, visibles dans le schéma d'annotation de la Figure 9.2. La définition de ce *frame* est la suivante : « Dans ce cadre, un Obstacle (Hindrance) rend plus difficile pour un Protagoniste

(Protagonist) d'accomplir l'Action désirée ou un Obstacle affecte directement une action où l'action peut être un acte naturel ou les actions de plusieurs individus dans un effort collectif. Il est important de noter que ces unités lexicales ne signifient pas que le Protagoniste n'a pas été en mesure d'achever l'Action ou que l'Action n'a pas eu lieu. L'Obstacle peut indiquer un manque d'accès ou un obstacle à la progression naturelle du protagoniste ou de l'action. » (*In this frame a Hindrance makes it more difficult for a Protagonist to complete their intended Action or a Hindrance directly affects an Action where the Action can be a natural act or the actions of many individuals in a collective endeavor. Crucially, these lexical units do not entail that the Protagonist was unable to complete the Action or that the Action did not occur. The Hindrance may indicate a lack of enablement or an obstacle on the natural progression of the Protagonist or Action*).

Frame Element	Core Type
Action	Core
Circumstances	Extra-Thematic
Concessive	Extra-Thematic
Degree	Extra-Thematic
Duration	Extra-Thematic
Explanation	Extra-Thematic
Frequency	Extra-Thematic
Hindrance	Core
Manner	Peripheral
Means	Peripheral
Place	Peripheral
Protagonist	Core
Time	Peripheral

FIGURE 9.2 – Schéma d'annotation de *Hindering*

L'élément ***Action*** est défini comme ce que le *Protagonist* souhaite effectuer mais qui est rendu difficile par l'élément *Hindrance* (l'obstacle). L'élément ***Hindrance*** est la force (ou absence de force) qui rend l'Action difficile ou moins probable d'être effectuée. Le ***Protagonist*** est l'entité douée de sensations qui tente d'effectuer l'*Action* rendue difficile par l'*Obstacle*. Nous pouvons observer deux exemples annotés avec ces éléments issus de la grille Figure 9.2 pour les phrases « Le nouveau barrage a entravé le flux de l'eau » et « Malcolm a été gêné par son problème d'alcoolisme » extraits de *FrameNet* en Figure 9.3. Ces exemples permettent d'illustrer différents types d'obstacles que peut recouvrir ce cadre, que ce soit un obstacle physique, comme dans l'exemple 1, ou plus abstrait, comme dans l'exemple 2.

FIGURE 9.3 – Exemple annoté pour le *frame Hindering*

Les différentes unités lexicales pour ce cadre les plus propices à être retrouvées dans les rapports d'anomalie sont : *hinder, obstruct* et *block* (« gêner, obstruer et bloquer »). Nous pouvons les retrouver dans la liste des marqueurs annotés automatiquement en Annexe B. Parmi les rapports catégorisés dans *Présence d'un obstacle*, nous pouvons trouver les descriptions d'anomalies suivantes :

— RACCORDEMENT DU FLEXIBLE HFX1605 SUR L'INTERFACE AD COTE PAC-02 EST **BLOQUE** PAR LE FLEXIBLE D'INTERCOMMUNICATION.
— PRESENCE D'UNE GRILLE ET D'UNE PLAQUE ALU **OBSTRUANT** LA LIGNE DE PRESSURISATION DU LOCAL 3.1 (CXX).
— SUR EAP 2 : SACHETS D'EPROUVETTES 4C VISSES A L'INTERIEUR SOUS LA FERRURE EAP (FLAMME VIOLETTE N°139) **GENE** LE PASSAGE DE L'OT FOURREAU.

Cependant, il existe certaines limites quant à l'adéquation de ce cadre avec la classe *Présence d'un obstacle*. En effet, le premier élément discordant que nous pouvons citer est le *Protagonist* qui n'est jamais évoqué dans les rapports (ou quasiment jamais si nous prenons le peu d'occurrences de la première personne que nous avons pu mentionner dans le chapitre 4). De ce fait, il n'existe pas d'entité douée de sensations souhaitant accomplir l'action mentionnée dans les énoncés. Par ailleurs, le terme *hindering* évoque une obstruction, un blocage qui entraîne une difficulté à effectuer une action. Cependant, dans la classe *Présence d'un obstacle*, nous incluons à la fois les difficultés, mais aussi les impossibilités à accomplir l'action en question. De même, nous incluons ces deux notions au sein de la classe *Action difficile ou impossible* qui, comme nous l'avons évoqué dans le chapitre 5, est proche de *Présence d'un obstacle*.

Concernant cette dernière limite, le *frame Preventing or letting* semble combler ce manque. Nous présentons donc ce second cadre dans la section suivante.

9.1.2 Le cadre *Preventing or letting*

Le *frame Preventing or letting* est aussi pertinent par rapport à la classe *Présence d'un obstacle* puisqu'il concerne les cas pour lesquels un obstacle empêche la réalisation d'un

événement. Le *frame* inclut aussi les situations pour lesquelles l'obstacle ne se réalise pas et on laisse l'événement se produire. Ce dernier cas ne recouvre cependant pas les situations de la classe de notre typologie. La définition de ce cadre est la suivante : « Un Obstacle Potentiel (Potential_hindrance) ou un Agent (via ses actions) empêchent un Événement (Event) de se produire, ou malgré la possibilité d'empêcher l'Événement, ne le font pas (*A Potential_hindrance or an Agent (via their actions) keeps an Event from taking place, or despite the possibility of prevent the Event, do not do so*).

Frame Element	Core Type
Agent	Core
Allowed_category	Extra-Thematic
Circumstances	Extra-Thematic
Degree	Extra-Thematic
Depictive	Extra-Thematic
Event	Core
Event_description	Extra-Thematic
Explanation	Extra-Thematic
Frequency	Extra-Thematic
Manner	Extra-Thematic
Means	Extra-Thematic
Period_of_iterations	Extra-Thematic
Place	Peripheral
Potential_hindrance	Core
Purpose	Extra-Thematic
Time	Peripheral

FIGURE 9.4 – Grille d'annotation pour le *frame Preventing or letting*

Nous retrouvons donc cette fois encore trois éléments noyaux qui sont : *Potential_hindrance, Agent* et *Event*, comme indiqué dans la Figure 9.4. L'élément ***Potential_hindrance*** est la force qui empêche l'Événement de se produire, ou inversement l'autorise à se produire. Cet élément rejoint *Hindrance* du cadre *Hindering*, mais dans le cas des rapports de cette catégorie, il n'est jamais question d'obstacle « potentiel ». L'***Agent*** est une entité douée de sensations qui empêche l'Événement de se réaliser. Il est le pendant de *Potential_hindrance* mais dans le cas où l'élément est une entité douée de sens. Enfin, l'***Event*** est ce qui ne se produit pas en raison de l'*Agent* ou du *Potential_hindrance*. Ce dernier possède des similarités avec l'élément *Action* du cadre *Hindering*. De ce cadre, le seul élément qui ne trouve pas d'équivalence dans le cadre *Preventing or letting* est le *Protagonist*. Cependant, ce dernier ne se réalise jamais dans le cadre des rapports d'anomalies et n'est donc pas pertinent pour notre utilisation de ces *frames*. Parmi les unités lexicales associées à ce *frame*, et que nous

retrouvons dans les rapports d'anomalies, nous pouvons citer « *prevent, stop, keep from* » (« empêcher, stopper »).

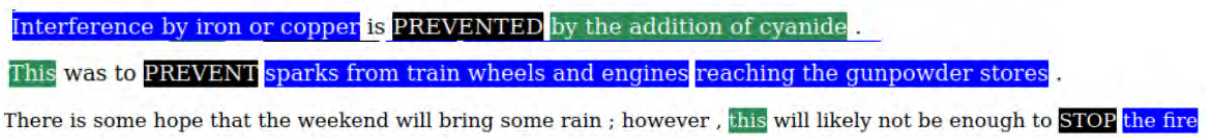


FIGURE 9.5 – Exemples de *FrameNet* annotés pour le *frame Preventing or Letting*

Dans la Figure 9.5, nous pouvons observer trois exemples issus de *FrameNet* annotés selon la grille du cadre *Preventing or letting* (nous avons uniquement choisi des exemples qui impliquent la présence d'un obstacle). Nous retrouvons les deux éléments noyaux que sont *Potential_hindrance* (en vert) et *Event* (en bleu). Nous pouvons voir que les situations recouvrent des domaines d'application variés comme celui de l'inhibition chimique avec le premier exemple. Nous observons aussi que l'élément *Event* peut couvrir une large séquence du texte, alors que l'élément *Potential_hindrance* est plus concis, c'est un élément précis de l'énoncé.

Parmi les rapports d'anomalies, nous listons ici quelques exemples issus de la catégorie *Présence d'un obstacle* qui sont couverts par ce cadre :

— RESSORTS D'EJECTION **EMPECHENT** L'INSTALLATION DES OUTILS DE MISE EN TENSION SANGLE (A LA DROITE DU BOULON)
— LA PRESENCE DU DOME RLH2 **EMPECHE** DE PASSER LA PERCEUSE POUR CHASSER LES RIVETS POUR AMENAGER LES LOGEMENTS DES BUSES DE PASSAGE DE CLOISON CONE CASE.
— LORS DU CONTROLE DE LA FONCTION DU TRANSFERT DES SERVO-DISTRIBUTEURS, LA PA **STOPPE** L'ESSAI (DECLENCHEMENT MCA) LORS DU CONTROLE DE LA 4EME VOIE (EPH2 VERIN V).

9.1.3 Deux *frames*, une catégorie

L'existence de ces deux *frames* peut entraîner plusieurs questions quant à la catégorisation de notre typologie, mais aussi quant à la notion d'obstacle.

Concernant le premier point, nous rappelons ici que la construction de la typologie s'est faite sans la prise en compte des *frames* définis par *FrameNet*. En effet, l'approche utilisée a été celle d'une linguistique de corpus. Nous avons choisi d'effectuer une étude *bottom-up* ou *corpus-driven*, en partant des données uniquement pour en faire ressortir des régularités, sans *a priori* préalable. Par ailleurs, le corpus que nous étudions est très éloigné de ceux ayant été utilisés pour la construction de *FrameNet* (c'est-à-dire le British National Corpus et

l'American National Corpus (Baker, 2009)). De ce fait, il n'est pas étonnant d'obtenir une catégorisation différente de celle de *FrameNet*. Bien que cela puisse soulever des questions, nous n'estimons pas pour autant que la catégorisation choisie pour la typologie d'expression d'un problème technique est à remettre en cause. Dans le cas de ces rapports, ces deux *frames* sont agglomérés en une catégorie, puisque nous n'effectuons pas de différenciation entre un obstacle qui « gêne » une action, et un obstacle qui « empêche » un événement, soit entre un obstacle partiel et un obstacle total. Cela renvoie aussi à une culture du REX dans laquelle il est important de rapporter une entrave qui pourrait avoir des conséquences importantes, et ce avant qu'elle se révèle être un obstacle complet. Par ailleurs, dans les rapports, cette différenciation n'est pas systématiquement exprimée. Dans l'exemple suivant, « PRESENCE D'UNE GRILLE ET D'UNE PLAQUE ALU OBSTRUANT LA LIGNE DE PRESSURISATION DU LOCAL 3.1 (CXX). », l'unité lexicale qui déclenche du *frame* est « OBSTRUANT » (*obstruct.v*), qui selon *FrameNet* dépend du *frame Hindering*. Cependant, dans le rapport, il n'est pas dit explicitement si l'obstruction est partielle ou totale. Est-ce que l'*Action* (qui n'est pas mentionnée dans le texte) n'est plus réalisable ou simplement gênée par cette obstruction ? Ainsi, l'appartenance à l'un ou l'autre *frame* est floue.

Pour notre travail, nous ne faisons donc pas de différenciation entre les rapports qui dépendent du *frame Hindering* ou de *Preventing or letting* puisqu'ils appartiennent tous à la même classe *Présence d'un obstacle*. En revanche, dans le cadre des opérations d'automatisation de l'annotation telle que nous l'avons mise en place, ou même d'une annotation manuelle si envisagée, il a été nécessaire de faire un choix. Il faut pouvoir passer des *frames* à une seule grille d'annotation. Nous n'avons pas souhaité étudier ces énoncés afin de faire émerger un *frame* comme nous avons pu le faire pour la catégorie *Fuite d'un liquide ou d'un gaz* puisque les deux *frames* de *FrameNet* nous semblent correspondre à la situation en question. De plus, s'appuyer sur cette ressource, issue d'un travail extensif et validé, permet de faciliter la mise en place d'une automatisation. Par ailleurs, nous avons aussi souhaité nous appuyer sur les « connaissances » d'un LLM à propos de *FrameNet*. Nous avons posé ainsi la question au LLM *Llama 3*, que nous employons pour l'annotation automatique, pour savoir s'il a connaissance de *FrameNet*, ce à quoi il a répondu par la positive :

> *Prompt* : Est-ce que tu connais FrameNet ?
>
> *Réponse* : Oui, je suis familiarisé avec FrameNet ! FrameNet est un projet de recherche en linguistique et en intelligence artificielle qui vise à créer une base de données de cadres sémantiques (frames) pour représenter les connaissances sur le monde réel.

> FrameNet a été créé par la Northwestern University et est maintenant maintenu par l'Université de Californie à Berkeley. Il contient des millions d'exemples de phrases annotées avec des informations sur les cadres sémantiques qui les sous-tendent, ainsi que des relations entre ces cadres.
>
> Les cadres sémantiques (frames) sont des modèles conceptuels qui capturent les relations entre les concepts et les événements du monde réel. Ils peuvent être utilisés pour analyser le sens de phrases, identifier les relations entre les concepts, et même générer du texte cohérent.
>
> FrameNet est un outil puissant pour les chercheurs en linguistique, en intelligence artificielle et en traitement automatique des langues naturelles (TALN). Il a été utilisé dans de nombreux domaines, tels que la génération de textes, l'analyse de sentiments, la reconnaissance de la parole et la compréhension du langage.
>
> Si vous avez des questions spécifiques sur FrameNet ou si vous cherchez à en savoir plus, n'hésitez pas à me demander !

Cependant, le fait d'être entre deux *frames* demande de faire un choix pour l'annotation. Nous avons décidé de faire usage de *Hindering*, usage que nous décrivons dans la section suivante. Ce *frame* est en effet parvenu à notre attention préalablement à celui de *Preventing or letting* et lors de l'élaboration de la méthodologie pour le traitement de cette catégorie.

9.2 Automatiser le repérage des éléments du *frame*

Dans cette section, nous présentons la méthodologie que nous avons mise en place pour effectuer une annotation automatique des cas de type *Présence d'un obstacle* avec un LLM. Cette tâche est similaire à du *Frame Semantic Role Labeling* (Gildea et Jurafsky, 2000), mais nous nous sommes permis de prendre certaines libertés avec le *frame* en question.

9.2.1 Quels sont les éléments à repérer ?

Dans la section précédente, nous avons pu voir comment deux *frames* recouvrent le périmètre de la catégorie *Présence d'un obstacle*. Nous avons à cette occasion établi que, pour cette tâche, nous avons choisi d'utiliser le cadre *Hindering* afin de guider la sélection des éléments à annoter dans le texte. Dans ce *frame*, il existe trois éléments noyaux : le *Protagoniste*, l'*Hindrance* et l'*Action*. Le premier, nous avons pu le voir, n'est pas présent dans les rapports et n'est donc pas à prendre en compte dans cette tâche. Le second, l'*Hindrance* est l'élément qui bloque la réalisation de l'action et n'est pas systématiquement présent dans

l'énoncé. Et le dernier, l'*Action* est l'élément qui ne se réalise pas, à cause de l'*Hindrance*, et peut représenter une large portion de texte.

En parallèle de la vision du *frame*, nous avons souhaité avoir une approche *bottom-up* afin de repérer les différents éléments qui constituent le *pattern* minimal de l'obstacle. Nous sommes partis d'exemples de rapports afin de repérer les éléments récurrents minimaux. Des exemples sont visibles dans la Table 9.1. Les éléments minimaux repérés constituant les situations d'obstacles apparaissent être :

— L'objet qui bloque : OBSTACLE ;
— L'action bloquée : ACTION ;
— L'objet patient de l'action bloquée : CIBLE.

Nous pouvons voir que ces exemples peuvent inclure toutes les combinaisons d'éléments et avec 1, 2 ou 3 éléments mentionnés. Nous n'avons cependant pas trouvé d'exemple dans le corpus avec uniquement l'action bloquée ou l'objet qui bloque. Nous pouvons ici poser la question de l'exemple 6, pour lequel seule l'action est présente. Selon notre typologie, une action difficile ou impossible à réaliser sans autre information appartient à une autre catégorie. La reconnaissance de ces catégories lors de la tâche d'étiquetage dépend cependant du marqueur et non des *frame elements* en présence. Ce même marqueur est présent sous forme de l'unité lexicale *hindrance* dans *FrameNet*, et du *frame Hindering*. Nous avons aussi établi que la catégorie *Action difficile ou impossible*, en bas du classement de la typologie, propose trop peu d'information pour être susceptible d'être formalisée en *vépole*. Nous savons donc déjà que, pour les cas où seule l'action est présente dans l'énoncé (en plus du marqueur) la formalisation ne sera pas possible. Nous pouvons aussi déjà supposer qu'avoir un seul élément, quel qu'il soit, risque d'être problématique pour une modélisation qui implique trois composants.

Afin de relier les éléments reconstitués dans l'élaboration des *patterns* minimaux de cette classe, mais aussi une approche *à la TRIZ* (en prévision de la modélisation), nous avons souhaité isoler différents éléments inclus dans l'élément *Action*, à savoir l'action dont il est question et ce sur quoi porte cette action. Nous avons pu voir que dans les *vépoles*, les champs (soit les actions) et les objets sur lesquels ils agissent sont distincts. Nous avons donc ajouté un troisième élément ***Target*** défini comme il suit : *Target* correspond au *patient* de l'action rendue plus difficile par l'*Hindrance*[75]. Dans la Figure 9.6, nous pouvons observer un exemple canonique que nous avons annoté avec la grille de *FrameNet* dans le premier cas, et avec l'ajout de l'élément *Target* dans le second. Nous avons ainsi séparé l'action de

75. Nous reprenons l'appellation *patient* telle que définie dans la Table 3.1 : « Affecté par l'action, subit un changement »

Index	Rapport	Patron
1	L'ACCES A LA PORTE DE LA BAIE REF-G.Y EST GENE PAR DES DALLES METALLIQUES ET DES TAPIS	**ACTION** à **CIBLE** EST GENE par **OBSTACLE**
2	BLOCAGE DU CONTREPOIDS SECONDAIRE DU BRAS LH2, PENDANT SA DESCENTE VERS L'AMORTISSEUR.	BLOCAGE du **CIBLE** pendant **ACTION**
3	LE CHEMINEMENT DU TUBE DE VENTILATION LE LONG DU FLEX D'ALIMENTATION SVE EMPECHE LA MISE EN PLACE CONFORMEMENT AU PLAN DE LA PT	**OBSTACLE** EMPECHE **ACTION**
4	OBSTRUCTION PARTIELLE DES EVENTS DES CAPOTS DAS : PRESENCE D'INSECTES (MOUSTIQUES) SUR LA FACE INTERNE DES TROIS MOUSTIQUAIRES.	OBSTRUCTION de **CIBLE** : **OBSTACLE**
5	BLOCAGE DANS SON LOGEMENT DE LA FERRURE D'IMMOBILISATION YN DU DAAR.	BLOCAGE de **CIBLE**
6		**ACTION** EST BLOQUÉ
7		**OBSTACLE** BLOQUE

TABLE 9.1 – *Patterns* minimaux de la présence d'un obstacle

son patient, à savoir ici « l'accès » et « la salle de bain ».

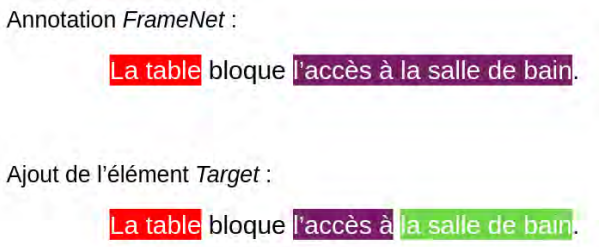


FIGURE 9.6 – Exemple annoté avec l'élément *Target*

Les trois éléments que nous cherchons à identifier automatiquement dans le texte sont donc : *Hindrance*, *Action* et *Target*. Notre objectif pour cette tâche est de savoir si un LLM est capable d'identifier ces différents éléments dans des textes techniques et bruités. Cette tâche pose déjà un certain nombre de difficultés. Nous faisons donc abstraction des éléments périphériques proposés dans *FrameNet* et qui peuvent être présents dans les rapports afin de nous concentrer sur ces éléments noyaux uniquement.

9.2.2 Méthodologie : une approche comparative

9.2.2.1 Présentation des *datasets*

La tâche que nous souhaitons effectuer est une annotation automatique de type *Frame Semantic Role Labeling* avec un LLM. Nous souhaitons à la fois évaluer la capacité d'un

LLM à faire ce type d'annotation, mais aussi l'impact des caractéristiques de notre texte (le domaine spécialisé et le bruit) sur ses performances. Nous avons déjà pu voir dans le chapitre 3 les travaux de Devasier *et al.* (2025) qui portent sur du *Frame Semantic Role Labeling* avec un LLM. Ils effectuent des tests sur deux *datasets* : FrameNet 1.7 (Ruppenhofer *et al.*, 2016) qui dispose d'un train set et d'un test set ayant été annoté en *frames* (Swayamdipta *et al.*, 2017) et YAGS (Yahoo ! Answers Gold Standard) (Hartmann *et al.*, 2017) afin de tester un corpus d'un domaine différent. Ils ont observé une baisse des résultats pour ce dernier, qu'ils attribuent d'une part à un manque de certains *frame elements* dans *FrameNet* qui correspondraient aux expressions trouvées dans le corpus. D'autre part, le corpus montre une grammaire non respectée et l'utilisation d'argot qui peuvent dégrader les performances du modèle.

Pour notre part, nous avons fait usage de cinq *datasets* avec différents textes pour la même tâche. Les cinq *datasets* utilisés pour cette tâche sont les suivants : un échantillon de rapports d'anomalies, les mêmes rapports d'anomalie normalisés, des textes extraits d'un autre corpus et deux jeux de textes synthétiques générés par un LLM. Nous présentons plus en détails ces *datasets* ci-dessous.

1. Le premier consiste en une sélection aléatoire de **25 rapports d'anomalies** catégorisés dans la classe *Présence d'un obstacle*. Pour cette tâche, nous avons fait le choix de prendre les rapports tels quels, sans segmentation. De ce fait, certains rapports peuvent inclure plusieurs phrases ou segments de phrases, dont certaines n'appartiennent pas à la catégorie, ou éventuellement à une autre catégorie. Par exemple, dans le rapport « LA PINOCHE ENTRE LE SUPPORT DE LA POE ET LA CHEMINEE EST CASSEE ET LA BAGUE SUPERIEURE EST BLOQUEE SUR LA CHEMINEE. », deux classes coexistent, *Dégradation - Usure - Saleté* et *Présence d'un obstacle*. Dans l'exemple « MONTAGE DU CAPOT N3 : SUR CAPOT N3 4 REP 223 / EPC 1 ECROU EST DESERTIE2/ LORS DE LA PRESENTATION DU CAPOT REP 222 D / EPC 1 EMBASE DE SUPPORT TYRAP S'EST DESERTI3/ LORS DE LA PRESENTATION DES CAPOTS 222 G ET 223 D EPC, DES TETE DE TYRAP EMPECHENT UN MONTAGE CORRECT. », trois anomalies sont mentionnées. Seule la troisième concerne la classe *Présence d'un obstacle*. Nous avons fait ce choix afin de voir si le LLM est capable d'identifier les situations qui relèvent bien du *frame* en question ou non. Cela permet de définir si, dans le cas d'une chaîne de traitement plus globale, il est nécessaire d'effectuer une étape supplémentaire de segmentation du corpus.
2. Le second *dataset* est constitué des mêmes **25 rapports d'anomalies normalisés**. Nous avons extrait ces rapports des résultats de la normalisation présentée au chapitre

7. Ce *dataset* nous permet de comparer l'impact du domaine de spécialité uniquement, en soustrayant les phénomènes de bruit de l'équation. Comme nous avons pu le voir dans le chapitre en question, cette normalisation n'est cependant pas parfaite. Dans la Figure 9.7, nous pouvons observer un exemple de rapport normalisé. Les éléments à corriger sont indiqués en rouge dans le premier rapport. Dans le second exemple, les éléments en vert sont ceux corrigés au nombre de six, en plus de la rectification de la casse. Il reste deux éléments rouges qui sont l'absence d'accent dans « placee » et l'ajout d'une erreur avec la suppression de l'acronyme « DAS ».

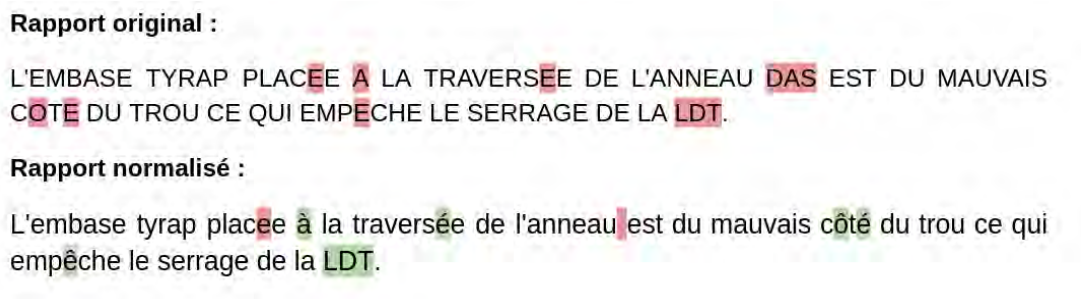


FIGURE 9.7 – Exemple de rapport normalisé des *datasets* 1 et 2

3. Pour ce troisième *dataset*, nous avons extrait **25 phrases du corpus frWaC** (Baroni *et al.*, 2009) qui regroupe une vaste collection de textes issus de sites web français (avec le domaine « .fr »). L'extraction s'est faite à partir de la recherche de marqueurs de cette catégorie (« empêche, bloque... ») au sein du corpus complet, puis l'extraction de 25 phrases complètes. Nous avons vérifié manuellement la correspondance entre ces phrases et les *frames Hindering* ou *Preventing or letting*. Cela nous a permis d'obtenir un *dataset* avec des données réelles et écologiques et de domaines variés tels que « Le magnésium réduit le stress en empêchant la montée du cortisol. » ou encore « Il doit rejoindre une planète peu connue et s'allier à des personnes qu'il ne connaît même pas pour empêcher (encore et toujours) la destruction de la Terre. »
4. Le quatrième *dataset* est composé de **25 phrases générées automatiquement par un LLM**. En effet, nous avons utilisé un LLM, ici Meta-Llama-3-8B (Grattafiori *et al.*, 2024) pour générer des phrases correspondant au cadre *Hindering*. Le prompt a été rédigé en anglais d'après les recommandations de Jin *et al.* (2024)[76]. Dans ce prompt (*cf. Figure 9.8*) nous avons inclus les éléments suivants :

76. Des travaux ultérieurs ont montré que l'utilisation de l'anglais ne produit pas systématiquement de meilleurs résultats lors du traitement de textes dans d'autres langues. Cela peut dépendre de la langue, de la tâche et du modèle.

— Le rôle que le modèle doit jouer et la tâche qu'il doit effectuer. Nous avons choisi d'utiliser le rôle de linguiste puisque la tâche en question, la génération de texte, doit s'appuyer sur le concept linguistique de *frame*, concept que nous ne définissons pas car nous avons pu voir que le modèle connaît l'existence de *FrameNet*. ;
— La définition du *frame Hindering* reprise de *FrameNet* afin de clarifier le type de phrases que nous désirons obtenir ;
— La définition et un exemple pour l'élément *Target*. Si cette information n'est pas nécessaire dans le simple objectif de générer des phrases correspondant au *frame*, elle est justifiée par le point suivant ;
— La demande d'avoir des phrases qui n'incluent pas nécessairement tous les éléments noyaux suivis d'un exemple avec uniquement l'élément *Target* présent. De manière générale, tous les éléments du *frame* ne se retrouvent pas tous nécessairement dans un énoncé. Dans les rapports d'anomalies, nous avons pu observer que cette tendance est fortement présente. Nous avons donc voulu demander spécifiquement dans le prompt des phrases n'incluant pas tous les éléments noyaux, et nous avons spécifié l'élément *Target* dans l'objectif d'obtenir des phrases avec une *Action* spécifiée sans *Target*.

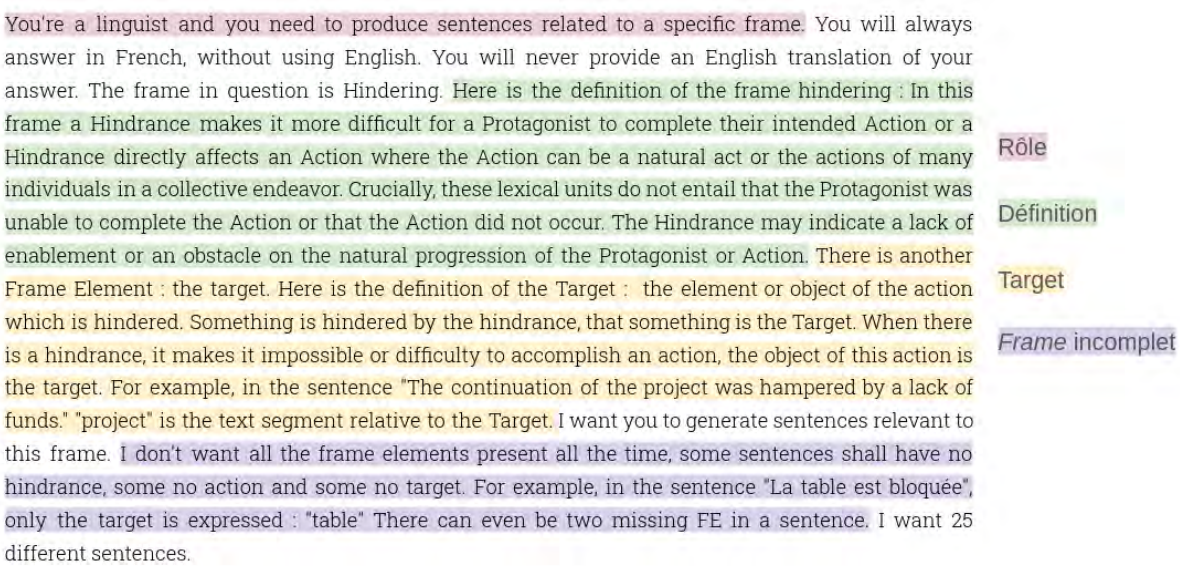


FIGURE 9.8 – prompt utilisé pour la génération de phrases avec un obstacle

5. Le dernier *dataset* est aussi constitué de 25 phrases générées automatiquement, mais cette fois, nous avons souhaité obtenir des **phrases synthétiques similaires aux rapports d'anomalies**. L'objectif était d'obtenir des phrases qui sont structurellement similaires aux rapports (avec des parenthèses, des guillemets...) mais en utilisant un

vocabulaire non spécialisé. Pour cela, nous avons inclus dans le prompt les éléments suivants (cf. Figure 9.9) :

— Le rôle similaire au *dataset* précédent ;
— Le type de structure désiré ainsi que l'utilisation d'un vocabulaire non spécialisé ;
— Le fait qu'il faut que la situation soit relative à un obstacle. Nous ne faisons pas mention de *frame* ici ;
— Deux exemples de vrais rapports d'anomalies.

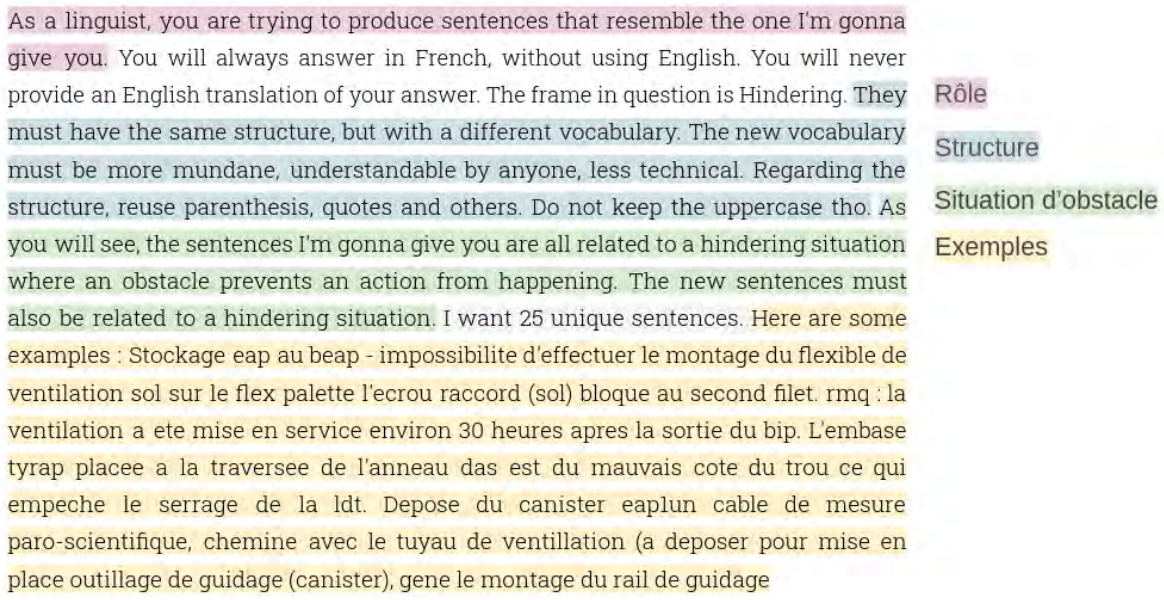


FIGURE 9.9 – prompt utilisé pour la génération de rapports avec un obstacle

Pour ces deux derniers *datasets* (c'est-à-dire les données synthétiques), nous avons des avis mitigés sur la qualité des phrases générées. En effet, nous observons que les structures et le vocabulaire utilisés sont répétitifs. Nous avons lancé le prompt (ou des variantes du prompt) à plusieurs reprises et retrouvons le même phénomène. Huang *et al.* (2025) ont observé des résultats similaires lors de l'utilisation de *few-shot prompting*, c'est-à-dire lorsque des exemples sont inclus dans le prompt. Dans la Figure 9.10, nous pouvons observer plusieurs exemples issus des données synthétiques. Nous avons surligné dans les mêmes couleurs les séquences qui réapparaissent dans plusieurs phrases. Comme nous pouvons le voir, plusieurs d'entre elles sont effectivement récurrentes. Par ailleurs, même lorsque le vocabulaire diffère, les structures et temps verbaux peuvent être similaires. De ce fait, la génération de données synthétiques montre très rapidement des limites. Pour le *dataset* 4, nous pouvons aussi constater que les domaines des exemples sont relativement abstraits comparés au type d'obstacle avec lesquels on traite dans les rapports (« démarrage du nouveau projet /

commerce », « expansion de l'entreprise «). Il n'est pas ici question d'obstacles physiques. Cela est probablement causé par l'exemple utilisé dans le prompt, issus de *FrameNet*, « *The project was hampered by a lack of funds* ». Dans le *dataset 5*, nous constatons que le respect de la consigne quant au fait de copier les rapports d'anomalie n'est pas suivie.

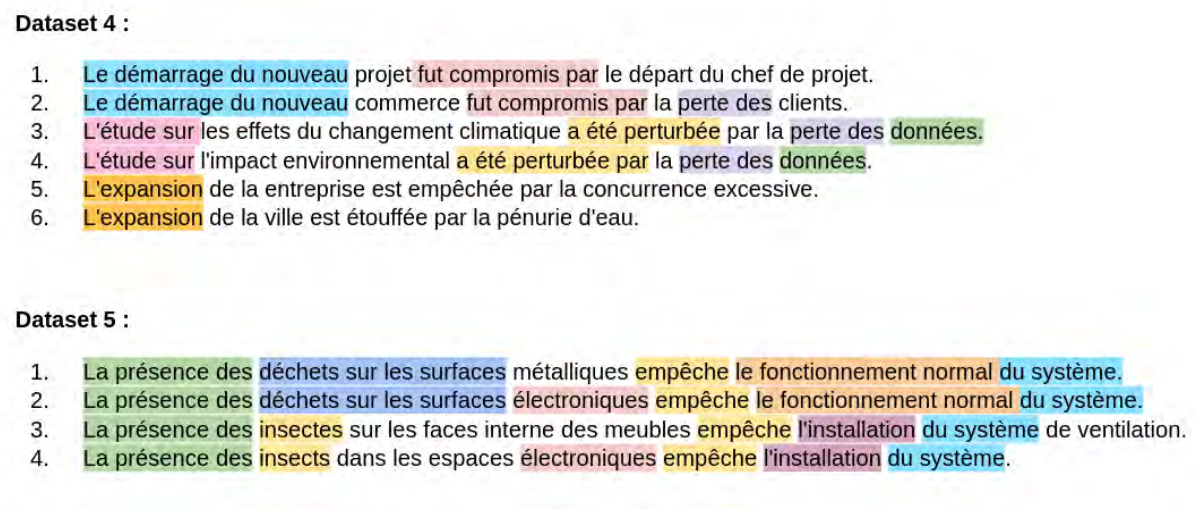


FIGURE 9.10 – Récurrences dans les données synthétiques

Ainsi, la génération automatique de phrases montre rapidement ses limites. Des tests de variation du prompt peuvent évidemment être envisagés afin d'obtenir des résultats plus satisfaisants. Concernant le *dataset 4*, la variation d'exemples permet, d'après quelques tests que nous avons effectués par la suite, d'obtenir des phrases traitant d'obstacles plus concrets. Pour le *dataset 5*, malgré plusieurs essais, nous n'avons pas réussi à obtenir des phrases plus similaires aux rapports. Une description plus détaillée de la structure et l'ajout d'exemples supplémentaires peuvent être envisagés.

9.2.2.2 Méthode d'annotation et d'évaluation

Pour annoter les différents *datasets*, nous avons élaboré un nouveau prompt présenté en Table 9.2. Celui-ci est aussi rédigé en anglais. Au sein de ce prompt, nous retrouvons les indications suivantes :

— L'attribution du rôle de linguiste de même que pour les prompts utilisés pour la génération de phrases ;
— La définition du *frame Hindering* issue de *FrameNet* ;
— La définition de la tâche : trouver les segments de textes relatifs aux trois éléments du *frame* ;
— Une définition pour chacun des éléments *Hindrance, Action* et *Target* ainsi qu'un exemple qui illustre ces trois éléments ;

— L'indication que les trois éléments ne sont pas systématiquement présents ainsi qu'un exemple ;
— L'indication d'une sortie au format *json* qui s'est révélée produire de meilleurs résultats (Devasier *et al.*, 2025) et permet un traitement *a posteriori* plus facile. La structure *json* à suivre est indiquée dans le script.

Description	Text
Rôle	You're a linguist and you are asked to annotate sentences according to the frame Hindering.
Définition du frame Hindering	Here is the definition of the frame hindering : In this frame a Hindrance makes it more difficult for a Protagonist to complete their intended Action or a Hindrance directly affects an Action where the Action can be a natural act or the actions of many individuals in a collective endeavor. Crucially, these lexical units do not entail that the Protagonist was unable to complete the Action or that the Action did not occur. The Hindrance may indicate a lack of enablement or an obstacle on the natural progression of the Protagonist or Action.
Tâche	I want you to find the text segment relative to three frame elements, Hindrance, Action and Target.
Définition des éléments du frame et exemple	Here is the definition of the frame element hindrance : The Hindrance is the force which makes the Action more difficult or less likely to be successfully completed. The Hindrance may be a lack of enabling force : The project was hampered by a lack of funds.'A lack of funds is the text segment relative to the hindrance in this exemple. Here is the definition of the frame element Action : The Action is what the Protagonist desires to do, which is made more difficult by the Hindrance. It can also be a natural act or collective endeavor, not linked to an overt Protagonist. Here is the definition of the frame element Target : the element or object of the action which is hindered. Something is hindered by the hindrance, that something is the Target. When there is a hindrance, It makes it impossible or difficuly to accomplish an action, the object of this action is the target. For example, in the sentence 'The continuation of the project was hampered by a lack of funds.' 'project' is the text segment relative to the Target and 'continuation' is the text segment related to the 'Action'.
Frames incomplets et exemples	There may be no text segment corresponding to the frame elements hindrance, action or target, even if the sentence is relative to the frame hindering. For example, the sentence Ṫhe door is blockedís a hindering case, but since we do not know what is blocking the opening of the door, there is no text segment to select for hindrance. If you encounter this kind of cases, write 'not present'.
Output	Return as json.

TABLE 9.2 – Prompt utilisé pour l'annotation

Ce prompt a été lancé cinq fois pour chacun des *datasets* compte tenu du non-déterminisme des modèles et nous avons gardé la température par défaut de l'outil utilisé (Ollama [77]), c'est-à-dire 0,8. Chaque phrase ou rapport a été traité indépendamment du reste de son *dataset*.

77. https://ollama.com/

Phrase	Des boîtes en carton empêchent l'ouverture de la porte verte	
	Target	Evaluation
Gold	porte	
Réponse 1	porte	Correct
Réponse 2	de la porte verte	Correct
Réponse 3	boîtes en carton	Incorrect

TABLE 9.3 – Évaluation de l'étiquetage automatique

Pour les cinq *datasets* (de 25 phrases ou rapports chacun), nous avons défini nous-même manuellement un *gold* pour l'évaluation de l'extraction en s'appuyant sur l'expertise que nous avons acquis sur nos données et sur la structure des énoncés. Pour cette évaluation, nous considérons que la réponse est correcte s'il existe un recouvrement au moins partiel entre le *gold* et le segment extrait, comme dans la Table 9.3. Pour le segment de texte qui fait office de *gold*, nous avons choisi de sélectionner une séquence de 1 à 2 mots non-vides qui correspondent à la tête du syntagme, plutôt que de sélectionner le segment de texte complet à la *Framenet*. Par exemple, pour la phrase « Des boîtes en carton empêchent l'ouverture de la porte verte », l'élément *Target* serait « porte verte ». Dans le *gold*, nous n'aurions sélectionné que « porte ». Ce choix entraîne donc une évaluation plus stricte puisque la tête du syntagme présente un segment plus réduit que le syntagme entier. Cependant, notre objectif n'est pas ici d'effectuer du *Frame Semantic Labeling* à proprement parler, mais plutôt de repérer les différents éléments du texte pertinents à la modélisation du problème et c'est la tête du syntagme qui est ici pertinente. Le *frame* est un guide dans cet objectif et non une grille stricte à laquelle l'extraction doit se conformer.

9.2.3 Résultats

Les résultats de l'extraction automatique sont présentés dans la Table 9.4. Nous avons fait la moyenne des 5 runs et présentons séparément les scores (précision, moyenne et F-mesure) pour chacun des éléments du *frame* modifié (*Hindrance, Action* et *Target*) ainsi que la moyenne sur les trois éléments. Pour calculer les résultats, nous évaluons la correspondance entre l'extraction et le gold au niveau du mot. Nous n'avons pas observé de cas avec des réponses vides. Il était indiqué dans le prompt de mentionner « not present » dans le cas où un *frame element* n'était pas trouvé dans l'énoncé. De même, dans le *gold*, nous avions indiqué « not present » lorsque l'élément n'est pas présent dans le texte. Nous avons cependant observé que le modèle a tendance à toujours vouloir donner une réponse issue des données et n'avons trouvé aucun cas d'utilisation de « not present » en réponse.

Les résultats montrent des scores de F-mesure satisfaisants pour tous les *datasets*, supérieurs à 0,6 (sauf pour l'élément *Hindrance* du *dataset FrWaC*). Cela montre un accord assez fort entre les segments extraits et les segments *gold*. Les *datasets* synthétiques (4 et 5) obtiennent des scores supérieurs à 0,9 (à l'exception du rappel de *Target* pour les phrases synthétiques). Les autres *datasets*, contenant des données réelles, montrent des scores plus faibles mais qui restent acceptables, notamment si nous prenons en compte la difficulté de la tâche et les particularités des données, avec entre 0,56 et 0,83 de F-mesure en fonction du *frame element*. Nous ne notons pas de tendance supplémentaire dans les scores entre les différents *datasets*. L'élément *Hindrance* semble un peu plus difficile à identifier. Cela peut être dû au fait que les éléments *Action* et *Target* sont habituellement suivis l'un par l'autre, facilitant ainsi l'identification de ces séquences.

Cependant, l'observation manuelle montre que les segments extraits par le modèle sont des séquences longues. Lors de l'élaboration du *gold*, nous avons choisi de sélectionner la tête du syntagme. Nous avons aussi pu voir que l'annotation de *FrameNet* peut comporter des segments annotés longs, qui impliquent des circonstants et des arguments, notamment pour l'élément *Action*. Dans ces cas, la présence de la tête dans le segment extrait permet de considérer comme que c'est un succès, mais cela s'éloigne du *gold* que nous avons conçu. En revanche, la séparation de cet élément en deux, avec l'ajout de *Target*, diminue significativement la taille des séquences annotées. De ce fait, un écart de taille des séquences était attendu, mais avec une certaine marge d'erreur. Si nous prenons l'exemple en Table 9.5, nous pouvons voir que dans le cas de l'élément *Hindrance*, la différence entre le *gold* et le segment extrait est présente mais raisonnable. En revanche, pour l'élément *Action*, le segment extrait est beaucoup trop large et inclut à la fois l'*Action*, la *Target* et des informations supplémentaires en lien avec le but de l'action. Pour *Target*, le segment extrait est tout simplement faux.

Cette observation nous a amenés à vouloir quantifier la différence dans le recouvrement du *gold* et de l'extraction automatique. Pour ce faire, nous avons calculé le coefficient de Dice qui permet de déterminer la similarité entre deux ensembles. Ce calcul se fait sur la base de caractères communs (ou non) entre le *gold* et la réponse du modèle. Si la réponse est la même, le score est de 1. Si le *gold* et la réponse du modèle n'ont pas de caractères communs, le résultat est de 0. C'est la taille relative des parties partagées et non partagées qui fait varier le score.

Les résultats sont présentés en Table 9.6 et, pour cette mesure, plus le score est élevé, plus le recouvrement est fort. Dans cette table, nous présentons, pour chaque *dataset* et pour chaque étiquette, les moyennes des scores d'extractions obtenus pour chaque segment, pour chacune des phrases/rapports pour tous les annotateurs et les 5 runs confondus. Ces moyennes ne prennent en compte que les cas où la réponse a été considérée comme correcte lors de l'évaluation de l'extraction (rappel, précision et F-mesure) afin de ne pas baisser

#	Datasets	Rappel	Précision	F-score
		Hindrance		
1	Rapports	0.57	0.87	0.69
2	Rapports normalisés	0.57	0.86	0.68
3	FrWac	0.43	0.83	0.56
4	Phrases générées par LLM	0.98	0.92	0.95
5	Rapports générés par LLM	0.98	1	0.99
		Action		
1	Rapports	0.84	0.81	0.83
2	Rapports normalisés	0.69	0.78	0.73
3	FrWac	0.68	0.59	0.63
4	Phrases générées par LLM	0.96	0.92	0.94
5	Rapports générés par LLM	0.98	1	0.99
		Target		
1	Rapports	0.56	0.93	0.70
2	Rapports normalisés	0.45	0.91	0.60
3	FrWac	0.73	0.95	0.82
4	Phrases générées par LLM	0.82	1	0.90
5	Rapports générés par LLM	0.98	1	0.99
		Moyenne des 3 frame elements		
1	Rapports	0.66	0.87	0.74
2	Rapports normalisés	0.61	0.85	0.67
3	FrWac	0.61	0.79	0.60
4	Phrases générées par LLM	0.92	0.95	0.93
5	Rapports générés par LLM	0.98	1	0.97

TABLE 9.4 – Résultats de l'extraction automatique

artificiellement les résultats et de mesurer au-delà des résultats déjà obtenus.

Pour les *datasets* générés automatiquement, les scores avoisinent 0,7. Pour les données réelles, les scores chutent autour de 0,5 pour les rapports et 0,6 pour les phrases issues de *FrWaC*. Cela nous indique que, conformément à ce que l'observation manuelle des résultats indique, les segments de textes extraits pour les *datasets* 1 à 3 sont beaucoup plus longs que le *gold*. Le modèle n'est donc pas en mesure de délimiter correctement les différents empans du texte, et donc de fournir des résultats exploitables en vue d'une modélisation ou autres applications.

Rapport	LA PRESENCE DU DOME RLH2 EMPECHE DE PASSER LA PERCEUSE POUR CHASSER LES RIVETS POUR AMENAGER LES LOGEMENTS DES BUSES DE PASSAGE DE CLOISON CONE CASE.		
	Hindrance	**Action**	**Target**
Gold standard	DOME	PASSER	PERCEUSE
LLM labeling	la présence du dome RLH2	passer la perceuse pour chasser les rivets pour aménager les logements des bus de passage de cloison cone case	les bus de passage de cloison cone case

TABLE 9.5 – Exemple de sortie de l'extraction automatique

Ces résultats peuvent probablement être améliorés, notamment avec l'ajout d'exemples plus précis dans le prompt, conformes aux types de segments identifiés dans le *gold*. Cependant, nous avons vu que les travaux de Devasier *et al.* (2025) montrent aussi des limites dans l'utilisation de LLM pour l'extraction d'éléments d'un *frame* dans des corpus de domaines différents.

L'utilisation de rapports normalisés ne montre pas d'amélioration au niveau des scores de recouvrement (pas plus que pour l'identification). Ainsi, le bruit ne semble pas à l'origine des difficultés dans l'identification des segments, ce que montrent aussi les résultats sur le *dataset* issu de *FrWac* qui sont du même ordre. Notre hypothèse est que la présence de séquences de textes additionnels dans les données réelles, dépassant la couverture du *frame*, semble perturber les performances du LLM. Quand il est confronté à des données réelles, qui s'éloignent des structures simples et canoniques, le modèle éprouve des difficultés à délimiter et à identifier les segments. Le modèle montre une tendance à vouloir sélectionner des portions étendues du texte, même si celles-ci ne sont pas pertinentes. Par ailleurs, le modèle a aussi donné des réponses non conformes qui altèrent les résultats, comme des réponses ponctuelles en anglais, la modification dans les réponses de certains mots inconnus (l'acronyme « FE » devient « feu »), ou des modifications de la forme du mot pour une forme fléchie (un verbe conjugué passe à l'infinitif). Puisque notre évaluation se fait sur la correspondance de chaînes de caractères au niveau du mot, et non sur une extraction à partir de position des *offsets*, ces résultats sont donc considérés comme faux dans l'évaluation de l'annotation.

#	Datasets	Recouvrement
		Hindrance
1	Rapports	0,54
2	Rapports normalisés	0,47
3	FrWac	0,69
4	Phrases générées par LLM	0,72
5	Rapport générés par LLM	0,47
		Action
1	Rapports	0,41
2	Rapports normalisés	0,40
3	FrWac	0,47
4	Phrases générées par LLM	0,59
5	Rapport générés par LLM	0,73
		Target
1	Rapports	0,52
2	Rapports normalisés	0,54
3	FrWac	0,65
4	Phrases générées par LLM	0,85
5	Rapport générés par LLM	0,77
		Moyenne
1	Rapports	0,49
2	Rapports normalisés	0,47
3	FrWac	0,60
4	Phrases générées par LLM	0,72
5	Rapport générés par LLM	0,67

TABLE 9.6 – Résultats du calcul du coefficient de Dice

9.3 Vers une modélisation ?

La modélisation des cas de la catégorie *Présence d'un obstacle* pose un certain nombre de difficultés. Pour la catégorie *Fuite d'un liquide ou d'un gaz* vue dans le chapitre 8, nous avions pu établir un *vépole* type dont les différents éléments correspondaient à certains participants du *frame*. Pour la classe *Présence d'un obstacle*, nous verrons qu'un *vépole* type ne peut pas être directement établi à partir du texte, et que cette correspondance entre les éléments du *frame* de l'obstacle et les éléments du *vépole* n'est pas directe.

Afin d'illustrer ces difficultés, nous proposons de partir de deux exemples de phrases simples représentant la situation d'obstacle :

1. Des tapis bloquent l'accès à la chambre.
2. Des poussières empêchent d'éclairer la pièce correctement.

Nous nous concentrons sur des obstacles physiques plus susceptibles d'être traités par TRIZ. Pour chacune de ces phrases, nous pouvons repérer les différents éléments du *frame* tel que nous l'avons conçu jusqu'ici, c'est-à-dire *Hindrance* (ce qui bloque), *Action* (l'action empêchée) et *Target* (le patient de l'action). Ces annotations sont présentées en Table 9.7.

Phrase	**Hindrance**	**Action**	**Target**
Des tapis bloquent l'accès à la chambre.	tapis	accès	chambre
Des poussières empêchent d'éclairer la pièce correctement.	poussières	éclairer	pièce

TABLE 9.7 – Annotation des phrases d'exemple

Si nous reprenons le schéma d'un *vépole* minimal, nous savons qu'il est composé de trois éléments : l'outil, le champ et le produit. L'outil émet un champ (une action) sur le produit qui subit ainsi un changement d'état.

Cette formulation montre des similitudes avec les éléments du *frame* qui pourraient nous entraîner à vouloir effectuer une équivalence directe entre les éléments du *vépole* et les éléments du *frame*. Cependant, il existe une inadéquation majeure entre les éléments des deux modèles. L'élément outil dans le *vépole* est censé produire une action sur l'élément produit. Or, dans ces exemples, l'obstacle (les tapis ou les poussières) n'est pas directement vecteur de ces actions. Ce ne sont pas les poussières qui éclairent, ce n'est pas le tapis qui accède à la chambre. Les éléments *Target*, la chambre et la pièce, ne subissent pas non plus de changement d'état.

Dans les Figures 9.11 et 9.12, nous retrouvons les modélisations en *vépole* pour les deux problèmes préalablement énoncés. Dans les deux cas, nous pouvons voir que certains éléments nécessaires aux *vépoles* sont absents de l'énoncé.

Pour le premier exemple, nous inférons sur le problème décrit en supposant l'existence d'une porte donc la manœuvre est bloquée. Celle-ci vient, dans l'énoncé, spécifier l'objet directement impacté par l'obstacle (les tapis). Une version de cet énoncé avec toutes les informations présentes prendrait la forme suivante : « Des tapis **devant la porte** bloquent l'accès à la chambre. » Il manque donc des précisions quant au véritable objet impacté par l'obstacle, qui peuvent être apportées en précisant dans cet exemple la position de l'obstacle physique.

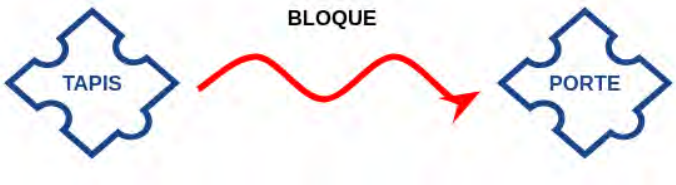


FIGURE 9.11 – Modélisation en *vépole* de l'exemple 1

Pour le second exemple, nous effectuons aussi des inférences quant à la situation afin d'ajouter dans l'énoncé l'information sur l'objet directement impacté par l'obstacle. Nous inférons que les poussières sont placées sur l'ampoule, entravant ainsi l'éclairage de la pièce. Nous partons de l'énoncé suivant : « Des poussières sur l'ampoule empêchent d'éclairer la pièce correctement ». La modélisation qui en découle, illustrée dans la Figure 9.12, est plus complexe que pour l'exemple de la porte bloquée. Elle implique ici trois *vépoles* dont deux *vépoles* néfastes (indiqués par la flèche ondulée rouge). C'est la poussière ici qui subit un changement d'état néfaste en étant entravée dans son déplacement « naturel ». Elle est agrégée, collée, à l'ampoule. Cette dernière est génératrice de l'action en empêchant la poussière de se déplacer. Cette même poussière est aussi outil d'un second *vépole* qui découle du premier et bloque ainsi la lumière, produite par l'ampoule. L'élément « lumière » ne figure cependant nulle part dans l'énoncé, que ce soit dans l'énoncé initial ou dans le second issu de l'inférence. Il est sous-entendu par le verbe « éclairer ». Ainsi, l'apport d'éléments supplémentaires quant à l'objet directement impacté par l'obstacle, c'est-à-dire l'ampoule, n'est pas suffisant comme il a pu l'être dans l'exemple précédent. Il est ici nécessaire d'avoir des connaissances du fonctionnement du système qui ne sont pas décrites dans le texte car présupposées connues.

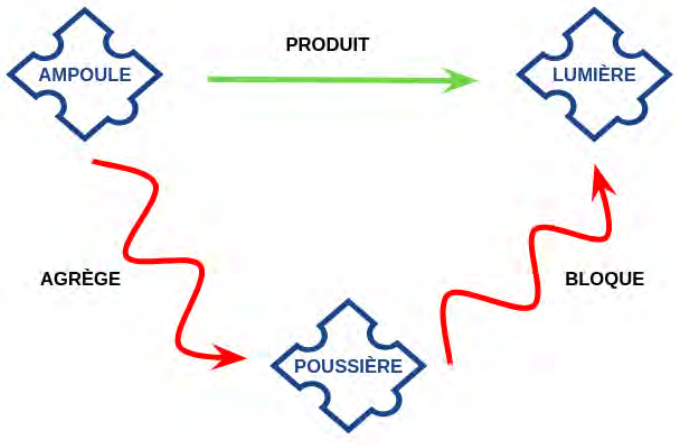


FIGURE 9.12 – Modélisation en *vépole* de l'exemple 2

Ainsi, dans les deux cas, les énoncés impliquant les trois éléments noyaux du *frame* sont insuffisants pour la modélisation en *vépole*. Dans le premier cas, une précision quant à l'objet directement impacté par l'obstacle est nécessaire pour combler les éléments manquants. Dans le second, la situation est plus délicate puisque les éléments manquants à la formalisation ne peuvent pas être comblés aisément par des précisions apportées par l'énoncé. Elles impliquent des éléments qui ne sont pas exprimés, car on présuppose d'une certaine connaissance du monde et du système qui dysfonctionne. Nous pouvons cependant noter que l'élément « Obstacle » (*Hindrance*) du frame (« tapis » et « poussières » dans les exemples) figure toujours en tant qu'élément outil du *vépole* qui implique le champ *bloque*. Nous pouvons donc, à partir des éléments extraits, reconstituer une partie d'au moins un des *vépoles* qui intervient dans la situation de dysfonctionnement qui est :

OBSTACLE (HINDRANCE) - BLOQUE - ?

Le reste du/des *vépoles* doit cependant être cherché parfois au-delà de l'énoncé et du *frame* tel qu'il est décrit.

Pour en revenir à nos rapports d'anomalies, certaines précisions quant à l'obstacle et sa disposition sont parfois mentionnés dans certains rapports (non majoritaires). Par exemple, dans le rapport « L'EMBASE TYRAP PLACEE A LA TRAVERSEE DE L'ANNEAU DAS EST DU MAUVAIS COTE DU TROU CE QUI EMPECHE LE SERRAGE DE LA LDT. », la disposition de l'obstacle est donnée avec précision avec la séquence « PLACEE A LA TRAVERSEE DE L'ANNEAU DAS EST DU MAUVAIS COTE DU TROU ». De même pour le rapport « UN CABLE DE MESURE PARO-SCIENTIFIQUE, CHEMINE AVEC LE TUYAU DE VENTILLATION (A DEPOSER POUR MISE EN PLACE OUTILLAGE DE GUIDAGE (CANISTER), GENE LE MONTAGE DU RAIL DE GUIDAGE », pour lequel la disposition est « CHEMINE AVEC LE TUYAU DE VENTILLATION ». Cependant, nous ne retrouvons pas ici de quelle façon le blocage par l'*Obstacle* (« L'EMBASE TYRAP » et « CABLE DE MESURE PARO-SCIENTIFIQUE ») a lieu. Cela n'est pas clairement exprimé dans le texte. En effet, dans l'énoncé, nous ne retrouvons pas l'information exprimant en quoi le fait que le « CABLE DE MESURE PARO-SCIENTIFIQUE, CHEMINE AVEC LE TUYAU DE VENTILLATION » est un problème, ni leur relation avec le « TUYAU DE VENTILLATION ». Peut-être qu'une connaissance poussée de ces systèmes, de leur agencement et disposition permettrait de déduire les relations existantes entre ces différents phénomènes. Ces textes étant écrits par des experts à l'intention d'autres experts, des connaissances du système sont présupposées. Ainsi, le texte seul n'exprime pas les différents éléments nécessaires au *vépole*.

Cela signifie que, même en présence des différents éléments du *frame*, une modélisation en *vépole* paraît difficile à produire et encore plus à automatiser. Les énoncés peuvent

contenir des informations pertinentes, mais la façon dont sont exprimées les relations entre les différents composants physiques intervenant dans la situation ne permet pas d'avoir une représentation claire du problème à partir du texte. Dans certains cas, des informations supplémentaires peuvent être récupérées dans les rapports (comme la position de l'obstacle pour « devant la porte ») et peuvent venir compléter le *vépole*. Cependant, dans d'autres cas, des éléments implicites manquent et ne peuvent être récupérés dans le texte. Les acteurs principaux de la situation ne sont pas tous mentionnés. Nous n'arrivons donc qu'à une modélisation partielle de la situation.

En conclusion de cette étude sur la catégorie *Présence d'un obstacle*, nous pouvons observer plusieurs limites également. En effet, la correspondance entre la catégorie et deux *frames* peut nous pousser à nous interroger sur la typologie. Celle-ci a été construite à partir d'une approche de linguistique de corpus outillée, et non à partir des *frames*. De ce fait, un manque d'alignement entre les deux n'est pas surprenante. Cette distinction effectuée par les deux *frames* se retrouve potentiellement dans le *vépole*. Nous pouvons envisager que la différence entre *Hindering*, qui fait référence à une gêne, et *Preventing or letting*, qui renvoie à un obstacle total, peut se retrouver à travers les types de relation du *vépole* « insuffisant » et « néfaste ». Cette hypothèse serait à confirmer par une étude plus poussée en ce sens, avec une mise en correspondance systématique d'énoncés appartenant à ces deux *frames* et d'une représentation en *vépole*. Ces différents arguments sont pourtant, selon notre opinion, détachés de la manière dont nous avons construit la typologie mais aussi de sa fonction. Celle-ci vise à représenter et à regrouper les différentes expressions des problèmes techniques dans le texte. Elle permet de faire le lien avec les *frames* mais n'a jamais été pensée comme un calque de la représentation sémantique à la *FrameNet* sur nos données, ni à classer les types de problèmes du corpus par rapport à des classes de *vépoles*.

Nous avons aussi pu voir que la modélisation n'est pas aisée. L'obstacle ne peut être formalisé aisément à l'aide d'un *vépole* type comme nous avons pu le faire pour la catégorie *Fuite d'un liquide ou d'un gaz*. Il nécessite la présence d'informations non exprimées dans le texte avec un manque de description de la nature du blocage qui a lieu.

La tâche d'extraction automatique a aussi montré des limites avec des difficultés, pour le LLM à repérer correctement les différents éléments du texte. L'application du prompt sur des données réelles a montré que les phrases non canoniques, et qui ne sont pas uniquement représentatives du *frame* en question, sont mal gérées par le modèle. Il se perd dans le contenu textuel et échoue à identifier correctement les différentes séquences.

Néanmoins, nous estimons avec cette étude avoir ouvert la porte vers des travaux intéressants concernant l'extraction automatique de type *Frame Semantic Role Labeling* de rapports d'anomalies avec un LLM, mais aussi sur l'adéquation entre TRIZ et les rapports d'anomalies. Nous avons pu établir comment repérer l'obstacle et les éléments centraux qui le constitue au sein des rapports d'anomalies qui était l'objectif fixé. Cependant, pour cette catégorie, le chemin vers le *vépole* laisse apercevoir d'autres éléments, au-delà du texte, à prendre en considération.

Par rapport à l'automatisation, nous avons pu identifier certains écueils. Une première piste viserait à l'amélioration du prompt afin de préciser le type d'attendu pour l'extraction. L'ajout d'exemples plus proches des données réelles, plutôt que l'utilisation d'exemples issus de *FrameNet* pourrait aussi permettre de meilleurs résultats. Une seconde piste serait d'envisager une segmentation des rapports afin de ne donner au modèle que la séquence de texte en rapport avec la situation d'obstacle, plutôt que le rapport complet qui peut inclure soit des phrases en rapport avec une autre classe de la typologie, soit des informations contextuelles sur l'anomalie.

Conclusion

Une des questions abordées qui sous-tend cette thèse est de savoir comment aborder le traitement de données spécialisées et bruitées. Pour formaliser les éléments sous forme de *vépole*, il nous faut d'abord repérer les éléments pertinents. Or, la complexité dans la compréhension des données en question pour des non-experts, et le bruit qu'elles comportent, a recentré la question afin de trouver un moyen de les traiter et d'en saisir le contenu, avec un accès limité à un expert et une absence de ressources externes dédiées.

Pour ce faire, nous avons donc exploré diverses approches de linguistique outillée et de TAL. Parmi celles-ci, nous retrouvons des techniques telles que le *Topic Modeling*, *Word2Vec* et l'analyse des fréquences lexicales afin d'explorer le corpus à gros grain autour de la notion de problème. C'est par l'analyse des résultats qui en sont ressortis que nous avons pu construire la typologie d'expression d'un problème technique fondée sur les marqueurs lexicaux qui font référence à la notion de problème. De ce fait, nous pensons que cette typologie peut être applicable à d'autres domaines techniques. Nous avons eu déjà eu l'occasion d'explorer cette possibilité par le biais d'un mémoire de master 1 effectué sous notre co-direction lors de l'année université 2024-2025 par Estelle Laurence, intitulé *Robustesse des projections syntaxiques face au bruit textuel : une étude expérimentale sur le corpus CoCoRep*. Dans celui-ci, nous retrouvons la classe *Action difficile ou impossible* et ses patrons, basés sur des marqueurs. La mise en application s'est faite sur un corpus de posts du forum *commentreparer.fr* dans lequel des utilisateurs requièrent de l'aide pour leurs appareils électroménagers défectueux. Les *patterns* retrouvés impliquent les mêmes marqueurs que ceux que nous avons pu identifier pour la catégorie dans notre corpus.

Par la suite, la mise en œuvre de cette typologie a notamment consisté en l'annotation d'un *dataset* par des non-experts. Celle-ci s'est focalisée sur la sélection et l'étiquetage des marqueurs selon la typologie. Nous avons pu voir dans quelle mesure des annotateurs non experts peuvent se frotter à des textes spécialisés et peu accessibles lorsque les annotations portent avant tout sur l'expression du problème, plutôt que son interprétation. Nous avons cependant pu remarquer la présence de vocabulaire général scientifique au sein des marqueurs (comme « perte », « invalide »...) *a posteriori* de l'annotation du *dataset*, par l'annotation

d'un échantillon par un expert. Une implication de l'expert en amont aurait pu permettre d'expliciter ces occurrences et donc une catégorisation des marqueurs plus fine. Cependant, le *fine-tuning* d'un modèle neuronal a montré des résultats satisfaisants étant donné la difficulté de la tâche.

Avec cette thèse, nous proposons aussi une méthode de normalisation des données avec l'utilisation d'un LLM génératif. Cette méthode permet de corriger le corpus des phénomènes de bruit sans pour autant impacter le lexique spécialisé. Cependant, si nous avons pu obtenir des scores satisfaisants avec le CER, une analyse qualitative montre tout de même des failles avec parfois des corrections non effectuées mais aussi la suppression d'éléments du texte, et ce, même avec les modèles les plus performants. Ainsi, l'utilisation des corpus qui en ressortent est à nuancer en fonction de l'utilisation qui est envisagée.

Par ailleurs, la tâche d'étiquetage des marqueurs a pu être reproduite sur le *dataset* normalisé, mais n'a pas montré d'augmentation dans les performances des modèles. Cela nous amène à questionner l'emploi de techniques avancées et coûteuses (en temps de traitement et du point de vue écologique) par rapport à la tâche en question. Cela démontre aussi la robustesse des modèles neuronaux de type *transformers* face au bruit.

Pour la normalisation, des perspectives peuvent aussi être ouvertes avec le *fine-tuning* des LLM. En les familiarisant avec un échantillon du corpus et avec une version nettoyée (plutôt que le *few-shot prompting* que nous avons mis en place), une amélioration des résultats peut être envisagée.

L'étude de cas de la catégorie *Fuite d'un liquide ou d'un gaz* nous a aussi permis d'explorer plusieurs aspects de ces données. Tout d'abord, nous avons pu voir en quoi les *frames* proposés par *FrameNet* montrent des limites face à des textes de domaines spécialisés. La situation typique qu'ils cherchent à décrire s'éloigne de celles que nous retrouvons dans les données en raisons des objectifs de communication qui diffèrent. Les énoncés de cette catégorie ont aussi montré une variété et une profusion dans les éléments énoncés qui sont cependant trop opaques pour être identifiés et qualifiés directement par un non-expert. De ce fait, nous avons choisi cette fois encore de nous concentrer sur la façon dont est exprimée la fuite pour en reconnaître le mode d'expression. Nous avons à nouveau fait usage d'outils et de méthodes de linguistique de corpus, à savoir de la fouille de motifs par de l'effacement de texte de façon itérative et de l'analyse de structures de dépendances syntaxiques. Cela nous a permis de faire émerger les patrons récurrents autour de l'expression de la fuite dans un environnement technique. Un entretien avec un expert est venu confirmer les éléments récupérés et expliciter les éléments qui restaient obscurs. Ces différentes approches nous ont permis de constituer le *frame de la fuite dans un environnement technique*.

Lors de l'étude de la catégorie *Présence d'un obstacle*, nous avons aussi pu percevoir les limites du *frame* de *FrameNet* mais cette fois en raison de l'adéquation entre le *frame* et la conception du problème telle que nous l'avons envisagée lors de la constitution de la typologie. Nous mettons ici le doigt sur un point particulier de la thèse qui est l'utilisation des *frames* comme point d'appui et non comme pilier de notre démarche, ni comme objectif du travail. Nous n'avons pas cherché à créer des *frames* propres au domaine, ni à étudier la représentation des *frames* dans le corpus. Nous avons plutôt décidé de nous appuyer sur cette théorie afin de guider notre analyse fine des éléments de l'énoncé. De ce fait, des écarts entre la typologie et la catégorisation en *frames* ne sont pas étonnants. Cependant, dans le but d'opérationaliser le traitement des énoncés, nous avons tout de même utilisé un des *frames* proposés par *FrameNet* (avec quelques modifications). Cela nous a permis d'effectuer une annotation de type *Frame Semantic Role Labeling* en utilisant un LLM. Cette méthode possède un avantage dans l'étude des textes de spécialité puisqu'elle permet de se passer d'expert, dont la disponibilité est souvent limitée, pour l'annotation de textes techniques. Elle a cependant montré ses limites puisque le modèle échoue dans l'identification et la délimitation des segments. Nous pensons toutefois que cette piste reste à approfondir. Un *fine-tuning* du modèle pourrait être envisagé pour spécifier le modèle sur les données et la tâche en question et permettrait peut-être une meilleure compréhension des textes non canoniques. Le type de réponse recherchée pourrait aussi être fourni au modèle. Enfin, une segmentation préalable du corpus peut être envisagée pour cibler les éléments à extraire.

Au travers de cette thèse, il était aussi question de voir dans quelle mesure les REX peuvent être sujets à un traitement automatique vers TRIZ, et en l'occurrence la constitution d'un *vépole*. Lors du choix des classes pour nos deux cas d'étude, nous avons pu constater que certaines catégories étaient peu propices à la modélisation en *vépole*. Certaines fournissent de trop peu d'informations, comme *Action difficile ou impossible* ou *Dispositif qui ne fonctionne pas*, et d'autres sont trop éloignées de l'objet du problème, tel que *Signal qui s'est déclenché* ou encore *Configuration hors spécification*. Nous pouvons donc déjà voir des limites à la formalisation des REX pour un certain nombre de ces rapports.

Les deux cas d'étude sur lesquels nous nous sommes concentrés ont amené des conclusions qui diffèrent quant à la modélisation. Dans le cas de la catégorie *Fuite d'un liquide ou d'un gaz*, nous avons pu mener l'étude jusqu'à un *vépole* type représentant le problème en question. Celui-ci est quasi-complet, à l'exception de la différenciation entre liquide et gaz qui pourrait être effectuée grâce à des listes fermées d'équivalences. Par ailleurs, nos travaux se sont concentrés sur l'analyse de l'expression de la fuite et non sur l'automatisation d'un repérage de ces éléments. Une poursuite serait à mener pour étudier cette question avec,

par exemple, l'entraînement d'un modèle d'étiquetage de séquence comme nous avons pu le faire avec les marqueurs. Cela nécessiterait toutefois une campagne d'annotation manuelle de la part d'experts. Une seconde approche pourrait consister à reproduire l'annotation automatique par un LLM que nous avons effectuée pour la catégorie *Présence d'un Obstacle*, mais en prenant en compte les considérations évoquées précédemment. Dans le cas de la fuite, le *frame* que nous avons dégagé couvre une grande partie des éléments de l'énoncé, sauf peut-être certains circonstants. Nous pouvons donc supposer que cela limiterait la sélection de segments trop importants et aiderait dans la délimitation des segments du texte à catégoriser. Pour la catégorie *Présence d'un obstacle*, nous avons pu voir que malgré la quantité d'informations retrouvées dans l'énoncé, celles-ci ne permettent pas une modélisation du problème. Nous ne retrouvons dans le texte une description de l'anomalie qui est parcellaire dans la représentation de la situation évoquée, ce qui empêche sa formalisation en *vépole*.

Avec cette thèse, nous avons donc exploré un champ différent des travaux qui mêlent le TAL et TRIZ puisque nous n'en avons pas rencontré, à notre connaissance qui se concentre sur les REX. En effet, de même que les brevets sont à l'origine de la méthode TRIZ, de nombreux travaux se sont concentrés sur ce type de documents. Dans les REX, il est fait constat d'une anomalie qui est souvent un dysfonctionnement technique. Ainsi, l'idée de vouloir modéliser ce dysfonctionnement en *vépole*, formalisme qui permet à représenter un problème technique, semble plausible. Cependant, nous avons pu voir plusieurs limites à cette formalisation qui laissent à penser qu'un traitement direct du texte, tel qu'il se trouve dans les FA du CNES, vers un *vépole*, n'est que partiellement envisageable. Cela pose aussi la question du choix des *vépoles* comme point d'accès à TRIZ, et notamment en fonction des catégories identifiées. En revanche, si la modélisation en *vépole* est un objectif défini des rapports, alors le travail que nous avons effectué peut être un point de départ vers des consignes d'aide à la rédaction. En effet, l'étude de ces énoncés et la tentative de modélisation nous ont permis de repérer les informations manquantes et les écueils dans la formulation des descriptions (nous parlons ici d'écueils du point de vue de l'objectif de modélisation uniquement). Ils ouvrent aussi une perspective quant à l'automatisation de l'utilisation de la méthode TRIZ à partir de REX. De plus, la question qui peut se poser aujourd'hui est celle de l'utilisation de LLM génératifs pour la générations de *vépoles* directement à partir du texte. Nous avons pu mener quelques expérimentations et l'aspect des résultats produits ressemble bien à la modélisation visée. Cependant, il serait nécessaire de disposer une évaluation experte afin de déterminer si les *vépoles* produits sont corrects.

Bibliographie

ABEILLÉ, A., CLÉMENT, L. et TOUSSENEL, F. (2003). *Building a Treebank for French*, pages 165–187. Springer Netherlands, Dordrecht.

ABRAMOVICI, M. (1999). *La prise en compte de l'organisation dans l'analyse des risques industriels : méthodes et pratiques*. Thèse de doctorat, Cachan, Ecole normale supérieure.

AL SHAROU, K., LI, Z. et SPECIA, L. (2021). Towards a better understanding of noise in natural language processing. *In* MITKOV, R. et ANGELOVA, G., éditeurs : *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62, Held Online. INCOMA Ltd.

ALIZADEH, M., KUBLI, M., SAMEI, Z., DEHGHANI, S., ZAHEDIVAFA, M., BERMEO, J., KOROBEYNIKOVA, M. et GILARDI, F. (2024). Open-source llms for text annotation : a practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8.

ALMANSBA, H. et BRAHAM CHAOUCHE, S. (2013). *Contribution à l'amélioration de la performance d'une Supply-Chain en utilisant TRIZ et SCOR - Application : Schlumberger Algérie*. Thèse de doctorat, École Nationale Polytechnique (Alger).

ALTSHULLER, G. et SEREDINSKI, A. (2004). *40 Principes d'innovation : TRIZ Pour Toutes Applications*. Seredinski (Avraam).

ALTSHULLER, G., SHULYAK, L. et RODMAN, S. (1999). *The Innovation Algorithm : TRIZ, Systematic Innovation and Technical Creativity*. Technical Innovation Center.

AMALBERTI, R. et BARRIQUAULT, C. (1999). Fondements et limites du retour d'expérience. *In Annales des ponts et chaussées*, volume 91, pages 67–75.

ANDREANI, V., FABRE, C., RAYNAL, C. et TANGUY, L. (2013). Techniques de tal au service de la constitution d'une base de rex et de son analyse. Rapport technique, Institut pour la Maîtrise des Risques.

ANTOUN, W., SAGOT, B. et SEDDAH, D. (2023). Data-Efficient French Language Modeling with CamemBERTa. *In 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 5174–5185, Toronto, Canada. In Findings of the Association for

Computational Linguistics : ACL 2023, pages 5174–5185, Toronto, Canada. Association for Computational Linguistics.

BAKER, C. F. (2009). La sémantique des cadres et le projet FRAMENET : une approche différente de la notion de « valence ». *Langages*, 176(4):32–49.

BAKER, C. F., FILLMORE, C. J. et LOWE, J. B. (1998). The berkeley framenet project. *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1 (ACL '98)*, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

BANARESCU, L., BONIAL, C., CAI, S., GEORGESCU, M., GRIFFITT, K., HERMJAKOB, U., KNIGHT, K., KOEHN, P., PALMER, M. et SCHNEIDER, N. (2013). Abstract meaning representation for sembanking. *In Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

BARONI, M., BERNARDINI, S., FERRARESI, A. et ZANCHETTA, E. (2009). The WaCky Wide Web : A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43:209–226.

BATISTA, D. S. (2018). Named-entity evaluation metrics based on entity-level. https://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/. Consulté le 18 juin 2022.

BELTAGY, I., LO, K. et COHAN, A. (2019). SciBERT : A pretrained language model for scientific text. *In* INUI, K., JIANG, J., NG, V. et WAN, X., éditeurs : *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

BENGIO, Y., SIMARD, P. et FRASCONI, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

BERDYUGINA, D. et CAVALLUCCI, D. (2021). Automatic Extraction of Potentially Contradictory Parameters from Specific Field Patent Texts. *In* BORGIANNI, Y., BRAD, S., CAVALLUCCI, D. et LIVOTOV, P., éditeurs : *Creative Solutions for a Sustainable Development*, pages 150–161, Cham. Springer International Publishing.

BERTOLUCI, G. (2001). *Proposition d'une méthode d'amélioration de la cohérence des processus industriels*. Thèse de doctorat, Arts et Métiers ParisTech.

BHATIA, V. K. (1993). *Analysing Genre : Language Use in Professional Settings*. Applied Linguistics and Language Study. Longman, London.

BIBER, D. (1988). *Variation across Speech and Writing*. Cambridge University Press.

BIKAUN, T., HODKIEWICZ, M. et LIU, W. (2024a). MaintNorm : A corpus and benchmark model for lexical normalisation and masking of industrial maintenance short text. *In* van der GOOT, R., BAK, J., MÜLLER-EBERSTEIN, M., XU, W., RITTER, A. et BALDWIN, T., éditeurs : *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, pages 68–78, San Ġiljan, Malta. Association for Computational Linguistics.

BIKAUN, T. K., FRENCH, T., STEWART, M., LIU, W. et HODKIEWICZ, M. (2024b). MaintIE : A fine-grained annotation schema and benchmark for information extraction from maintenance short texts. *In* CALZOLARI, N., KAN, M.-Y., HOSTE, V., LENCI, A., SAKTI, S. et XUE, N., éditeurs : *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10939–10951, Torino, Italia. ELRA and ICCL.

BLEI, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

BLEI, D. M., NG, A. Y. et JORDAN, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

BOAS, H. C. (2009). 3. Semantic frames as interlingual representations for multilingual lexical databases. *In* BOAS, H. C., éditeur : *Multilingual FrameNets in Computational Lexicography : Methods and Applications*, pages 59–100. De Gruyter Mouton.

BOLDING, Q., LIAO, B., DENIS, B., LUO, J. et MONZ, C. (2023). Ask language model to clean your noisy translation data. *In* BOUAMOR, H., PINO, J. et BALI, K., éditeurs : *Findings of the Association for Computational Linguistics : EMNLP 2023*, pages 3215–3236, Singapore. Association for Computational Linguistics.

BOLDRINI, J.-C. (2005). *L'accompagnement des projets d'innovation. Le suivi de l'introduction de la méthode TRIZ dans des entreprises de petite taille.* Thèse de doctorat, Université de Nantes.

BOULDING, K. (1961). *The Image : Knowledge in Life and Society*. Ann Arbor paperbacks. University of Michigan Press.

BOURIGAULT, D., AUSSENAC-GILLES, N. et CHARLET, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, 18:87–110.

BOUTET, J. (2001a). La part langagière du travail : bilan et évolution. *Langage et société*, 98(4):17–42.

BOUTET, J. (2001b). Les mots du travail. *In Langage et Travail*, chapitre 8, pages 203–30. A. Borzeix et B. Fraenkel (Eds.).

BRUNDAGE, M. P., SEXTON, T., HODKIEWICZ, M., DIMA, A. et LUKENS, S. (2021). Technical language processing : Unlocking maintenance knowledge. *Manufacturing Letters*, 27:42–46.

BULTEY, A., BEUVRON, F. D. B. D. et ROUSSELOT, F. (2007). A substance-field ontology to support the TRIZ thinking approach. *International Journal of Computer Applications in Technology*, 30(1/2):113.

BURCHARDT, A., ERK, K., FRANK, A., KOWALSKI, A., PADÓ, S. et PINKAL, M. (2009). Using framenet for the semantic analysis of german : Annotation, representation, and automation. *In* BOAS, H. C., éditeur : *Multilingual FrameNets in Computational Lexicography : Methods and Applications*, pages 209–244. De Gruyter Mouton, Berlin, New York.

CANDITO, M., AMSILI, P., BARQUE, L., BENAMARA, F., de CHALENDAR, G., DJEMAA, M., HAAS, P., HUYGHE, R., MATHIEU, Y. Y., MULLER, P., SAGOT, B. et VIEU, L. (2014). Developing a French FrameNet : Methodology and first results. *In* CALZOLARI, N., CHOUKRI, K., DECLERCK, T., LOFTSSON, H., MAEGAARD, B., MARIANI, J., MORENO, A., ODIJK, J. et PIPERIDIS, S., éditeurs : *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1372–1379, Reykjavik, Iceland. European Language Resources Association (ELRA).

CANDITO, M. et SEDDAH, D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus : Syntactic annotation and use for a parser lexical domain adaptation method) [in French]. *In* ANTONIADIS, G., BLANCHON, H. et SÉRASSET, G., éditeurs : *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, pages 321–334, Grenoble, France. ATALA/AFCP.

CASCINI, G., FANTECHI, A. et SPINICCI, E. (2004). Natural language processing of patents and technical documentation. *Document Analysis Systems VI*, 3163:89–92.

CASSE, C. et CAROLY, S. (2014). Concevoir un dispositif de Retour d'Expérience intégré à l'activité collective. *In 49 ème congrès international. Société d'Ergonomie de Langue Française*.

CATINO, M. (2008). A Review of Literature : Individual Blame vs. Organizational Function Logics in Accident Analysis. *Journal of Contingencies and Crisis Management*, 16:53–62.

CAVALLUCCI, D. (1999). Triz : l'approche altshullerienne de la créativité. *Techniques de l'ingénieur*, 2(A5211):1–18.

CAVALLUCCI, D., LUTZ, P. et KUCHEAVARY, D. (2002). Converging in problem formulation : a different path in design. *In Proceedings of the Design Engineering Technical Conferences and Computer and Information in Engineering Conference*, Montreal, Canada.

CHARNOCK, R. (1999). Les langues de spécialité et le langage technique : considérations didactiques. *ASp. la revue du GERAS*, pages 281–302.

CHEN, C., LIU, G., ZHANG, Y., LI, Y. et TAO, F. (2017). SAO-based semantic mining of patents for semi-automatic construction of a customer job map. *Sustainability*, 9(8):1386.

CHEVREAU, F.-R. et WYBO, J.-l. (2007). Approche pratique de la culture de sécurité. Pour une maîtrise des risques industriels plus efficace. *Revue française de gestion*, 174(5):171–189.

CHOI, S., YOON, J., KIM, K., LEE, J. Y. et KIM, C.-H. (2011). SAO network analysis of patents for technology trends identification : A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. *Scientometrics*, 88(3):863–883.

CHOUDHRY, R. M., FANG, D. et MOHAMED, S. (2007). The nature of safety culture : A survey of the state-of-the-art. *Safety Science*, 45(10):993–1012.

CHOULIER, D. et DRAGHICI, G. (2000). Triz : une approche de résolution des problèmes d'innovation dans la conception de produits. *Récupéré sur http ://www. mec. utt. ro/˜ draghici/draghici_mef00. pdf*.

CHUANG, J., MANNING, C. et HEER, J. (2012). Termite : Visualization techniques for assessing textual topic models.

CHYBOWSKI, L. (2017). The usage of the Miniature Dwarfs method in the improvement of passenger ship construction. *Scientific Journals of the Maritime University of Szczecin*, 51:28–34.

CONDAMINES, A. (1997). Langue spécialisée ou discours spécialisé ? *In* LAPIERRE, L., OORE, I. et RUNTE, H., éditeurs : *Mélanges de linguistique offerts à Rostislav Kocourek*, pages 171–184. Les presses d'Alfa.

CONDAMINES, A. (2005). Linguistique de corpus et terminologie. *Langages*, 39(157):36–47.

CONDAMINES, A. et REBEYROLLE, J. (1997). Point de vue en langue spécialisée. *Meta*, 42(1):174–184.

CONNEAU, A., KHANDELWAL, K., GOYAL, N., CHAUDHARY, V., WENZEK, G., GUZMÁN, F., GRAVE, E., OTT, M., ZETTLEMOYER, L. et STOYANOV, V. (2020). Unsupervised cross-lingual representation learning at scale. *In* JURAFSKY, D., CHAI, J., SCHLUTER, N. et TETREAULT, J., éditeurs : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

COOPER, M. (2000). Towards a model of safety culture. *Safety Science*, 36(2):111–136.

CRUSE, D. A. (1973). Some thoughts on agentivity. *Journal of Linguistics*, 9(1):11–23.

DAL PONT, J.-P. (2003). Sécurité et gestion des risques. *Techniques de l'ingénieur. Sécurité et gestion des risques*, SE12:SE12–1.

DANIELLOU, F., BOISSIÈRES, I. et SIMARD, M. (2010). *Les facteurs humains et organisationnels de la sécurité industrielle : un état de l'art*. Les cahiers de la sécurité industrielle. FonCSI.

DAS, D., CHEN, D., MARTINS, A. F. T., SCHNEIDER, N. et SMITH, N. A. (2014). Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.

DEVASIER, J., MEDIRATTA, R. et LI, C. (2025). Can LLMs Extract Frame-Semantic Arguments ? *ArXiv :2502.12516*.

DEVLIN, J., CHANG, M.-W., LEE, K. et TOUTANOVA, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. *In* BURSTEIN, J., DORAN, C. et SOLORIO, T., éditeurs : *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

DIMA, A., LUKENS, S., HODKIEWICZ, M., SEXTON, T. et BRUNDAGE, M. P. (2021). Adapting natural language processing for technical text. *Applied AI Letters*, 2(3):e33.

DIS, E., BOLLEN, J., ZUIDEMA, W., ROOIJ, R. et BOCKTING, C. (2023). Chatgpt : five priorities for research. *Nature*, 614:224–226.

DJEMAA, M. (2017). *Stratégie domaine par domaine pour la création d'un FrameNet du français : annotations en corpus de cadres et rôles sémantiques.* Thèse de doctorat, Sorbonne Paris Cité.

DJEMAA, M., CANDITO, M., MULLER, P. et VIEU, L. (2016). Corpus annotation within the French FrameNet : a domain-by-domain methodology. *In* CALZOLARI, N., CHOUKRI, K., DECLERCK, T., GOGGI, S., GROBELNIK, M., MAEGAARD, B., MARIANI, J., MAZO, H., MORENO, A., ODIJK, J. et PIPERIDIS, S., éditeurs : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3794–3801, Portorož, Slovenia. European Language Resources Association (ELRA).

DOLBEY, A. (2009). *BioFrameNet : A FrameNet Extension to the Domain of Molecular Biology*. Thèse de doctorat, University of California, Berkeley.

DOWTY, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

DUBOIS, J. et DUBOIS-CHARLIER, F. (1997). *Les verbes français*. Larousse-Bordas, Paris, France.

DUPONT, G. et VERVEN, R. (1999). Analyse et mise en oeuvre de la méthode de conception triz. *Rapport de PFE*, 56.

EL BOUKKOURI, H. (2020). Ré-entraîner ou entraîner soi-même ? Stratégies de pré-entraînement de BERT en domaine médical. *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition).*, Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL:29–42.

EL BOUKKOURI, H., FERRET, O., LAVERGNE, T. et ZWEIGENBAUM, P. (2019). Embedding Strategies for Specialized Domains : Application to Clinical Entity Recognition. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, pages 295–301, Florence, Italy. Association for Computational Linguistics.

ERDMANN, K. O. (1922). *Die Bedeutung Des Wortes : Aufsätze Aus Dem Grenzgebiet Der Sprachpsychologie Und Logik.* s.n., Leipzig.

ETTINGER, A., HWANG, J., PYATKIN, V., BHAGAVATULA, C. et CHOI, Y. (2023). "you are an expert linguistic annotator" : Limits of LLMs as analyzers of Abstract Meaning Representation. *In* BOUAMOR, H., PINO, J. et BALI, K., éditeurs : *Findings of the Association for Computational Linguistics : EMNLP 2023*, pages 8250–8263, Singapore. Association for Computational Linguistics.

EVERSHEIM, W. (2009). *Innovation Management for Technical Products : Systematic and Integrated Product Development and Production Planning*. Springer-Verlag, Berlin, Heidelberg.

FALZON, P. (1986). *Langages opératifs et compréhension opérative*. Thèse de doctorat, Université Paris Descartes.

FALZON, P. (1987). Langages opératifs et compréhension opérative. *Le Travail Humain*, n°50:281–286.

FALZON, P. (1989). *Ergonomie cognitive du dialogue*. Presses universitaires de Grenoble.

FAN, R.-E., CHANG, K.-W., HSIEH, C.-J., WANG, X.-R. et LIN, C.-J. (2008). Liblinear : A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874.

FELLBAUM, C., éditeur (1998). *WordNet : An Electronic Lexical Database*. MIT Press, Cambridge, MA.

FEY, V. et RIVIN, E. (2005). *Innovation on Demand : New Product Development Using TRIZ*. Cambridge University Press.

FILLMORE, C. J. (1967). The case for case. *Universals in Linguistic Theory*, pages 1–25.

FILLMORE, C. J. (1976). Frame Semantics and the Nature of Language*. *Annals of the New York Academy of Sciences*, 280(1):20–32.

FILLMORE, C. J. (1977a). The case for case reopened. *Grammatical Relations*, 8:59–81.

FILLMORE, C. J. (1977b). Scenes and frames semantics. *Fundamental Studies in Computer Science*, 59.

FILLMORE, C. J. (1982). Frame semantics. *Linguistics in the Morning Calm.–Seoul : Hanshin Publishing Co*, pages 111–137.

FILLMORE, C. J. (1985). Frames and the semantics of understanding. *Quaderni di semantica*, 6:222–254.

FILLMORE, C. J. (2007). Valency issues in framenet. *In* HERBST, T. et GÖTZ-VOTTELER, K., éditeurs : *Valency : Theoretical, Descriptive and Cognitive Issues*, pages 129–162. De Gruyter Mouton, Berlin, New York.

FILLMORE, C. J. et BAKER, C. (2012). A Frames Approach to Semantic Analysis. *In* HEINE, B. et NARROG, H., éditeurs : *The Oxford Handbook of Linguistic Analysis*, pages 313–340. Oxford University Press, 1 édition.

FRADIN, B. et WINTERSTEIN, G. (2012). Tuning agentivity and instrumentality : deverbal nouns in-oir revisited. *Paper delivered at Décembrettes*, 8:6–7.

FRANCIS, C., PLANCHETTE, G., SIGNORET, J. P., THING-LEO, G., CHANG, L. et MÉRIAN, Y. (2020). Risques et métiers du risque dans l'entreprise industrielle. *In Congrès Lambda Mu 22 " Les risques au cœur des transitions " (e-congrès) - 22e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement, Institut pour la Maîtrise des Risques*, Le Havre (e-congrès), France.

FUCKS, I. (2013). L'énigme de la culture de sécurité dans les organisations à risques : une approche anthropologique :. *Le travail humain*, Vol. 75(4):399–420.

FÜRSTENAU, Hagen et Lapata, M. (2009). Étiquetage sémantique des rôles semi-supervisé. *In* LASCARIDES, A., GARDENT, C. et NIVRE, J., éditeurs : *Actes de la 12e Conférence du Chapitre Européen de l'ACL (EACL 2009)*, pages 220–228, Athènes, Grèce. Association pour la linguistique computationnelle.

GADD, K. et GODDARD, C. (2011). *TRIZ for Engineers : Enabling Inventive Problem Solving*. Wiley.

GALAND, L., KURELA, M. et CLAVIJO, H. R. (2018). Techniques de TAL pour la recherche des "signaux faibles" et catégorisation des risques dans le REX SDF des lanceurs spatiaux. *In Congrès Lambda Mu 21, " Maîtrise des risques et transformation numérique : opportunités et menaces "*, Reims, France.

GENTILHOMME-KOUTYRINE, Y. (1994). Regards sur la terminologisation en lexicologie. *Meta*, 39(4):546–560.

GHANE, M., ANG, M. C., CAVALLUCCI, D., ABDUL KADIR, R., NG, K. W. et SOROOSHIAN, S. (2024). Semantic TRIZ feasibility in technology development, innovation, and production : A systematic review. *Heliyon*, 10(1):e23775.

GILBERT, C. (2001). Retours d'expérience : le poids des contraintes. *In Annales des mines*, volume 22.

GILDEA, D. et JURAFSKY, D. (2000). Automatic Labeling of Semantic Roles. *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong. Association for Computational Linguistics.

GILDEA, D. et JURAFSKY, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

GILDEA, D. et PALMER, M. (2002). The necessity of parsing for predicate argument recognition. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 239–246, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

GÓMEZ-PÉREZ, J. M., GARCÍA-SILVA, A., LEONE, R., ALBANI, M., FONTAINE, M., PONCET, C., SUMMERER, L., DONATI, A., ROMA, I. et SCAGLIONI, S. (2022). Artificial Intelligence and Natural Language Processing and Understanding in Space : A Methodological Framework and Four ESA Case Studies.

GORDON, A. et SWANSON, R. (2007). Generalizing semantic role annotations across syntactically similar verbs. *In* ZAENEN, A. et van den BOSCH, A., éditeurs : *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 192–199, Prague, Czech Republic. Association for Computational Linguistics.

GORDON, W. J. J. (1965). *Stimulation des facultés créatrices dans les groupes de recherche par la méthode synectique*. Éditions Hommes et techniques, Suresnes.

GRATTAFIORI, A., DUBEY, A., JAUHRI, A., PANDEY, A., KADIAN, A., AL-DAHLE, A., LETMAN, A., MATHUR, A., SCHELTEN, A., VAUGHAN, A., YANG, A., FAN, A., GOYAL, A., HARTSHORN, A., YANG, A., MITRA, A., SRAVANKUMAR, A., KORENEV, A., HINSVARK, A., RAO, A., ZHANG, A., RODRIGUEZ, A., GREGERSON, A., SPATARU, A., ROZIERE, B., BIRON, B., TANG, B., CHERN, B., CAUCHETEUX, C., NAYAK, C., BI, C., MARRA, C., MCCONNELL, C., KELLER, C., TOURET, C., WU, C., WONG, C., FERRER, C. C., NIKOLAIDIS, C., ALLONSIUS, D., SONG, D., PINTZ, D., LIVSHITS, D., WYATT, D., ESIOBU, D., CHOUDHARY, D., MAHAJAN, D., GARCIA-OLANO, D., PERINO, D., HUPKES, D., LAKOMKIN, E., ALBADAWY, E., LOBANOVA, E., DINAN, E., SMITH, E. M., RADENOVIC, F., GUZMÁN, F., ZHANG, F., SYNNAEVE, G., LEE, G., ANDERSON, G. L., THATTAI, G., NAIL, G., MIALON, G., PANG, G., CUCURELL, G., NGUYEN, H., KOREVAAR, H., XU, H., TOUVRON, H., ZAROV, I., IBARRA, I. A., KLOUMANN, I., MISRA, I., EVTIMOV, I., ZHANG, J., COPET, J., LEE, J., GEFFERT, J., VRANES, J., PARK, J., MAHADEOKAR, J., SHAH, J., van der LINDE, J., BILLOCK, J., HONG, J., LEE, J., FU, J., CHI, J., HUANG, J., LIU, J., WANG, J., YU, J., BITTON, J., SPISAK, J., PARK, J., ROCCA, J., JOHNSTUN, J., SAXE, J., JIA, J., ALWALA, K. V., PRASAD, K., UPASANI, K., PLAWIAK, K., LI, K., HEAFIELD, K., STONE, K., EL-ARINI, K., IYER, K., MALIK, K., CHIU, K., BHALLA, K., LAKHOTIA, K., RANTALA-YEARY, L., van der MAATEN, L., CHEN, L., TAN, L., JENKINS, L., MARTIN, L., MADAAN, L., MALO, L., BLECHER, L., LANDZAAT, L., de OLIVEIRA, L., MUZZI, M., PASUPULETI, M., SINGH, M., PALURI, M., KARDAS, M., TSIMPOUKELLI, M., OLDHAM, M., RITA, M., PAVLOVA, M., KAMBADUR, M., LEWIS, M., SI, M., SINGH, M. K., HASSAN, M., GOYAL, N., TORABI, N., BASHLYKOV, N., BOGOYCHEV, N., CHATTERJI, N.,

Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L.,

KHABSA, M., AVALANI, M., BHATT, M., MANKUS, M., HASSON, M., LENNIE, M., RESO, M., GROSHEV, M., NAUMOV, M., LATHI, M., KENEALLY, M., LIU, M., SELTZER, M. L., VALKO, M., RESTREPO, M., PATEL, M., VYATSKOV, M., SAMVELYAN, M., CLARK, M., MACEY, M., WANG, M., HERMOSO, M. J., METANAT, M., RASTEGARI, M., BANSAL, M., SANTHANAM, N., PARKS, N., WHITE, N., BAWA, N., SINGHAL, N., EGEBO, N., USUNIER, N., MEHTA, N., LAPTEV, N. P., DONG, N., CHENG, N., CHERNOGUZ, O., HART, O., SALPEKAR, O., KALINLI, O., KENT, P., PAREKH, P., SAAB, P., BALAJI, P., RITTNER, P., BONTRAGER, P., ROUX, P., DOLLAR, P., ZVYAGINA, P., RATANCHANDANI, P., YUVRAJ, P., LIANG, Q., ALAO, R., RODRIGUEZ, R., AYUB, R., MURTHY, R., NAYANI, R., MITRA, R., PARTHASARATHY, R., LI, R., HOGAN, R., BATTEY, R., WANG, R., HOWES, R., RINOTT, R., MEHTA, S., SIBY, S., BONDU, S. J., DATTA, S., CHUGH, S., HUNT, S., DHILLON, S., SIDOROV, S., PAN, S., MAHAJAN, S., VERMA, S., YAMAMOTO, S., RAMASWAMY, S., LINDSAY, S., LINDSAY, S., FENG, S., LIN, S., ZHA, S. C., PATIL, S., SHANKAR, S., ZHANG, S., ZHANG, S., WANG, S., AGARWAL, S., SAJUYIGBE, S., CHINTALA, S., MAX, S., CHEN, S., KEHOE, S., SATTERFIELD, S., GOVINDAPRASAD, S., GUPTA, S., DENG, S., CHO, S., VIRK, S., SUBRAMANIAN, S., CHOUDHURY, S., GOLDMAN, S., REMEZ, T., GLASER, T., BEST, T., KOEHLER, T., ROBINSON, T., LI, T., ZHANG, T., MATTHEWS, T., CHOU, T., SHAKED, T., VONTIMITTA, V., AJAYI, V., MONTANEZ, V., MOHAN, V., KUMAR, V. S., MANGLA, V., IONESCU, V., POENARU, V., MIHAILESCU, V. T., IVANOV, V., LI, W., WANG, W., JIANG, W., BOUAZIZ, W., CONSTABLE, W., TANG, X., WU, X., WANG, X., WU, X., GAO, X., KLEINMAN, Y., CHEN, Y., HU, Y., JIA, Y., QI, Y., LI, Y., ZHANG, Y., ZHANG, Y., ADI, Y., NAM, Y., YU, WANG, ZHAO, Y., HAO, Y., QIAN, Y., LI, Y., HE, Y., RAIT, Z., DEVITO, Z., ROSNBRICK, Z., WEN, Z., YANG, Z., ZHAO, Z. et MA, Z. (2024). The llama 3 herd of models. arXiv :2407.21783.

GROSS, M. (1975). *Méthodes en syntaxe : régime des constructions complétives*. Actualités scientifiques et industrielles. Hermann, Paris, France.

GRUBER, J. S. (1965). *Studies in Lexical Relations*. Thèse de doctorat, Massachusetts Institute of Technology.

GUILBERT, L. (1973). La spécificité du terme scientifique et technique. *Langue française*, 17(1):5–17.

GULDENMUND, F. (2000). The nature of safety culture : a review of theory and research. *Safety Science*, 34(1):215–257.

HADJ-MABROUK, H. (2004). Retour d'expérience et facteur humain. application à la sécurité des transports ferroviaires. *In 39e Congrès annuel de l'Association Québécoise du Transports et des Routes (AQTR'2004)*, pages 1–20.

HADJ-MABROUK, H. et HAMDAOUI, F. (2008). Apport du retour d'expérience à l'analyse des risques. *16ème Congrès "Lambda Mu", Maîtrise des risques et sûreté de fonctionnement, IMdR, Institut pour la Maîtrise des Risques*.

HALE, A. (2000). Culture's confusions. *Safety Science*, 34(1):1–14.

HARRIS, Z. (1990). *Mathematical Structures of Language*. Interscience Publisher.

HARTMANN, S., KUZNETSOV, I., MARTIN, T. et GUREVYCH, I. (2017). Out-of-domain FrameNet semantic role labeling. *In* LAPATA, M., BLUNSOM, P. et KOLLER, A., éditeurs : *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, pages 471–482, Valencia, Spain. Association for Computational Linguistics.

HENDRIX, G. G., THOMPSON, C. W. et SLOCUM, J. (1973). Language processing via canonical verbs and semantic models. *In International Joint Conference on Artificial Intelligence*.

HESSLOW, G. (1988). The problem of causal selection. *In Contemporary Science and Natural Explanation : Commonsense Conceptions of Causality.*, pages 11–32. New York University Press, New York, NY, US.

HIRTZ, J., STONE, R. B., MCADAMS, D. A., SZYKMAN, S. et WOOD, K. L. (2002). A functional basis for engineering design : Reconciling and evolving previous efforts. *Research in Engineering Design*, 13(2):65–82.

HOCHREITER, S. et SCHMIDHUBER, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

HODKIEWICZ, M. et HO, M. (2016). Cleaning historical maintenance work order data for reliability analysis. *Journal of Quality in Maintenance Engineering*, 22:146–163.

HUANG, Y., WU, S., GAO, C., CHEN, D., ZHANG, Q., WAN, Y., ZHOU, T., XIAO, C., GAO, J., SUN, L. et ZHANG, X. (2025). Datagen : Unified synthetic dataset generation via large language models. *In The Thirteenth International Conference on Learning Representations*.

ILEVBARE, I. M., PROBERT, D. et PHAAL, R. (2013). A review of TRIZ, and its benefits and challenges in practice. *Technovation*, 33(2):30–37.

INSTITUT POUR UNE CULTURE DE SÉCURITÉ INDUSTRIELLE (ICSI) (2017). Culture de sécurité : Les essentiels de l'icsi. Consulté le 07/03/2025.

JIANG, A. Q., SABLAYROLLES, A., MENSCH, A., BAMFORD, C., CHAPLOT, D. S., de las CASAS, D., BRESSAND, F., LENGYEL, G., LAMPLE, G., SAULNIER, L., LAVAUD, L. R., LACHAUX, M.-A., STOCK, P., SCAO, T. L., LAVRIL, T., WANG, T., LACROIX, T. et SAYED, W. E. (2023). Mistral 7b.

JIN, Y., CHANDRA, M., VERMA, G., HU, Y., DE CHOUDHURY, M. et KUMAR, S. (2024). Better to ask in english : Cross-lingual evaluation of large language models for healthcare queries. *In Proceedings of the ACM Web Conference 2024*, WWW '24, page 2627–2638, New York, NY, USA. Association for Computing Machinery.

JOHNSON, C. W. (2003). *Failure in Safety-Critical Systems : A Handbook of Accident and Incident Reporting*. University of Glasgow Press, Glasgow, Scotland. Available in electronic form.

JURAFSKY, D. et MARTIN, J. H. (2024). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd édition. Online manuscript released August 20, 2024.

KIM, S., PARK, I. et YOON, B. (2020). SAO2Vec : Development of an algorithm for embedding the subject–action–object (SAO) structure using Doc2Vec. *PLOS ONE*, 15(2):e0227930.

KINGSBURY, P. et PALMER, M. (2002). From treebank to propbank. *In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands, Spain. European Language Resources Association (ELRA).

KIPPER, K., KORHONEN, A., RYANT, N. et PALMER, M. (2006). Extending verbnet with novel verb classes. *In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy. European Language Resources Association (ELRA).

KITTREDGE, R. I. (2005). Sublanguages and Controlled Languages. *In* MITKOV, R., éditeur : *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

KOCOUREK, R. (1991). *La Langue Française de la Technique et de la Science*. La documentation française.

KURELA, M., BACQUÉ, M. et LAURENT, R. (2020). Classification automatique des faits techniques pour la conformité des lanceurs spatiaux. *In Congrès Lambda Mu 22 " Les risques au cœur des transitions " (e-congrès) - 22e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement, Institut pour la Maîtrise des Risques*, Le Havre (e-congrès), France.

KÖPF, A., KILCHER, Y., von RÜTTE, D., ANAGNOSTIDIS, S., TAM, Z.-R., STEVENS, K., BARHOUM, A., DUC, N. M., STANLEY, O., NAGYFI, R., ES, S., SURI, S., GLUSHKOV, D., DANTULURI, A., MAGUIRE, A., SCHUHMANN, C., NGUYEN, H. et MATTICK, A. (2023). Openassistant conversations – democratizing large language model alignment. *arXiv preprint arXiv :2304.07327*.

LEE, J., YOON, W., KIM, S., KIM, D., KIM, S., SO, C. H. et KANG, J. (2019). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

LERAT, P. (1995). *Les Langues Spécialisées*. Presses Universitaires de France.

LEVESON, N. (2004). A new accident model for engineering safer systems. *Safety Science*, 42(4):237–270.

LEVIN, B. (1993). *English Verb Classes and Alternations : A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

LEVIN, B. et RAPPAPORT HOVAV, M. (2005). *Argument Realization*. Research Surveys in Linguistics. Cambridge University Press.

L'HOMME, M.-C. (2023). Managing polysemy in terminological resources. *Terminology International Journal of Theoretical and Applied Issues in Specialized Communication*.

L'HOMME, M.-C., ROBICHAUD, B. et SUBIRATS, C. (2020). Building multilingual specialized resources based on FrameNet : Application to the field of the environment. *In* TORRENT, T. T., BAKER, C. F., CZULO, O., OHARA, K. et PETRUCK, M. R. L., éditeurs : *Proceedings of the International FrameNet Workshop 2020 : Towards a Global, Multilingual FrameNet*, pages 85–92, Marseille, France. European Language Resources Association.

LIM, S., LECOZE, J.-C. et DECHY, N. (2002). Intégration des aspects organisationnels dans le rex. l'accident majeur, un phénomène complexe à étudier. Rapport Technique 366988, INERIS.

LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L. et STOYANOV, V. (2019). Roberta : A robustly optimized bert pretraining approach. cite arxiv :1907.11692.

LIVOTOV, P. (2008). TRIZ and Innovation Management. *INNOVATOR*.

LÖNNEKER-RODMAN, B. et BAKER, C. F. (2009). The framenet model and its applications. *Natural Language Engineering*, 15(3):415–453.

MAHOWALD, K., IVANOVA, A. A., BLANK, I. A., KANWISHER, N., TENENBAUM, J. B. et FEDORENKO, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

MANN, D. (2002). *Hands-On Systematic Innovation*. Creax Press, Ieper, Belgium.

MARCUS, M. P., SANTORINI, B. et MARCINKIEWICZ, M. A. (1993). Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

MARTIN, L., MULLER, B., ORTIZ SUÁREZ, P. J., DUPONT, Y., ROMARY, L., de la CLERGERIE, É., SEDDAH, D. et SAGOT, B. (2020). CamemBERT : a tasty French language model. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

MATHET, Y., WIDLÖCHER, A. et MÉTIVIER, J.-P. (2015). The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, 41(3):437–479.

MICHELOT, C. (2015). L'invention de la créativité. *Phronesis*, 4(2):54–61.

MIKOLOV, T., CHEN, K., CORRADO, G. S. et DEAN, J. (2013). Efficient estimation of word representations in vector space. *In International Conference on Learning Representations*.

MILETIC, F. (2022). *An investigation into contact-induced semantic shifts in Quebec English : conciliating corpus-based vector models and variationist sociolinguistic inquiry*. Theses, Université Toulouse le Mirail - Toulouse II.

MILLER, G. A. (1995). Wordnet : A lexical database for english. *Communications of the ACM*, 38(11):39–41.

MOEHRLE, M., WALTER, L., GERITZ, A. et MÜLLER, S. (2005). Patent-Based Inventor Profiles as a Basis for Human Resource Decisions in Research and Development. *R&D Management*, 35:513–524.

MOEHRLE, M. G. (2005). What is TRIZ? From Conceptual Basics to a Framework for Research. *Creativity and Innovation Management*, 14(1):3–13.

MONTMAIN, J., PENALVA, J.-M., AKHARRAZ, A., CHAPURLAT, V., COUTURIER, P., CRAMPES, M., DRAY, G., JANAQI, S., MARC, I., PONCELET, P., RICCIO, P.-M., RUNTZ, P., VACHER, B. et VASQUEZ, M. (2003). État de l'art sur les théories de la décision et méthodologies de l'approche système. Rapport technique RMT03-018, Ministère de l'Équipement, des Transports, du Logement, du Tourisme et de la Mer, France.

MORTUREUX, M.-F. (1995). Les vocabulaires scientifiques et techniques. *Les Carnets du Cediscor*, 3:13–25.

MUELLER, J. et THYAGARAJAN, A. (2016). Siamese recurrent architectures for learning sentence similarity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

NI, X., SAMET, A. et CAVALLUCCI, D. (2021). Replicating TRIZ Reasoning Through Deep Learning. *In* BORGIANNI, Y., BRAD, S., CAVALLUCCI, D. et LIVOTOV, P., éditeurs : *Creative Solutions for a Sustainable Development*, volume 635, pages 330–339. Springer International Publishing, Cham.

NI, X., SAMET, A. et CAVALLUCCI, D. (2022). Similarity-based approach for inventive design solutions assistance. *Journal of Intelligent Manufacturing*, 33(6):1681–1698.

OHARA, K., FUJII, S., ISHIZAKI, S., OHORI, T., SAITO, H. et SUZUKI, R. (2004). The Japanese FrameNet project ; an introduction. *In* FILLMORE, C. J., PINKAL, M., BAKER, C. F. et ERK, K., éditeurs : *Proceedings of the Workshop on Building Lexical Resources from Semantically Annotated Corpora*, pages 9–12, Lisbon. LREC 2004, LREC 2004.

OHARA, K. H. (2009). *6. Frame-based contrastive lexical semantics in Japanese FrameNet : The case of risk and kakeru*, pages 163–182. De Gruyter Mouton, Berlin, New York.

OPENAI, ACHIAM, J., ADLER, S., AGARWAL, S., AHMAD, L., AKKAYA, I., LEONI ALEMAN, F., ALMEIDA, D., ALTENSCHMIDT, J., ALTMAN, S., ANADKAT, S., AVILA, R., BABUSCHKIN, I., BALAJI, S., BALCOM, V., BALTESCU, P., BAO, H., BAVARIAN, M., BELGUM, J., BELLO, I., BERDINE, J. et AL. (2023). GPT-4 Technical Report. *arXiv e-prints*, page arXiv :2303.08774.

ORTIZ SUÁREZ, P. J., SAGOT, B. et ROMARY, L. (2019). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. *In* BAŃSKI, P., BARBARESI, A., BIBER, H., BREITENEDER, E., CLEMATIDE, S., KUPIETZ, M., LÜNGEN, H. et ILIADI, C., éditeurs : *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom. Leibniz-Institut für Deutsche Sprache.

PALMER, M., GILDEA, D. et KINGSBURY, P. (2005). The proposition bank : A corpus annotated with semantic roles. *Computational Linguistics*, 31(1):71–106.

PALMER, M., GILDEA, D. et XUE, N. (2010). *Semantic Role Labeling*. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham.

PAN, S. J. et YANG, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359.

PERROW, C. (2011). *Normal Accidents : Living with High Risk Technologies - Updated Edition*. Princeton University Press.

PERSING, I. et NG, V. (2009). Semi-supervised cause identification from aviation safety reports. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2 - Volume 2*, ACL '09, pages 843–851, USA. Association for Computational Linguistics.

PHAL, M. (1970). Le vocabulaire général d'orientation scientifique : essai de définition et méthode d'enquête. *Les langues de spécialités, Analyse Linguistique et Recherche Pédagogique, AIDELA*, pages 94–115.

PLANK, B. (2016). What to do about non-standard (or non-canonical) language in NLP. *In Proceedings of KONVENS 2016*.

PRADHAN, S., HACIOGLU, K., KRUGLER, V., WARD, W., MARTIN, J. et JURAFSKY, D. (2005). Support vector learning for semantic argument classification. *Machine Learning*, 60:11–39.

PUSTEJOVSKY, J., BUNT, H. et ZAENEN, A. (2017). *Designing Annotation Schemes : From Theory to Model*, pages 21–72. Springer Netherlands, Dordrecht.

QI, P., ZHANG, Y., ZHANG, Y., BOLTON, J. et MANNING, C. (2020). Stanza : A python natural language processing toolkit for many human languages. *In Association for Computational Linguistics (ACL) System Demonstrations*, pages 101–108.

QUILLIAN, M. (1966). *Semantic Memory*. AFCRL (Series). Air Force Cambridge Research Laboratories, Office of Aerospace Research, United States Air Force.

QUINIOU, S., CELLIER, P., CHARNOIS, T. et LEGALLOIS, D. (2012). Fouille de données pour la stylistique : cas des motifs séquentiels émergents. *In 11es Journées Internationales d'Analyse Statistique des Données Textuelles (JADT'12)*, pages 821–833.

RAGAB-ZOUAOUA, D. (2012). *Lois d'évolution de TRIZ pour la conception des futures générations des produits : proposition d'un modèle*. Thèse de doctorat, Arts et Métiers ParisTech.

RANTANEN, K. et DOMB, E. (2002). *Simplified TRIZ : New Problem-Solving Applications for Engineers and Manufacturing Professionals*. CRC press.

REASON, J. T. (1997). *Managing the risks of organizational accidents*. Ashgate, Aldershot, Hants, England ;.

ROCHE, C. (2022). La définition des termes : une approche conceptuelle. *https ://hal.science/hal-03916622*. working paper or preprint.

ROGERS, A., KOVALEVA, O. et RUMSHISKY, A. (2020). A primer in BERTology : What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

ROSCH, E. H. (1973). Natural categories. *Cognitive Psychology*, 4(3):328–350.

ROTH, M. et LAPATA, M. (2015). Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460.

ROTH, M. et WOODSEND, K. (2014). Composition of word representations improves semantic role labelling. *In* MOSCHITTI, A., PANG, B. et DAELEMANS, W., éditeurs : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP})*, pages 407–413, Doha, Qatar. Association for Computational Linguistics.

ROUSSELOT, F., BULTEY, A. et ZANNI, C. (2008). Formalisation des connaissances capitalisées en conception inventive. *In Actes des 19es Journées Francophones d'Ingénierie des Connaissances (IC 2008)*, pages 173–182. 19es Journées Francophones d'Ingénierie des Connaissance.

RUPPENHOFER, J., ELLSWORTH, M., PETRUCK, M., JOHNSON, C. et SCHEFFCZYK, J. (2016). *FrameNet II : Extended Theory and Practice*. http://framenet.icsi.berkeley.edu.

RUSSO, D. et DUCI, S. (2015). From Altshuller's 76 Standard Solutions to a New Set of 111 Standards. *Procedia Engineering*, 131:747–756.

SANACORE, D. (2023). *Une Histoire de Famille : Description Morphosémantique Des Lexèmes Construits et Des Relations Dérivationnelles*. Thèse de doctorat, Toulouse II.

SAVRANSKY, S. (2000). *Engineering of Creativity : Introduction to TRIZ Methodology of Inventive Problem Solving*. CRC Press.

SCHMIDT, T. (2009). *4. The Kicktionary – a multilingual lexical resource of football language*, pages 101–134. De Gruyter Mouton, Berlin, New York.

SHULYAK, L. (1998). *Introduction to TRIZ*, pages 15–22. Worcester, Mass. : Technical Innovation Center.

SIMMONS, R. (1974). Semantic networks : Their computation and use for understanding english sentences. *In Computer Models of Thought and Language*, pages 61–113. W.H. Freeman Co.

SNOW, R., O'CONNOR, B., JURAFSKY, D. et NG, A. (2008). Cheap and fast – but is it good ? evaluating non-expert annotations for natural language tasks. *In* LAPATA, M. et NG, H. T., éditeurs : *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

SOMERS, H. (1998). An attempt to use weighted cusums to identify sublanguages. *In Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning - NeMLaP3/CoNLL '98*, page 131, Sydney, Australia. Association for Computational Linguistics.

SOUCHKOV, V. (2019). A brief history of TRIZ. https://the-trizjournal.com/a-brief-history-of-triz/ [Consulté le : (05/12/2024)].

STALMASZCZYK, P. (1996). Theta roles and the theory of theta-binding. *In Papers and Studies in Contrastive Linguistics*, volume 31, page 97–110.

SUBIRATS, C. (2009). *5. Spanish FrameNet : A frame-semantic analysis of the Spanish lexicon*, pages 135–162. De Gruyter Mouton, Berlin, New York.

SUBIRATS, C. et PETRUCK, M. R. L. (2003). Surprise : Spanish framenet ! *International Congress of Linguists Workshop on Frame Semantics*, page 6.

SURDEANU, M., HARABAGIU, S., WILLIAMS, J. et AARSETH, P. (2003). Using predicate-argument structures for information extraction. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 8–15, Sapporo, Japan. Association for Computational Linguistics.

SWAYAMDIPTA, S., THOMSON, S., DYER, C. et SMITH, N. A. (2017). Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. *arXiv preprint arXiv :1706.09528*.

SWIER, R. S. et STEVENSON, S. (2004). Unsupervised Semantic Role Labelling. *Proceedings of the 2004 conference on empirical methods in natural language processing*.

TANGUY, L. et REBEYROLLE, J. (2020). Les titres des publications scientifiques en français : fouille de texte pour le repérage de schémas lexico-syntaxiques. *Le Français Moderne - Revue de linguistique française*, 1:137–156. https://hal.science/hal-02867968.

TANGUY, L., TULECHKI, N., URIELI, A., HERMANN, E. et RAYNAL, C. (2016). Natural language processing for aviation safety reports : From classification to interactive analysis. *Computers in Industry*, 78:80–95.

TEA, C. (2009). *Retour d'expérience et données subjectives : quel système d'information pour la gestion des risques ?* Thèse de doctorat, Arts et Métiers ParisTech.

THOMAS, A., GAIZAUSKAS, R. et LU, H. (2024). Leveraging LLMs for post-OCR correction of historical newspapers. *In* SPRUGNOLI, R. et PASSAROTTI, M., éditeurs : *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 116–121, Torino, Italia. ELRA and ICCL.

TONELLI, S. (2010). *Semi-automatic techniques for extending the FrameNet lexical database to new languages*. phd, Universit'a Ca' Foscari, Venezia.

TORRENT, T. T., HOFFMANN, T., ALMEIDA, A. L. et TURNER, M. (2024). *Copilots for Linguists : AI, Constructions, and Frames*. Elements in Construction Grammar. Cambridge University Press, Cambridge.

TOUTANOVA, K., HAGHIGHI, A. et MANNING, C. (2005). Joint learning improves semantic role labeling. *In* KNIGHT, K., NG, H. T. et OFLAZER, K., éditeurs : *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 589–596, Ann Arbor, Michigan. Association for Computational Linguistics.

TULECHKI, N. (2015). *Natural language processing of incident and accident reports : application to risk management in civil aviation*. Theses, Université Toulouse le Mirail - Toulouse II.

USUGA CADAVID, J. P., GRABOT, B., LAMOURI, S., PELLERIN, R. et FORTIN, A. (2022). Valuing free-form text data from maintenance logs through transfer learning with CamemBERT. *Enterprise Information Systems*, 16(6):1790043.

Van der GOOT, R., LJUBEŠIĆ, N., MATROOS, I., NISSIM, M. et PLANK, B. (2018). Bleaching text : Abstract features for cross-lingual gender prediction. *In* GUREVYCH, I. et MIYAO, Y., éditeurs : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 383–389, Melbourne, Australia. Association for Computational Linguistics.

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. u. et POLOSUKHIN, I. (2017). Attention is all you need. *In* GUYON, I., LUXBURG, U. V., BENGIO, S., WALLACH, H., FERGUS, R., VISHWANATHAN, S. et GARNETT, R., éditeurs : *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

VAUGHAN, D. (1990). Autonomy, interdependence, and social control : Nasa and the space shuttle challenger. *Administrative Science Quarterly*, 35(2):225–257.

VERBITSKY, M. (2004). Semantic triz. *TRIZ journal-Español*.

VERGELY, P. (2004). *Analyse Linguistique de l'expression Du Dysfonctionnement Technique : Le Cas Des Échanges Entre Chefs de Salle et Maintenance Opérationnelle Dans La Navigation Aérienne*. Thèse de doctorat, Université Toulouse II-Le Mirail.

VERHAEGEN, P.-A., JORIS, D., VANDEVENNE, D. et DUFLOU, R. J. (2011). Identifying candidates for design-by-analogy. *Computers in Industry*, 62(4):446–459.

WANG, B., WEI, C., LIU, Z., LIN, G. et CHEN, N. F. (2024). Resilience of large language models for noisy instructions. *In* AL-ONAIZAN, Y., BANSAL, M. et CHEN, Y.-N., éditeurs : *Findings of the Association for Computational Linguistics : EMNLP 2024*, pages 11939–11950, Miami, Florida, USA. Association for Computational Linguistics.

WANG, Y. et ZHAO, D. (2021). Cross-domain function analysis and trend study in chinese construction industry based on patent semantic analysis. *Technological Forecasting and Social Change*, 162:120331.

WARNIER, M. (2018). *Contribution de La Linguistique de Corpus à La Constitution de Langues Contrôlées Pour La Rédaction Technique : L'exemple Des Exigences de Projets Spatiaux*. Thèse de doctorat, Université Toulouse le Mirail - Toulouse II.

WENZEK, G., LACHAUX, M.-A., CONNEAU, A., CHAUDHARY, V., GUZMÁN, F., JOULIN, A. et GRAVE, E. (2020). CCNet : Extracting high quality monolingual datasets from web crawl data. *In* CALZOLARI, N., BÉCHET, F., BLACHE, P., CHOUKRI, K., CIERI, C., DECLERCK, T., GOGGI, S., ISAHARA, H., MAEGAARD, B., MARIANI, J., MAZO, H., MORENO, A., ODIJK, J. et PIPERIDIS, S., éditeurs : *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

WIDLÖCHER, A. et MATHET, Y. (2009). La plate-forme glozz : environnement d'annotation et d'exploration de corpus. *In Actes de la 16e Conférence Traitement Automatique des Langues Naturelles (TALN'09), session posters*, pages 10–p.

WIEGMANN, D. A., ZHANG, H., von THADEN, T., SHARMA, G. et MITCHELL, A. (2002). A synthesis of safety culture and safety climate research. *Urbana-Champaign (Ill) : Aviation research lab institute of aviation*.

WILKS, Y. (1973). An artificial intelligence approach to machine translation. *In Computer Models of Thought and Language*, pages 114–151. R. C. Schank and K. M. Colby.

WITTGENSTEIN, L. (1953). *Philosophical Investigations*. Basil Blackwell, Oxford.

WU, H., SHEN, G. Q., LIN, X., LI, M. et LI, C. Z. (2021). A transformer-based deep learning model for recognizing communication-oriented entities from patents of ict in construction. *Automation in Construction*, 125:103608.

WYBO, J.-L., COLARDELLE, C., POULOSSIER, M.-P. et CAUCHOIS, D. (2001). Méthodologie de partage d'expérience de gestion d'incidents. *Récents Progrès en Génie des Procédés*, 15(58):115–128.

WYBO, J. L., GODFRIN, V., COLARDELLE, C. et GUINET, V. (2003). Méthodologie de retour d'expérience des actions de gestion des risques. *Conférence MATE*.

YANG, A., YANG, B., ZHANG, B., HUI, B., ZHENG, B., YU, B., LI, C., LIU, D., HUANG, F., WEI, H., LIN, H., YANG, J., TU, J., ZHANG, J., YANG, J., YANG, J., ZHOU, J., LIN, J., DANG, K., LU, K., BAO, K., YANG, K., YU, L., LI, M., XUE, M., ZHANG, P., ZHU, Q., MEN, R., LIN, R., LI, T., TANG, T., XIA, T., REN, X., REN, X., FAN, Y., SU, Y., ZHANG, Y., WAN, Y., LIU, Y., CUI, Z., ZHANG, Z. et QIU, Z. (2024). Qwen2.5 technical report. *https ://arxiv.org/abs/2412.15115*.

YANG, C., ZHU, D. et WANG, X. (2017). SAO semantic information identification for text mining. *International Journal of Computational Intelligence Systems*, 10(1):593–604.

ZHANG, J., HAVERALS, W., NAYDAN, M. et KERNIGHAN, B. (2024). Post-ocr correction with openai's gpt models on challenging english prosody texts. *In DocEng 2024 - Proceedings of the 2024 ACM Symposium on Document Engineering*, DocEng 2024 - Proceedings of the 2024 ACM Symposium on Document Engineering. Association for Computing Machinery, Inc. Publisher Copyright : © 2024 Copyright held by the owner/author(s). ; 2024 ACM Symposium on Document Engineering, DocEng 2024 ; Conference date : 20-08-2024 Through 23-08-2024.

ZOUAOUA, D., CRUBLEAU, P., CHOULIER, D. et RICHIR, S. (2013). Application of evolution's laws. *In ETRIA's TRIZ Future Conference 2013*, Paris, France.

ZWEIGENBAUM, P. et GRABAR, N. (2002). Accentuation de mots inconnus : application au thesaurus biomédical MeSH. *In* PIERREL, J.-M., éditeur : *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 53–62, Nancy, France. ATALA.

ANNEXES

Annexe A

Guide d'annotation

Guide d'annotation des marqueurs des Fiches Anomalies

Mariame Maarouf

Avril 2023

Table des matières

1 Présentation

Ce guide d'annotation a pour objectif de présenter les consignes d'annotations de Fiches Anomalies (FA). Cette annotation cible l'expression d'un problème dans un corpus de documents de type REX (Retour d'EXpérience) dans un environnement technique et industriel.

Le corpus d'annotation ici présent est constitué de FA produites lors de phases d'exploitation de lanceurs Ariane 5. Pour chaque écart à l'attendu rencontré, même minime, une FA est produite par un opérateur.

La partie du corpus concernée par l'annotation est le champ "Description de l'anomalie". Dans celui-ci se trouve un constat du problème effectué sur le vif par un opérateur et n'implique généralement aucun diagnostic du problème. En revanche, nous pouvons observer des degrés de précision différents dans l'exposé du problème, et ces précisions permettent l'élaboration d'une catégorisation de ces descriptions d'un problème.

Les données ont été rédigées en majuscules et sans accents, et la ponctuation et la présence d'espaces n'est pas toujours respectée. Ces particularités peuvent en entacher leur compréhension et être source d'ambiguïtés. Dans la suite de ce document, on fera référence à une FA pour parler de l'énoncé qui est séparé des autres par un saut de ligne. Une FA peut être composée de plusieurs phrases et comporter plusieurs lignes (entre une et trois lignes en général).

2 Annotation en type et en marqueur

Une typologie d'expression d'un problème (cf. Table 2) a été produite à partir de l'étude de ce corpus de FA. Cette typologie est composée de neufs types identifiés. L'objectif de cette tâche d'annotation est de repérer, pour chaque FA, le ou les marqueurs d'expression d'un problème présents, et de leur assigner le type auquel il correspond. Ils peuvent être définis comme un marqueur, dans le texte, qui réfère, qui est propre à la catégorie de la typologie auquel il se rattache, comme présenté dans les exemples des parties ci-dessous. Cette typologie se veut dégressive. C'est-à-dire que plus le type est haut dans le tableau 2, plus le problème exprimé est spécifique et, généralement, facile à catégoriser. En revanche, plus le type est bas, plus le problème exprimé est générique et peu précis. Les types "Obstacle", "Dégradation" et "Élément manquant" sont des catégories similaires à "Hors Spécification" mais dont le type d'écart est plus spécifique, plus précis que pour ce dernier.

Index	Type	Étiquette
1	Fuite	Fuite
2	Signal qui s'est déclenché	Signal
3	Obstacle	Obstacle
4	Dégradation - Usure - Saleté	Dégradation
5	Élément manquant	Absence
6	Hors spécification	HorsSpec
7	Dispositif qui ne fonctionne pas	FonctionnePas
8	Action difficile ou impossible	Impossible
9	État du monde	EtatDuMonde

TABLE 1 – Types d'expression du problème

Afin de procéder à l'annotation, nous proposons la méthode suivante. Tout d'abord, commencer par la lecture de la FA dans son entièreté afin d'en vérifier sa compréhension. Si l'énoncé est jugé incompréhensible par l'annotateur, il est demandé d'identifier cette FA *via* un tag "Incompréhension" sur le premier token de l'énoncé. Si l'énoncé est compréhensible, repérer les marqueurs qui participent à son intelligibilité et au repérage du type dont il est question, puis procéder à l'annotation. Si l'énoncé est compréhensible, mais qu'aucun marqueur n'est percevable, deux cas sont possibles. Soit il correspond au type 9 "État du monde" (voir section 2.9), soit il contient uniquement des informations d'ordre contextuel. Pour les cas du type "État du monde", aucun marqueur ne peut être annoté, la procédure est d'étiqueter le premier terme de la FA avec l'étiquette "État du monde".

Les descriptions de ces problèmes sont des constats plutôt sommaires de la situation, parfois pauvre en information. Toutefois, afin de déterminer quel type correspond au marqueur, il est important de s'appuyer uniquement sur le contenu textuel de l'énoncé, de rester aussi proche que possible du segment à annoter et d'éviter toute inférence sur la situation décrite.

2.1 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"...

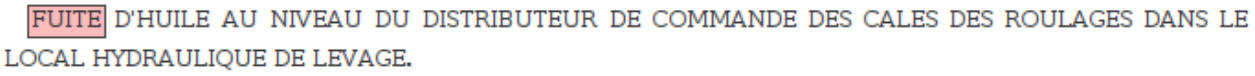


FIGURE 1 – Exemple d'annotation pour le type "Fuite"

2.2 Signal qui s'est déclenché

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"...

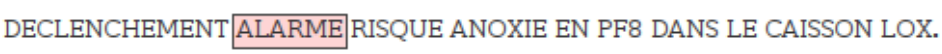


FIGURE 2 – Exemple d'annotation pour le type "Signal - alerte qui s'est déclenché"

2.3 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"...

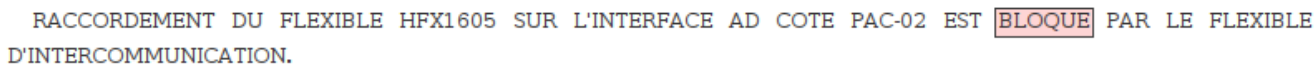


FIGURE 3 – Exemple d'annotation pour le type "Obstacle"

2.4 Dégradation - Usure - Saleté

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".

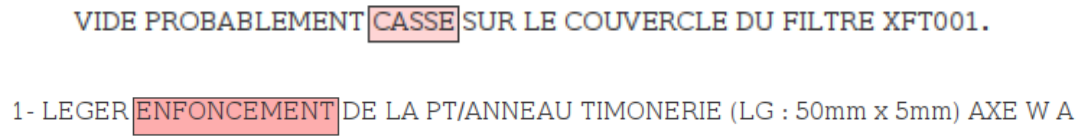


FIGURE 4 – Exemples d'annotation pour le type "Dégradation-Usure-Saleté"

2.5 Élément manquant

Objet, dispositif, document ou tout autre élément manquant. Les marqueurs seront de l'ordre de : "sans", "absence", "perte"...

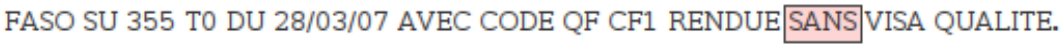


FIGURE 5 – Exemple d'annotation pour le type "Objet manquant"

2.6 Hors spécification

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"...

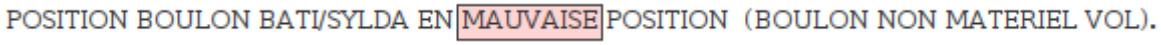


FIGURE 6 – Exemple d'annotation pour le type "Hors Spécification"

2.7 Dispositif qui ne fonctionne pas

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"...

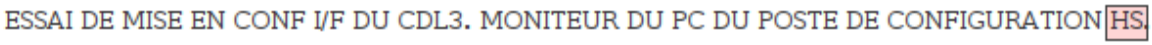
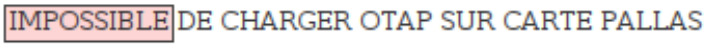


FIGURE 7 – Exemple d'annotation pour le type "Dispositif qui ne fonctionne pas"

2.8 Action difficile ou 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"...



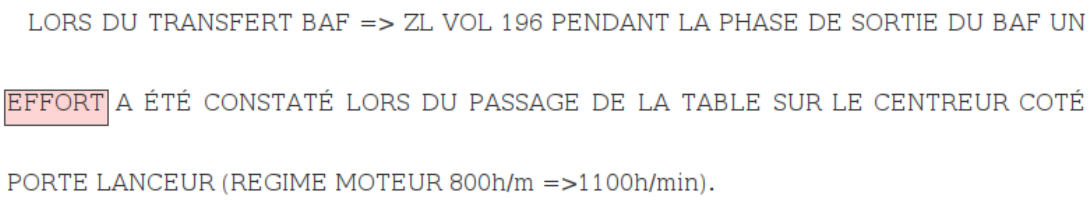


FIGURE 8 – Exemples d'annotation pour le type "Impossible d'accomplir une action"

2.9 État du monde

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".

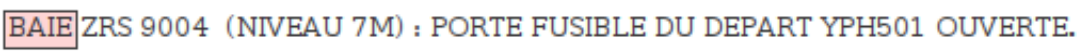


FIGURE 9 – Exemples d'annotation pour le type "État du monde"

2.10 Présence de multiples marqueurs

Dans une FA, plusieurs marqueurs du même, ou de types différents, peuvent être présents ; ils doivent tous être annotés. En revanche, un terme porteur de la notion de problème selon notre typologie n'est pas nécessairement un marqueur. Cette différenciation s'effectue à partir de la prise en compte du contexte d'apparition du mot. Ainsi, dans l'exemple Figure 10, nous pouvons remarquer la présence du terme "EXCEPTION" deux fois. Dans le premier cas, il est marqueur d'un problème et déclencheur du type "Signal qui s'est déclenché". Dans le second cas, il fait partie du message d'alarme et n'est pas à annoter (dans certains cas, la FA peut être constituée uniquement d'un message d'alarme, par défaut, le terme serait cette fois à annoter).

FIGURE 10 – Deux "EXCEPTION", un déclencheur

2.10.1 Plusieurs déclencheurs du même type

Plusieurs déclencheurs du même type peuvent être trouvés au sein d'une même FA. Dans cette situation, annoter toutes les occurrences de marqueur (cf. Figure 11) et ce, même si plusieurs marqueurs sont présents dans la même phrase.

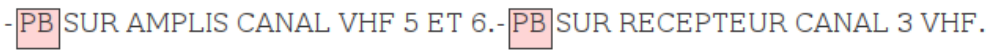


FIGURE 11 – Exemple d'annotation pour plusieurs marqueurs du même type dans une FA

2.10.2 Plusieurs déclencheurs de type différents

Des FA peuvent aussi posséder plusieurs énoncés qui appartiennent à des catégories différentes. Dans ces cas, il est important de se focaliser sur l'énoncé en question pour déterminer son type et non sur l'ensemble de la FA, qui contribue cependant à la compréhension générale du problème. Ainsi, dans l'exemple Figure 12, trois déclencheurs peuvent être repérés : "N'EST PAS ACCESSIBLE", "DEVANT", "EST EGALEMENT GENE". Les deux derniers renvoient à la catégorie "Obstacle", un élément ou une action est physiquement bloqué par un objet. En revanche, le premier déclencheur relève du type "Impossible d'accomplir une action". Il est important, pour typer ce marqueur, de se focaliser principalement sur la partie de l'énoncé qui est couverte par ce marqueur ; "DES TAPIS SONT STOCKES DEVANT LA PORTE" est un second énoncé dans lequel on peut trouver un second type de problème.

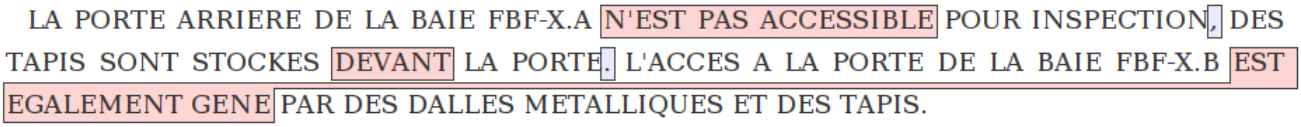


FIGURE 12 – Exemple d'annotation pour plusieurs marqueurs de types différents dans une FA

2.10.3 Déclencheurs disjoints

Dans certains cas, des marqueurs "disjoints" peuvent être présents dans le texte, afin d'être compris en tant que déclencheur dans l'énoncé, il est nécessaire que les deux éléments existent (cf. Figure 13). Pour ces occurrences, une relation de type "PartOf" est à ajouter entre les deux éléments annotés. De toute évidence, les deux éléments appartiennent au même type d'expression d'un problème ("Hors spécifications" dans le cas de cet exemple).

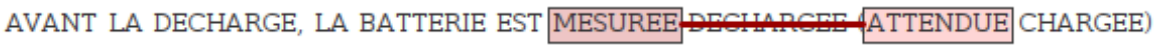


FIGURE 13 – Exemple d'annotation pour les marqueurs disjoints

2.11 Quelques cas d'ambiguïtés

Certains cas d'ambiguïtés peuvent se présenter. Lorsqu'une hésitation apparaît entre deux types, il est demandé de choisir celui qui est le plus précis (plus le type est bas dans le tableau 2, moins le type est précis). En revanche, il ne faut jamais inférer le type de problème, uniquement se baser sur ce qui est écrit dans le texte. Par exemple :

— "VENTILATEURS BRUYANTS SUR LES TIROIRS DE LA BAIE N° 9 DANS LA MFEG" : Cet énoncé pourrait soit appartenir au type "Dispositif qui ne fonctionne pas", soit à "Dégradation". "BRUYANT" n'est pas suffisamment précis et n'appartient pas au vocabulaire de la dégradation et l'absence d'échelle de temps ne peut laisser penser à un phénomène d'usure. Ainsi, on choisira de le catégoriser comme un dispositif qui ne fonctionne pas correctement.

— "DRUCK DPI610 N° 6280 SONDE INTERNE INSTABLE." : Ici, l'énoncé peut soit correspondre au type "Dispositif qui ne fonctionne pas", soit au type "Hors Spécification". Le marqueur "Instable" n'indique pas un empêchement à faire fonctionner la sonde et sans information en ce sens, la catégorisation de cet énoncé dépendra plutôt du type "Hors spécification", la sonde devrait être stable mais ne l'est pas.

— "ABSENCE DE PROCEDURE AU FORMAT ARIANESPACE POUR LE RETOUR EN SECURITE (PROCEDURE FORMAT CSG)" : Si le marqueur "ABSENCE" peut evidemment indiquer l'appartenance au type "Absence d'un élément", la présence de standards dans la FA ("FORMAT ARIANESPACE" par opposition à "FORMAT CSG") peut faire pencher en faveur du type "Hors spécification". Le problème est-il l'absence d'une procédure au format ARIANESPACE ou le fait que la procédure soit au format CSG ? En l'absence de marqueur propre à la notion de hors spécification et étant donné que le type "Absence" possède un degré de précision plus élevé, on choisira de catégoriser la FA comme étant de type "Absence".

3 Utilisation de Glozz

L'outil utilisé pour effectuer cette tâche d'annotation est Glozz [1]. Un manuel est à disposition en ligne si besoin [2].

1. http ://www.glozz.org/
2. http ://glozz.free.fr/glozzManual_1_0.pdf

3.1 Chargement des fichiers

Le document à annoter sera fourni en accompagnement de ce guide. Afin de charger le corpus dans l'outil, procéder comme suit :

1. Cliquer sur le bouton "Open corpus"
2. Charger le document corpus au format ".ac" + charger le document ".aa" fournis
3. Cliquer sur le bouton "load model"
4. Charger le document ".aam" fourni
5. Cliquer sur le bouton "style editor" et charger le document "styleSheet_expPb.as"

FIGURE 14 – Barre d'outils Glozz

3.2 Annotation

Afin de créer une annotation :

1. Cliquer sur le bouton "Create a new simple unit" / "Ajout d'une étiquette" (voir Figure 15)
2. Cliquer sur le type d'étiquette souhaité ("Déclencheur")
3. Sélectionner la zone de texte à étiqueter
4. Choisir le type correspondant dans le menu déroulant de la zone "Choix du marqueur"

Si besoin d'ajouter une relation entre deux éléments d'un marqueur disjoint :

1. Annoter les deux marqueurs concernés comme décrit ci-dessus
2. Cliquer sur le bouton "Edit relation" / "Ajout d'une relation"
3. Cliquer sur l'étiquette "PartOf"
4. Sélectionner les deux éléments à relier

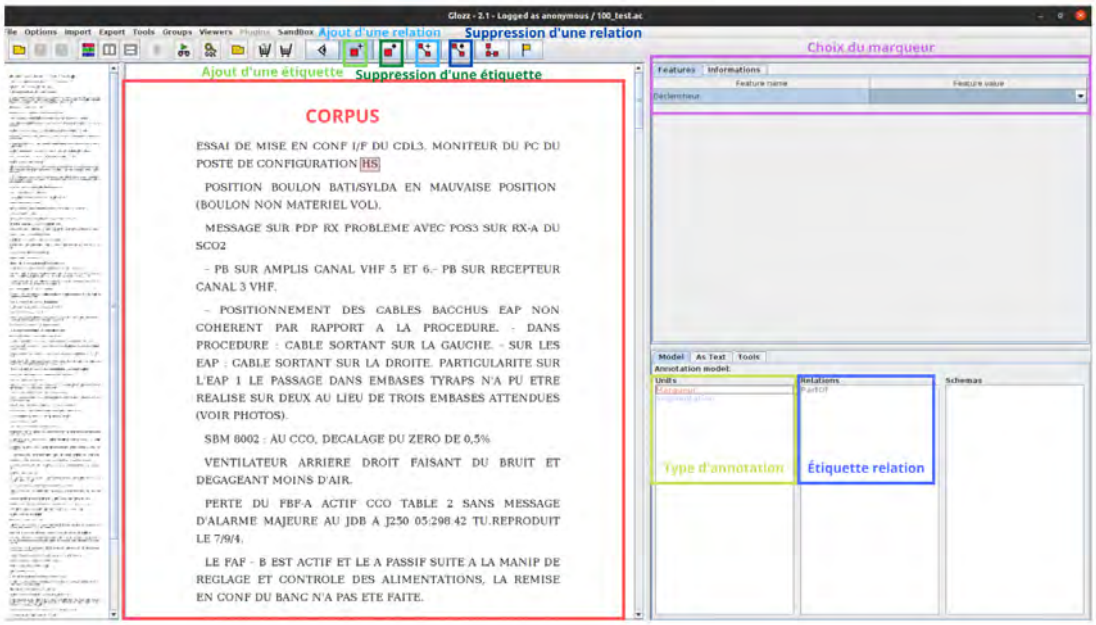


FIGURE 15 – Interface Glozz

Type	Étiquette
Fuite	Fuite
Signal qui s'est déclenché	Signal
Obstacle	Obstacle
Dégradation - Usure - Saleté	Dégradation
Élément manquant	Absence
Hors spécification	HorsSpec
Dispositif qui ne fonctionne pas	FonctionnePas
Impossibilité d'accomplir une action	Impossible
État du monde	EtatDuMonde

TABLE 2 – Correspondance types / étiquettes

Annexe B

Listes de marqueurs automatiquement extraits

Nous fournissons ici les listes des marqueurs issus de l'étiquetage automatique du modèles appris sur le corpus complet des descriptions d'anomalies. Pour circonvenir aux incohérences d'étiquetage, nous avons sélectionné uniquement les *subwords* dont les étiquettes BI qui se suivent sont de la même catégorie. En cas d'incohérence, le segment n'est donc pas inclut dans ces listes.

Le nombre de marqueurs extrait par catégories au total est :

— Fuite d'un liquide ou d'un gaz : 13 ;
— Signal qui s'est déclenché : 83 ;
— Présence d'un obstacle : 53 ;
— Dégradation - Usure - Saleté : 720 ;
— Élément manquant : 346 ;
— Configuration hors spécification : 2887 ;
— Dispositif qui ne fonctionne pas : 487.

Pour la catégorie *Configuration hors spécification*, nous retrouvons 2887 marqueurs. Étant donné ce nombre élevé, nous avons choisi d'inclure uniquement les 720 tokens les plus fréquents dans la liste ci-dessous (720 étant le nombre le plus élevé de marqueurs suivant, pour la catégorie *Dégradation - Usure - Saleté*). Nous n'incluons pas les marqueurs annotés pour la catégorie *État du monde* puisqu'ils ne sont pas liés à la catégorie du point de vue sémantique.

La catégorie *Fuite d'un liquide ou d'un gaz* est celle avec le moins de marqueurs (13), et ce sont tous des variations de *fuite, fuyard ou écoulement*. Nous retrouvons aussi quelques erreurs d'annotation avec *fusibles, fusées et fu*.

La catégorie *Présence d'un obstacle* est celle qui suit avec 53 marqueurs. Nous y retrouvons plusieurs variations de *bloque, empeche, gene, obture et obstrue*. Nous remarquons aussi quelques erreurs avec notamment *fuite, est, ping* ou encore *abscence*.

IL est intéressant de noter la grande variété des marqueurs de la catégorie *Dégradation - Usure - Saleté* qui couvrent les différentes façon dont un composant peut être dégradé.

Pour la catégorie *Élément manquant*, la liste semble montrer un ensemble de sous-listes avec de nombreuses variations autour d'un lemme pour en exprimer la négation. Il en va de même pour la catégorie *Dispositif qui ne fonctionne pas* avec en plus des variations autour du mot « problème » et de « perte de ».

Action difficile ou impossible

- accessibilite
- accessible
- accessibles
- aucun
- aucune
- aucune possibilite
- aurait pu
- avec possibilite
- beaucoup de difficulte
- bilite
- dangereux
- de possibilite
- difficile
- difficile , voire imlpossible
- difficilement
- difficilement accessible
- difficilement realisable
- difficiles
- difficulte
- difficultes
- est
- est difficile
- est difficilement
- extrememement
- grande
- grandes difficultes
- grosse
- grosse difficulte
- il
- illisi
- illisibilite
- illisible
- illisibles
- illsible
- imposible
- impossbilite
- impossib
- impossibibilite
- impossibiite
- impossibil
- impossibilite
- impossibilte
- impossible
- impossibles
- impossiblite
- impossilite
- incapable
- indemontable
- indemontables
- indispensable
- interdit
- introuvable
- introuvables
- inutilisable
- irrealisable
- isolable
- l '
- mpossibilite
- n
- n ' a pa pu
- n ' a pas ete possible
- n ' a pas permis
- n ' a pas pu
- n ' a pas pu etre
- n ' a pu
- n ' a pu etre
- n ' arrive pas
- n ' arrive plus
- n ' avons pas pu
- n ' ayant pu
- n ' est accessible qu
- n ' est pas accessible
- n ' est pas capable
- n ' est pas possible
- n ' est pas realisable
- n ' est plus accessible
- n ' est plus possible
- n ' est possible
- n ' est possible non plus
- n ' est realisable
- n ' etaient pas possibles
- n ' etait pas possible
- n ' ont pas
- n ' ont pas permis
- n ' ont pas pu
- n ' ont pas pu etre
- n ' ont pu
- n ' ont pu etre
- ne parvient pas
- ne passe pas
- ne permet
- ne permet pas
- ne permet plus
- ne permettait pas
- ne permettant
- ne permettant pas
- ne permettent pas
- ne permettra pas
- ne peut
- ne peut pas
- ne peut plus
- ne peuvent
- ne pouvait pas
- ne savent plus
- ne serait utilisable
- ne sont pas

- ne sont pas accessibles
- ne sont plus accessibles
- no impossible
- non accessible
- non accessibles
- non imputable
- non possibilite
- non possible
- non realisable
- non realisables
- non utilisable
- obligation
- on
- pas
- pas d ' accessibilite
- pas de
- pas de possibilite
- pas moyen
- pas possible
- peu utilisable
- plus
- plus de possibilite
- plus difficile
- possibilite
- possible
- probable
- realisables
- refuse
- rien
- risque
- sans possibilite
- sont difficiles
- susceptible
- susceptibles
- tres difficile
- tres difficilement accessibles
- tres difficiles
- trop facile
- uniquement accessible

Présence d'un obstacle

- abscence
- battante
- blocage
- bloque
- bloque derriere
- bloque en
- bloque par
- bloquee
- bloquee par
- bloquees
- bloquees par
- bloquent
- bloquer
- bloques
- bloquesle
- bouche
- bouchee
- cache
- cache derriere
- coince
- coincee
- disparition
- empeche
- empeche d
- empeche de
- empechent
- empechent de
- est
- fermeture
- fuite
- gene
- genent
- morts
- ne permet pas d ' avoir des mouvements
- ob
- obstruant
- obstruction
- obstrue
- obtura
- obturant
- obturateurs
- obturation
- obture
- obturee
- obtures
- occasion
- perdent
- perimee
- ping
- suppression
- tombes
- touche
- vides

Élément manquant

- a indisponible
- a pas de
- abandon
- absence
- absence d
- absence de
- absent
- absent de
- absente
- absentes
- absents
- acquisition
- aucun
- aucun defaut
- aucune
- beaucoup de
- besoin
- bien present
- c ' est trouvee posee
- d ' absence
- d ' absence d
- d ' absence de
- d ' acquisition
- defaut d ' acquisition
- disparition
- disponible
- disponibles
- en
- en absence
- en absence de
- en acquisition
- en defaut
- en defaut pas d
- en defaut plus de
- en l ' absence
- en l ' absence de
- enleve
- enlevee
- enlevees
- enlevement
- enlever
- enleves
- erreur d ' acquisition
- est
- et pas de
- existence
- existent
- incomplete
- indi
- indisponibilite
- indisponibilites
- indisponible
- indisponibles
- inexis
- inexistance
- inexistant
- inexistante
- inexistantes
- inexistants
- jeu
- l
- l ' absence
- l ' absence de
- l ' acquisition
- lack of
- lost
- mais
- manqua
- manquait
- manquant
- manquante
- manquantes
- manquants
- manque
- manque d
- manque de
- manquent
- manquent de
- manques
- minimum d ' acquisition
- moins d
- n
- n '
- n ' a
- n ' a aucun
- n ' a aucune
- n ' a aucune donnee disponible
- n ' a ete
- n ' a eut
- n ' a fournit aucun
- n ' a pas
- n ' a pas d
- n ' a pas de
- n ' a pas genere de
- n ' a pas prise
- n ' a pas recu
- n ' a pas recu d
- n ' a plus aucun
- n ' a plus d
- n ' a plus de
- n ' a rien
- n ' affiche aucun
- n ' affiche pas
- n ' affiche plus de
- n ' affiche plus rien
- n ' affiche rien
- n ' apparaissent pas
- n ' apparait
- n ' apparait pas
- n ' apparait plus
- n ' atteint
- n ' avait pas
- n ' avait pas de
- n ' avait trouve aucun
- n ' avions pas
- n ' avons pas eu
- n ' ayant pas

— n ’ ayant pas d
— n ’ ayant pas de
— n ’ effectue aucune
— n ’ es pas audible
— n ’ est
— n ’ est affiche
— n ’ est pas
— n ’ est pas acquis
— n ’ est pas c
— opie
— n ’ est pas cree
— n ’ est pas defini
— n ’ est pas disponible
— n ’ est pas en
— n ’ est pas en acquisition
— n ’ est pas executable
— n ’ est pas ide
— n ’ est pas pose
— n ’ est pas present
— n ’ est pas presente
— n ’ est pas recue
— n ’ est pas reparable
— n ’ est pas visible
— n ’ est pas vu
— n ’ est pas vue
— n ’ est plus
— n ’ est plus disponible
— n ’ est plus en
— n ’ est plus visible
— n ’ est presente
— n ’ est prevu
— n ’ etaient pas presents
— n ’ etait pas disponible
— n ’ etait pas present
— n ’ etait pas presente
— n ’ etait plus disponible
— n ’ etant disponible
— n ’ existait pas
— n ’ existe
— n ’ existe aucun
— n ’ existe par
— n ’ existe pas
— n ’ existe pas assez de
— n ’ existe pas de
— n ’ existe pas encore
— n ’ existe plus
— n ’ existe plus de
— n ’ existent pas
— n ’ existent plus
— n ’ indique
— n ’ indique aucune
— n ’ indique plus
— n ’ indique plus de
— n ’ ont
— n ’ ont aucun
— n ’ ont pas
— n ’ ont pas d
— n ’ ont pas de
— n ’ ont pas encore
— n ’ ont pas ete posees
— n ’ ont plus aucun
— n ’ ont plus d ’ image
— n ’ ont plus de
— n ’ ont plus de vue
— n ’ ont recu
— n ’ utilise pas de
— n ’ y a
— n ’ y a aucun
— n ’ y a aucune
— n ’ y a pas
— n ’ y a pas assez de
— n ’ y a pas d
— n ’ y a pas de
— n ’ y a pas des
— n ’ y a pas ete
— n ’ y a pas eu
— n ’ y a pas eu d
— n ’ y a pas eu de
— n ’ y a pas ni
— n ’ y a plus
— n ’ y a plus aucune
— n ’ y a plus d
— n ’ y a plus de
— n ’ y a plus rien
— n ’ y a qu
— n ’ y a que
— n ’ y a rien
— n ’ y ai pas d
— n ’ y avait pas
— n ’ y avait pas d
— n ’ y avait pas de
— n ’ y avait plus d
— n ’ y avait plus de
— n ’ y en a
— n ’ y en a pas
— n ’ y en avait qu
— n ’ y pas de
— n affiche
— ne
— ne comporte aucun

— ne comporte pas d
— ne comporte pas de
— ne constatent
— ne contiennent
— ne contient aucun
— ne contient pas
— ne convient pas d
— ne demande pas de
— ne demandent pas de
— ne dispose pas
— ne disposent pas de
— ne donne
— ne donne aucun
— ne donne aucune
— ne donne pas
— ne donne pas de
— ne en defaut
— ne fait pas
— ne fournit pas de
— ne fournit plus d
— ne genere pas de
— ne generent pas d
— ne laisse aucun
— ne mentionne
— ne montre pas
— ne montrent aucun
— ne possedait pas
— ne possedait pas de
— ne possede pas
— ne possede pas de
— ne possede que
— ne possedent pas d
— ne possedent pas de
— ne pouvons pas afficher d
— ne presentait aucune
— ne presentait pas de
— ne presente pas
— ne presente pas d
— ne presente pas de
— ne prevoit aucune
— ne recevait aucun
— ne recoit aucun
— ne recoit pas de
— ne recoivent aucune
— ne revele aucune
— ne s ' affiche
— ne s ' affiche pas
— ne semble pas present
— ne soit presente
— ne sont pas disponibles
— ne sont pas presentes
— ne sont pas presents
— ne sont pas recues
— ne sont plus presents
— ne st pas active
— ne voit aucun
— ni de
— nombreuses
— non
— non - acquisition
— non acquisition
— non apparition
— non d ' acquisition
— non detection
— non disponible
— non en place
— non ob
— non passage
— non perte
— non presence
— non present
— non presents
— non reception
— non recu
— non recue
— non recues
— non retrouves
— non suppression
— non utilise
— non visible
— pas
— pas assez
— pas assez de
— pas d
— pas d '
— pas d ' acquisition
— pas d ' apparition
— pas d ' image
— pas de
— pas de defaut
— pas de mention

- pas de perte
- pas de reponse
- pas disponible
- pas present
- pb d ' acquisition
- perdu
- perte
- perte acquisition
- perte d
- perte d ' accessibilite
- perte d ' acquisition
- perte d ' image
- perte de
- perte de l ' acquisition
- perte du
- pertes
- plus
- plus aucun
- plus d
- plus d ' image
- plus de
- plus rien
- point d ' acquisition
- retrait
- rien
- rien n ' apparait
- sans
- sans aucune
- seule
- seulement
- seules
- sont
- sont absents
- trop
- trop de
- une absence
- vide
- y a

Dégradation - Usure - Saleté

- abaisse
- abaissee
- abime
- abimee
- abimees
- abimer
- abimes
- aborte
- abscence
- accorvitas
- accrocha
- accrochage
- accroche
- accrochee
- accrocs
- acorvitas
- adeki
- adosse
- affaiblie
- affaiblissement
- affaisement
- affaisse
- affaissee
- affaissees
- affaissement
- affaissements
- affaisser
- affaisses
- affouillement
- affouillements
- affranchi
- alimente
- alimentes
- allongement
- allume
- allumee
- allumees
- allumes
- allure
- amalgamee
- amelio
- amovi
- amovible
- amovibles
- ancienne
- anciens
- anticorosion
- ap
- aretes
- arpyat
- arrachage
- arrache
- arrache de
- arrachee
- arrachees
- arrachement
- arrachements
- arracher
- arraches
- arraches de
- asprie
- asshe
- attaquee
- attenuation
- baisse
- baissee
- bandes noires
- beepe
- beeper
- bipee
- blanchatre
- blanchatres
- blanches
- blesse
- blesses
- blessure
- blindage
- bouchee
- bouchees
- bouge
- boulonnes
- bourrage
- branlant
- branlants
- brise
- brisee
- brouilles
- bru
- bruite
- bruitee
- bruites
- brulante
- brule
- brulee
- brulees
- brules
- brulle
- brulures
- brusque
- bruyant
- bruyante
- bruyants
- bulles
- cadenassee
- cass
- cassage
- cassant
- cassants
- casse
- cassede
- cassee
- cassees
- cassent
- casser
- casses
- cassure
- cassures
- change
- changee
- changement
- changer
- changes
- chap
- chevauchement
- choc
- chocs
- chute
- cintree
- cisaille
- cisaillement
- clampage
- claquage
- claque
- clignote
- coiffe
- coincee
- collision
- coloration
- comprime
- configurees

— consumme
— contamines
— corrige
— corrode
— corrodee
— corrodees
— corrodes
— corrompu
— corrompue
— corrompus
— corrosion
— couchee
— coule
— coulissants
— coulisse
— coulure
— coulures
— coupe
— coupee
— coupees
— coupes
— crapauds
— craquelee
— craqueles
— crep
— creve
— crique
— croise
— croisee
— croisees
— croisement
— croises
— crottes
— croute
— debranchees
— debroche
— dechire
— dechiree
— dechirees
— dechirement
— dechires
— dechirure
— dechirures
— declenche
— declipsees
— deco
— decoince
— decolle
— decollee
— decollement
— decolles
— decompose
— decon
— deconfiguration
— deconfigure
— deconne
— deconnecte
— deconnectee
— deconnectees
— deconnectes
— deconnexion
— decoupe
— decoupee
— decoupes
— decousues
— decrochage
— decroche
— decrochee
— def
— defaite
— deformation
— deformations
— deformations importantes
— deforme
— deformee
— deformees
— deforment
— deformer
— deformes
— deg
— degagem
— degager
— degainee
— degommage
— degondee
— degonfle
— degonflee
— degradation
— degradations
— degrade
— degradee
— degradees
— degrader
— degrades
— degrafe
— degrafee
— degrafees
— degrafes
— degraisse
— degraphe
— degration
— delamination
— delove
— deluge
— demanteles
— demon
— demontee
— demontees
— den
— denude
— denudes
— deplacee
— deplacees
— deplo
— deplombe
— deplombee
— deplombees
— deplombes
— deraccorde
— deraccordee
— deraccordes
— deracine
— deraillement
— derangement
— derangements
— des
— desa
— desactivation
— desafleurement
— desagregation
— desagrege
— desamorcage
— desamorce
— desamorcee
— desarmee
— desaxes
— descellee
— dese
— deshydratee
— desintegration
— desintegres
— desinter
— desobtures
— desolidari
— desolidarisation
— desolidarise
— desolidarisee

- desolidarisees
- desolidarisent
- desolidarises
- desordonne
- desoude
- dessere
- desseree
- desserees
- desseres
- desserra
- desserrage
- desserrages
- desserres
- desserti
- dessertie
- dessertis
- dessolidarisee
- dessoude
- dessoudee
- dessoudes
- destabilisee
- destraces
- destruction
- det
- detache
- detachee
- detaches
- deterioration
- deteriore
- deterioree
- deteriorees
- deteriorer
- deteriores
- deterioresur
- deterrio
- deterrioree
- detorsadee
- detruit
- detruite
- detruites
- detruits
- deve
- devisse
- devissee
- devorer
- diminue
- diminuer
- diminution
- discontinu
- discontinuites
- discordance
- discordante
- disfonctionnement
- dissociation
- distordue
- distordues
- distorsion
- ecaillee
- echappe
- eclat
- eclate
- eclatee
- eclatees
- eclatement
- eclisse
- ecrase
- ecrasee
- ecrasement
- ecrases
- ecrassement
- efface
- effacee
- effacees
- effaces
- effiloche
- effondrees
- effondrement
- effrite
- effritee
- egradation
- ejecte
- ejectee
- ejectes
- ejection
- en torsion
- enclenche
- enclipse
- endom
- endomage
- endomagees
- endommage
- endommagee
- endommagees
- endommagement
- endommager
- endommages
- enfiche
- enfon
- enfonce
- enfoncee
- enfoncees
- enfoncem
- enfoncement
- enfoncements
- enfoncer
- entaille
- entame
- epaisseur
- eraflee
- eraflure
- eraflures
- ero
- erode
- erodes
- eronee
- erosion
- errone
- erronee
- erronees
- erronnee
- etalonnage
- etalonne
- etalonnes
- etanche
- etanchees
- etancheite
- etoupe
- evacue
- eventree
- evolution
- exceeds
- explose
- exploses
- explosion
- fatigue
- fen
- fendu
- fendue
- fendues
- fendus
- fissur
- fissuration
- fissurationde
- fissurations
- fissure
- fissuree
- fissurees
- fissures

— fissures importantes
— foiree
— foirees
— fondu
— fondue
— fondues
— fondus
— foudre
— fourniture
— fracturee
— fragilisation
— freinee
— froisse
— frotte
— frottement
— fuite
— fume
— fumee
— fur
— furtif
— gondolee
— gonflage
— gresillements
— grille
— grillee
— grillees
— grilles
— grip
— gripee
— grippage
— grippe
— grippee
— grippees
— grippes
— grosse
— grosse degradation
— hausse
— hauts
— hernie
— heurt
— heurte
— humide
— ignores
— illumine
— impact
— impacte
— impacts
— importantes
— incertitude
— incessant
— incruste
— inerte
— infiltration
— inondation
— inonde
— inondee
— interrompu
— interruption
— inversion
— irradie
— irreelles
— irrisees
— la degradation
— la rupture
— lache
— legerement
— lite
— lourde
— malfonction
— marque
— masse
— meurtrissures
— microfuite
— migre
— modification
— modifie
— modifiee
— morsures
— netto
— nettoye
— nettoyees
— noircis
— non etanche
— non etancheite
— obstruction
— obstrue
— obstruee
— obstruees
— obtruee
— ondules
— otees
— outillage
— oxy
— oxydation
— oxydations
— oxyde
— oxydee
— oxydees
— oxydes
— oxygne
— pen
— penchee
— pend
— pendu
— pendule
— pendulent
— penduler
— pendules
— perce
— percee
— percees
— perces
— percut
— percute
— percutee
— percutees
— percuter
— perdu
— perdue
— perforation
— perfore
— peri
— perime
— pertubation
— perturbation
— perturbations
— perturbe
— perturbee
— perturber
— perturbera
— piegee
— pince
— pincee
— pinces
— piquee
— pl
— plantage
— plie
— pliee
— pliees
— plies
— pliure
— pliures

- pliures importantes
- pollue
- polluees
- pollution
- pollutions
- positionne
- positionnes
- pourries
- projection
- rabattue
- rabattus
- raccorde
- raccordee
- rainure
- ralent
- rapprochee
- ratissee
- ravinements
- rayee
- rayes profondement
- rayure
- rayures
- rebondit
- rebouche
- recouverts
- recupere
- referme
- refoulement
- refroidissement
- relachee
- remplacement
- rempli
- remplie
- renforce
- reouverte
- repetes
- retablie
- retardee
- revoir
- rom
- rompu
- rompue
- rompus
- ronges
- roses
- rotules
- rouille
- rouillee
- rouillees
- rouilles
- roule
- roulee
- roulement
- roulements
- rouler
- rousses
- rupture
- rupture de
- s ' est
- saccade
- satu
- saturait
- saturation
- saturee
- saute
- sautillement
- scelles
- scotche
- scotchee
- scotchees
- se
- se croise
- se dechirent
- se deforme
- se grippe
- seches
- secourue
- sectionne
- sectionnee
- sectionnees
- sectionnes
- securicides
- separation
- shunte
- shuntes
- sombre
- sombres
- souffle
- soufflee
- souilles
- souleve
- soulevee
- soulevement
- spirale
- stoppe
- stoppee
- suintement
- suiveuse
- suppression
- suppri
- supprimes
- surplombant
- surveille
- suspendu
- tarees
- toiture
- tombe
- tombee
- tordu
- tordue
- tordues
- tordus
- torsade
- torsadee
- torsades
- torsion
- touches
- tourne
- trace
- trace d
- trace d ' oxydation
- trace de
- trace de choc
- trace de corrosion
- trace de rouille
- trappe
- trappes
- traversantes
- tremblante
- tremblantes
- tremblements
- trodu
- tronquee
- trouee
- tyrap
- tyrape
- verdatre
- verdatres
- vieillissement
- vrille
- vrillee
- zesfuite

Fuite d'un liquide ou d'un gaz

- ecoulement
- fu
- fuient
- fuit
- fuite
- fuite de
- fuites
- fusees
- fusibles
- fuyard
- fuyarde
- fuyardes
- fuyards

Signal qui s'est déclenché

- acquisition
- activation
- affiche
- alarm
- alarme
- alarme majeure
- alarmes
- alerte
- allume
- apparition
- apparition d ' erreurs
- appel
- attention
- bas
- basculement
- basculements
- bru
- brusque
- bruyant
- cables
- choc
- chocs
- cl
- claquement
- clignotement
- confusion
- consommation
- d ' erreur
- d ' erreurs
- deb
- dec
- declenche
- declenchee
- declenchement
- declenches
- deplacement
- descente
- desequilibre
- destabilisation
- detection
- disco
- ech
- echauffement
- echo
- ecoulement
- epaulement
- errones
- flous
- inc
- inf
- inst
- investigation
- jeu
- l
- l ' alarme
- l ' alarme majeure
- l ' erreur
- message
- message d '
- message d ' erreur
- messages
- messages d ' erreur
- messages d ' erreurs
- mouvement
- mouvements
- non
- nuisant
- parasit
- penurie
- pollution
- remplacement
- risque d ' erreur
- risques d ' erreurs
- ruisselement
- s
- s ' est
- s ' est declenchee
- se declenche
- sifflement
- source d ' erreurs
- suppression
- traces de chocs
- voyant

Dispositif qui ne fonctionne pas

- - arret
- a
- a disparu
- a panne
- a plante
- absence de fonctionnement
- arret
- arret fonctionnement
- arrete
- bloque
- bouche
- bruyant
- collantes
- couche
- court
- court -
- court - circuit
- crame
- def
- defailla
- defaillance
- defaillance fonctionnement
- defaillant
- defaillante
- defaillants
- defaut
- defaut de fonctionnement
- defaut fonctionnement
- defect
- defecteux
- defectueu
- defectueuese
- defectueuse
- defectueuses
- defectueux
- defectueuxada
- deffectueuse
- deffectueux
- defilement
- defixe
- demarre
- desactionne
- desactive
- disparu
- disparues
- disparus
- dyffonctionnement
- dysfoctionnement
- dysfonctionnee
- dysfonctionnement
- dysfonctionnment
- dysjonction
- echoue
- echoue egalement
- en
- en court -
- en court - circuit
- en defaut
- en derangement
- en echec
- en erreur
- en fail
- en fin de
- en halte
- en mauvais etat
- en panne
- en ruine
- en service
- erreur
- errones
- eteint
- failed error
- failure
- fatal error
- fige
- figee
- figes
- fonctionne
- fonctionne aleatoirement
- fonctionnel
- fonctionnelle
- fonctionnelles
- fonctionnement
- fonctionnement aleatoire
- fonctionnement correct
- h
- h - s
- h .
- h . s
- h . s -
- h . s .
- h . t
- hors
- hors - services
- hors activation
- hors d ' usage
- hors fonction
- hors service
- hors service .
- hors services
- hors utilisation
- hs
- hs +
- hs -
- hs - >
- hs :
- hs a
- hs en
- hs ne descent pas
- hsplus
- ht
- ht .
- image
- inact
- inactif
- inactifs
- inactive
- inactivite
- indisponible
- inexploitable
- inexploitables
- inoperant
- instable

— interdiction
— inutilisable
— invalide
— l
— l ’
— l ’ arret
— l ’ etat faute
— l . s
— mal
— manque de sensibilite
— mauvais fonction
— mauvais fonctionnement
— mauvais usage
— n
— n ’
— n ’ a pas demarre
— n ’ a pas fonctionne
— n ’ a pas fonctionnee
— n ’ acquiert pas
— n ’ active
— n ’ affiche plus
— n ’ ali
— n ’ alimente plus
— n ’ apparait jamais
— n ’ apparait pas
— n ’ arrete pas
— n ’ arrive pas
— n ’ assurent
— n ’ assurent plus
— n ’ effectue pas
— n ’ enregistre pas
— n ’ est pas
— n ’ est pas actif
— n ’ est pas actionne
— n ’ est pas active
— n ’ est pas deroule
— n ’ est pas devenu actif
— n ’ est pas efficace
— n ’ est pas ejectee
— n ’ est pas fonctionnel
— n ’ est pas operationnel
— n ’ est pas passe actif
— n ’ est pas plante
— n ’ est pas realise
— n ’ est pas utilisee
— n ’ est plus fonctionnel
— n ’ est plus fonctionnelle
— n ’ est plus manoeuvrable
— n ’ est plus operation
— n ’ etait actif
— n ’ evolue pas
— n ’ evoluent pas
— n ’ existe pas
— n ’ existe plus
— n ’ imprime
— n ’ imprime pas
— n ’ imprime pas .
— n ’ imprime plus
— n ’ imprime plus .
— n ’ imprime plus rien
— n ’ ont pas fonctionne
— n ’ ouvre pas
— n . m
— n . s
— ne
— ne ’ ouvre
— ne bascule pas
— ne boite pas
— ne boot pas
— ne boot plus
— ne boote pas
— ne bouge pas
— ne chauffe pas
— ne clignote pas
— ne communique pas
— ne communique plus
— ne communiquent plus
— ne compile plus
— ne compte plus
— ne connait pas
— ne coulent pas
— ne cree pas
— ne date plus
— ne debit
— ne debite pas
— ne debitent pas
— ne demarre
— ne demarre pas
— ne demarre pas seule
— ne demarre plus
— ne demarrent plus
— ne demarreront pas
— ne dialogue plus
— ne disparait pas
— ne ferme pas
— ne ferme plus

— ne fonction
— ne fonctione
— ne fonctionnait pas
— ne fonctionnait plus
— ne fonctionnant pas
— ne fonctionne
— ne fonctionne pas
— ne fonctionne pas bien
— ne fonctionne pas normale
— ne fonctionne plus
— ne fonctionne plus que
— ne fonctionne que
— ne fonctionne toujours pas
— ne fonctionnement pas
— ne fonctionnement plus
— ne fonctionnenet pas
— ne fonctionnenet plus
— ne fonctionnent
— ne fonctionnent pas
— ne fonctionnent pas bien
— ne fonctionnent plus
— ne fontionne pas
— ne fournit pas
— ne genere pas
— ne lance pas
— ne le retrouve jamais disponible
— ne maintien
— ne manoeuvre pas
— ne manoeuvre plus
— ne manoeuvrent pas
— ne marcha
— ne marche pas
— ne marche plus
— ne marchent pas
— ne marchent plus
— ne monte pas
— ne pas
— ne passe pas
— ne peut
— ne pivote pas
— ne prend pas
— ne presente pas
— ne protege plus
— ne protegent plus
— ne re
— ne reagissent
— ne reagit pas
— ne reagit plus
— ne reconnait pas
— ne reconnait plus
— ne remonte pas
— ne rentre pas
— ne repart pas
— ne repond pas
— ne repond plus
— ne repondent pas
— ne reprend pas
— ne reste pas
— ne reste plus
— ne retourne pas
— ne retrouve pas
— ne revient pas
— ne s
— ne s ’
— ne s ’ abaissent plus
— ne s ’ active pas
— ne s ’ affiche pas
— ne s ’ arrete pas
— ne s ’ arrete plus
— ne s ’ effectue pas
— ne s ’ effectue plus
— ne s ’ engage
— ne s ’ entend
— ne s ’ est pas arrete
— ne s ’ est pas ouverte
— ne s ’ est pas replie
— ne s ’ est plus produit
— ne s ’ etait pas lance
— ne s ’ execute pas
— ne s ’ extrait
— ne s ’ imprime pas
— ne s ’ initialise pas
— ne s ’ ouvre pas
— ne s ’ ouvre pas completement
— ne s ’ ouvre pas totalement .
— ne s ’ ouvre plus

- ne se
- ne se charge pas
- ne se connecte pas
- ne se connecte plus
- ne se declenche
- ne se deroule pas
- ne se desserre plus
- ne se deve
- ne se fait pas
- ne se fait plus
- ne se ferme pas
- ne se ferme pas .
- ne se ferme pas completement
- ne se ferme plus
- ne se ferment plus
- ne se laisse pas
- ne se lance pas
- ne se leve pas
- ne se leve plus
- ne se met pas
- ne se met pas en marche
- ne se met plus
- ne se met plus en charge
- ne se mettent pas
- ne se montait pas
- ne se monte
- ne se monte pas
- ne se montent pas
- ne se positionne pas
- ne se produit pas
- ne se rallume pas
- ne se rallume plus
- ne se reenclenche plus
- ne se referme pas
- ne se referme pas .
- ne se referme seulement
- ne se rep
- ne se replie pas
- ne se retrouve pas
- ne se soit pas ouverte
- ne se stabilise pas
- ne se synchronise plus
- ne se synchronisent plus
- ne se visse
- ne se visse pas
- ne se visse pas entierement
- ne se vissent pas
- ne semble pas fonctionner
- ne signale pas
- ne sonne pas
- ne sont pas fonctionnels
- ne sont pas receptionnees
- ne sont pas visibles
- ne sont plus operationnels
- ne sort pas
- ne sortant pas
- ne sortent pas
- ne tiennent plus
- ne tient
- ne tient pas
- ne touche pas
- ne tour
- ne tourne pas
- ne tourne plus
- ne tournent pas
- ne tournent plus
- ne traversent pas
- ne voient pas
- ne voient plus
- ne voit
- ne voit jamais
- ne voit pas
- ne voit plus
- ne voit toujours pas
- non
- non - arret
- non actif
- non activation
- non active
- non apparition
- non arret
- non chargement
- non consommation
- non demarre
- non demarree
- non deroule
- non deroulee
- non deroules
- non detection
- non diffusion
- non disparition
- non en place
- non evolution
- non execute
- non existante
- non fonction

— non fonctionnel
— non fonctionnelle
— non fonctionnels
— non fonctionnement
— non fournie
— non mise en service
— non operationnel
— non operationnelle
— non operationnelles
— non ouverture
— non reactive
— non realisation
— non reception
— non receptionne
— non receptionnee
— non recuperation
— non retour
— non retrait
— non rotation
— non sortie
— non transmission
— non utilisation
— non utilise
— non visible
— non vue
— ont disparu
— ouvert
— ouverte
— pan
— panne
— panne tache
— pannes
— pas
— pas de
— pas de fonctionnement
— pas de reception
— pas de reponse
— pas en fonctionnement
— pb
— pb de fonctionnement
— persiste
— perte
— perte actif
— perte de la sante
— perte de sensibilite
— perte sante
— perte sensibilite
— plante
— plus de fonctionnement
— probleme
— retire
— rien ne se passe
— s ’
— s ’ arrete
— s ’ arrete totalement
— s ’ arretent
— s ’ est active
— s ’ est arrete
— s ’ est detache
— s ’ est plante
— s ’ est plantee
— s ’ etait plante
— s ’ eteint
— s ’ ouvre
— sans
— sans effet
— sans fonction
— satu
— saturee
— se
— se bloque
— se bloquent
— se bloquerait
— se defait
— se fait entendre
— se fige
— se leve
— se met
— se met en halte
— se met en marche
— se plante
— se replie
— se touchent
— sensibilite
— trouble
— use

Configuration hors spécification

- 100
- <
- a
- a anomalie
- a cause
- a declenche
- a defaut
- a faute
- a l ' envers
- a plus de
- a revoir
- aberrantes
- abimees
- absolus
- acquise a
- allumage
- allume
- allumee
- alors qu
- alors qu '
- alors qu ' attendue
- alors qu ' elles
- alors qu ' on attend
- alors que
- anomali
- anomalie
- anomalies
- anormal
- anormale
- anormalement
- anormalement elevee
- anormalement long
- anormales
- anormaux
- arret
- atteinte
- attend
- attendu
- attendu a
- attendu allume
- attendue
- attendue a
- attendue allumee
- attendues
- attendues ouverte
- attendus
- attendus a
- atypique
- atypiques
- au
- au lieu
- au lieu d
- au lieu d '
- au lieu d ' etre
- au lieu de
- au lieu de la
- au lieu des
- au lieu du
- au profit
- augmentation
- avant
- bagotements
- bagotte
- bagottement
- bagottements
- bague
- baisse
- besoin
- bien que
- blessure
- bloque
- choc
- chute
- comme demande
- commute
- comportement
- comportement atypique
- compte tenu du
- conformement
- contrainte
- contraintes importantes
- contrairement
- contrairement a
- contrairement aux
- contre -
- correct
- correcte
- correctement
- coude
- critere
- d ' indisponibilite
- dangereuse
- dangereux
- de
- dec
- decala
- decalage
- decale
- decale a
- decale de
- decale par rapport
- decalee
- decalee par rapport
- decalees
- decales
- decales par rapport
- decallage
- decalle
- decharge
- dechargee
- declenche
- decolle
- decollee
- decollees
- decolles
- decrochage
- defa
- defaits
- defaut
- defauts
- defectueux

- deja
- delta
- demande
- demandee
- demandees
- demandes
- denude
- denudes
- depasse
- depasse de
- depassee
- depassement
- depassement de
- depassement limite
- depassements
- depassent
- depot
- depressurise
- depressurisees
- deregle
- dereglee
- derive
- desalignement
- desolidarisation
- desolidarise
- desolidarisee
- desserre
- desserree
- desserrees
- desserres
- dessoude
- dessoudee
- dessoudees
- dessoudes
- detendu
- detendus
- differe
- difference
- different
- different de
- differente
- differente de
- differentes
- differentes de
- differents
- differents de
- differents entre
- difficultes
- discordance
- distribues
- divers
- doivent
- doute
- douteux
- ec
- ecart
- ecartement
- ecarts
- ecco
- echec
- eclat
- egale a
- egaux
- en
- en alarme
- en anomalie
- en attendant
- en bagottement
- en contrainte
- en defaut
- en dessous
- en dessous de
- en ecart
- en ecart de
- en ecart par rapport
- en ecarts
- en erreur
- en interference
- en lieu et place
- en lieu et place de
- en mauvais etat
- en mauvaise
- en retard
- en retard par rapport
- erreur
- erreurs
- erronee
- erronees
- error
- est
- est acquise a
- est inversee
- est sorti de
- et inversement
- et non
- et non cote
- et non en
- et non pas
- et pas
- etat
- etat attendu
- etat faute
- eteint
- eteinte
- eteintes
- eteints
- exception
- excessif
- excessive
- exigence
- faible
- faibles
- failed
- fatale
- fault
- fausse
- fausses
- faute
- floue
- fluctu
- fluctuante
- fluctuation
- fluctuations
- fluctue
- gel
- habit
- hors
- hors chrono
- hors critere
- hors criteres
- hors date limite
- hors de
- hors du
- hors echelle
- hors famille
- hors gabarit

- hors limite
- hors limite par rapport
- hors limites
- hors normes
- hors securite
- hors seuil
- hors seuils
- hors spec
- hors specif
- hors specification
- hors specification evalue
- hors specifications
- hors tension
- hors tolerance
- hors tolerance par rapport
- hors tolerances
- hors tolerence
- hors validite
- hors zone
- impacts
- inaccessible
- inaccessibles
- inact
- inactif
- inadapte
- inadaptee
- inadaptees
- inadaptes
- inadvertance
- inapproprie
- inaudible
- inaudibles
- incoherence
- incoherences
- incoherent
- incoherente
- incoherentes
- incoherents
- incompatibilite
- incompatible
- incompatibles
- incomplet
- incomplete
- incomplets
- inconnu
- inconnue
- incorrect
- incorrecte
- incorrectement
- incorrectes
- incorrects
- inefficace
- inferieur
- inferieur a
- inferieure
- inferieure a
- inferieures
- inferieures a
- inferieurs
- inferieurs a la limite
- inoperant
- inoperante
- inoperants
- instabilite
- instable
- instables
- insuffisance
- insuffisant
- insuffisante
- insuffisants
- interdit
- interdite
- interference
- interferences
- inutile
- inutilis
- invalide
- invalides
- invalidite
- inverse
- inversee
- inversees
- inverses
- inversion
- jeu
- jeu important
- jusqu ' a
- l
- l '
- l ' attendu
- l ' envers
- l ' etat faute
- lieu
- limitation
- limite
- limitee
- mais
- mais pas
- mal
- mal fixe
- mal fixee
- mal fixees
- mal fixes
- mal oriente
- mal positionne
- mal positionnee
- mal regle
- mal serre
- malgre
- mauvais
- mauvais etat
- mauvaise
- mauvaises
- maximum
- mesuree
- mesuree a
- moins
- n '
- n ' a
- n ' a pas
- n ' a pas ete
- n ' a pas ete accrochee
- n ' a pas ete applique
- n ' a pas ete effectue
- n ' a pas ete faite
- n ' a pas ete realisee
- n ' a pas ete respecte
- n ' a pas ete retrouve
- n ' est
- n ' est pas
- n ' est pas a

- n ' est pas accroche
- n ' est pas adapte
- n ' est pas celle
- n ' est pas celui
- n ' est pas compatible
- n ' est pas conforme
- n ' est pas dans
- n ' est pas demande
- n ' est pas en
- n ' est pas etanche
- n ' est pas fixe
- n ' est pas fixee
- n ' est pas protege
- n ' est pas protegee
- n ' est pas reference
- n ' est pas respecte
- n ' est pas respectee
- n ' est pas serre
- n ' est pas serree
- n ' est plus
- n ' est plus fixe
- n ' est plus fixee
- n ' etait pas
- n ' etait pas en
- n ' ont pas
- n ' ont pas ete
- ne
- ne contient pas
- ne correspond pas
- ne correspondent
- ne correspondent pas
- ne couvre pas
- ne declenche pas
- ne depasse pas
- ne depassent pas
- ne devrait pas
- ne doit
- ne doivent pas
- ne donne pas
- ne figurent pas
- ne passe
- ne passe pas
- ne passent pas
- ne passent plus
- ne peut
- ne peut -
- ne peut pas
- ne peuvent
- ne peuvent pas
- ne protege pas
- ne repond pas
- ne repondent
- ne respectent pas
- ne s '
- ne s ' allume pas
- ne s ' allume plus
- ne s ' allument pas
- ne semble pas
- ne sont
- ne sont pas
- ne sont pas adaptes
- ne sont pas compatibles
- ne sont pas conformes
- ne sont pas en place
- ne sont pas fixees
- ne sont pas fixes
- ne sont pas presentes
- ne sont pas proteges
- ne sont pas reperes
- ne sont pas serres
- ne sont plus
- ne sont toujours pas
- ne tient pas
- ne tient plus
- ne tient que
- non
- non - conforme
- non a jour
- non accessible
- non accrochage
- non acquis
- non active
- non adapte
- non adaptee
- non adaptes
- non allume
- non apparie
- non appel
- non application
- non applique
- non assuree
- non atex
- non atteint
- non atteinte
- non attendu
- non attendue
- non attendues
- non attendus
- non autorise
- non basculement
- non bl
- non bloquant

- non bloque
- non branches
- non coherente
- non compatible
- non compatibles
- non conforme
- non conforme a
- non conformes
- non conformite
- non connecte
- non connectes
- non continuite
- non correct
- non correcte
- non correspondance
- non coupes
- non demande
- non demandee
- non demandees
- non demarrage
- non demarre
- non demontable
- non depose
- non disponible
- non effectue
- non effectuee
- non effectues
- non en place
- non etanche
- non etanches
- non fermees
- non fermeture
- non fixe
- non fixee
- non fixees
- non fixes
- non flammes
- non fournies
- non fournis
- non freine
- non freinee
- non garantie
- non graissees
- non identifie
- non identifiee
- non identifiees
- non identifies
- non impression
- non indique
- non lineaire
- non mentionnees
- non mis
- non mise
- non monte
- non montee
- non montes
- non nominal
- non obtent
- non obtention
- non obtenu
- non obtenus
- non obture
- non obtures
- non ok
- non positionnee
- non precise
- non prevu
- non prevue
- non pris
- non prise
- non prises
- non protege
- non protegee
- non protegees
- non proteges
- non raccordee
- non raccordees
- non realise
- non realisee
- non realisees
- non reconnu
- non referencee
- non reglementaire
- non remontee
- non repere
- non reperee
- non reperes
- non respect
- non respecte
- non respectee
- non respectees
- non respectes
- non securises
- non serre
- non serree
- non serrees
- non serres
- non significatif
- non significative
- non specifie
- non stable
- non suivi
- non trouve
- non trouvee
- non utilise
- non utilisee
- non utilises
- non valide
- non valides
- non visible
- non visibles
- non vu
- nulle
- obtenu
- obtenu a
- obtenue
- ouverte
- par depassement
- par erreur
- par rapport
- parasite
- parasites
- pas
- pas adapte
- pas assez

- pas compatible
- pas conforme
- pas conforme a
- pas conformes
- pas correct
- pas correcte
- pas correctement
- pas de
- pas en
- pas etanche
- pas nominal
- pas reperes
- pas represente
- pas respecte
- pas serree
- pas strictement
- pas vu
- pb
- pb de
- perdu
- perdue
- perdus
- perte
- pic
- plus
- plus faible que
- plus importante
- plus long que
- pour
- pour attendu
- pour attendu a
- pour attendue
- pour des limites
- pour probleme
- pour spec
- pour une
- pour une limite
- pour une spec
- pour une specif
- pour une specification
- preconise
- presence
- probabilite
- probable
- probleme
- problemes
- quelque soit
- refus
- refuse
- refusee
- reglage
- resultat
- retard
- risque
- s ' allume
- s ' est
- sans
- saturees
- satures
- se
- se decolle
- seulement
- sont
- sorti de
- sortie de
- specif
- superieur
- superieur a
- superieure
- superieure a
- superieures
- superieures a
- superieurs
- superieurs a
- superieurs aux
- sur
- sur defaut
- suspect
- suspectee
- traces
- trop
- trop courte
- trop courtes
- trop courts
- trop eleve
- trop elevee
- trop faible
- trop faibles
- trop grand
- trop grande
- trop important
- trop lache
- trop long
- trop longs
- trop longue
- trop longues
- trop petit
- trop tendu
- trouve
- trouvee
- trouvees ouvertes
- un
- uniquement
- valide
- varie

Annexe C

Algorithme de correspondance

```
// ======= Initialisation des données =======

DÉFINIR rapport = "TESTS TYPE 1 DES RECHAUFFEURS AXES 600 ET 601SUITE A LA PERTE
    ..."
DÉFINIR rapport_corrige = "Tests type 1 des réchauffeurs axes 600 et 601. Suite
    à la perte..."
DÉFINIR wdiff_output = "Tests type 1 des [-RECHAUFFEURS-] {+réchauffeurs+} axes
    600 et [-601SUITE A-] {+601. Suite à+}..."


// ======= Découpage des segments =======

DÉFINIR intervalles = DIVISER wdiff_output PAR le motif "(\[-.*?-\] \{\+.*?\+\})
    "


// ======= Traitement des segments =======

DÉFINIR offset = 0
DÉFINIR decalage = 0
DÉFINIR translation = LISTE_VIDE

POUR CHAQUE interv DANS intervalles :
        SI interv COMMENCE_PAR "[" :
                CHERCHER le motif "\[-(.*?)-\] \{\+(.*?)\+\}" dans interv
                SI motif TROUVÉ :
                        EXTRAIRE rapport_text = groupe 1
                        EXTRAIRE rapportCorrige_text = groupe 2
                        DÉFINIR start = offset
```

```
                        DÉFINIR end = start + LONGUEUR(rapport_text)
                        SI LONGUEUR(rapport_text) > 0 :
                                DÉFINIR coef = LONGUEUR(rapportCorrige_text) /
    LONGUEUR(rapport_text)
                        SINON :
                                DÉFINIR coef = 1
                        AJOUTER (start, end, decalage, coef) À translation
                        decalage = decalage + LONGUEUR(rapportCorrige_text) -
    LONGUEUR(rapport_text)
                        offset = offset + LONGUEUR(rapport_text)
                SINON :
                        AFFICHER "Erreur de parsing :", interv
        SINON :
                DÉFINIR start = offset
                DÉFINIR end = start + LONGUEUR(interv)
                AJOUTER (start, end, decalage, 1) À translation
                offset = offset + LONGUEUR(interv)


// ======= Fonction de traduction des offsets =======

FONCTION translate(x, trans) :
        POUR CHAQUE segment DANS trans :
                EXTRAIRE start, end, translat, coef DE segment
                SI start <= x < end :
                        DÉFINIR local_adjust = ARRONDIR((x - start) * (coef - 1)
    )
                        RETOURNER ARRONDIR(x + translat + local_adjust)
        RETOURNER x


// ======= Utilisation et affichage =======

DÉFINIR (x, y) = (22, 80)
DÉFINIR xt = translate(x, translation)
DÉFINIR yt = translate(y, translation)
AFFICHER xt, yt
AFFICHER rapport[x:y], "-->", rapport_corrige[xt:yt]
```

```
POUR CHAQUE t DANS translation :
        EXTRAIRE start, end, dec, coef DE t
        AFFICHER (start, end, dec, coef)
        AFFICHER t[0]+t[2], ARRONDIR((t[1] + (t[0] - start) * (coef-1)))
        AFFICHER rapport[start:end], "-->", rapport_corrige[start + dec:end +
    ARRONDIR(dec * coef-1)]
```

Titre : Approches du dysfonctionnement technique dans les REX d'Ariane 5 : de l'analyse linguistique outillée de son expression vers la modélisation TRIZ du problème

Mots clés : Traitement Automatique des Langues, Résolution de problèmes techniques, Linguistique de corpus, TRIZ, Lanceurs spatiaux

Résumé : Les REX (Retours d'EXpérience) sont des documents textuels dont la visée est de rapporter un problème, ou un dysfonctionnement, et qui jouent un rôle important dans la maîtrise des risques au sein d'une organisation. Plusieurs travaux de TAL (Traitement Automatique des Langues) ont donc vu le jour afin de capitaliser les connaissances qu'ils abritent. Par ailleurs, des méthodes de résolution de problèmes techniques ont été développées, comme la méthode TRIZ, et présentent un intérêt non négligeable pour les dysfonctionnements qui peuvent être rapportés dans les REX. De ce fait, un partenariat s'est créé entre le CNES qui cherche à exploiter ses REX liés aux lanceurs spatiaux, et la société MeetSYS, spécialisée dans la méthode TRIZ pour la capitalisation du savoir expert. Cette thèse s'est vue comme l'opportunité d'explorer l'utilisation du TAL et de la linguistique de corpus pour l'extraction fine d'information dans les REX d'Ariane 5 en vue de modéliser un dysfonctionnement technique sous forme de vépole (formalisme propre à TRIZ). Cela signifie être capable de partir d'un texte brut, spécialisé et bruité vers un formalisme conçu indépendamment des données en question. À cette fin, une démarche en plusieurs étapes a été mise en place en vue de se rapprocher autant que possible de ce formalisme. L'un des piliers sur lequel s'appuie cette démarche est la sémantique des cadres, et la ressource FrameNet qui en découle, qui nous permet d'identifier et de qualifier les éléments textuels qui constituent le problème. Nous explorons dans cette thèse plusieurs approches de TAL et de linguistique de corpus dans l'étude des REX, soit des textes spécialisés et bruités, pour identifier les structures sémantiques qui composent l'expression d'un dysfonctionnement technique. Nous mêlons ainsi des techniques comme le Topic Modeling, word2vec et de l'analyse lexicale outillée pour de l'exploration de corpus, du fine-tuning de modèles neuronaux pour de l'étiquetage automatique, l'utilisation de LLMs pour de la normalisation et de l'annotation automatique, mais aussi de l'analyse syntaxique et de la reconnaissance de patrons pour l'analyse fine des structures langagières. Dans un premier temps, une analyse du corpus nous a permis de dégager une typologie d'expressions d'un dysfonctionnement technique en neuf classes. Elle est basée sur la détection de marqueurs lexicaux au sein de la description de l'anomalie qui a été repérée et décrite. À partir de cette typologie, nous avons pu effectuer une annotation des marqueurs lexicaux spécifiques au sein du corpus. Celle-ci nous a permis d'explorer l'utilisation d'annotateurs non experts du domaine sur des données spécialisées et, par la suite, d'entraîner un modèle neuronal à base de transformers pour l'étiquetage automatique des rapports d'anomalies. Nous avons aussi mené une étude afin de normaliser automatiquement ces rapports pour en supprimer le bruit, avant de tester l'impact de cette normalisation sur l'entraînement du modèle. Cette étude n'ayant pas montré d'améliorations sur la tâche d'étiquetage automatique nous entraîne à interroger la pertinence de la normalisation des données bruitées, et notamment en fonction de la tâche visée. Par la suite, nous avons pu focaliser notre étude sur deux catégories de la typologie qui sont la Fuite d'un liquide ou d'un gaz et la Présence d'un obstacle. Pour la première, nous avons mis en place une approche impliquant plusieurs méthodes complémentaires de linguistique de corpus afin de faire émerger un frame de la fuite dans un environnement technique. Nous avons ainsi pu identifier les différents éléments qui composent l'expression de la fuite. Pour la catégorie Présence d'un obstacle, nous avons utilisé des LLMs génératifs pour l'annotation automatique de ces textes. Par ce biais, nous avons pu explorer les capacités et les limites d'un LLM à effectuer une annotation de type Frame Semantic Role Labeling, mais aussi à traiter un texte spécialisé et bruité.

Title: Analysing technical dysfunction in Ariane 5 anomaly reports : from its linguistic expression to TRIZ modeling of the problem

Key words: Natural Language processing, Technical problem solving, Corpus linguistic, TRIZ, Space launchers

Abstract: Anomaly reports are textual documents whose purpose is to report a problem or malfunction, and play an important role in risk management within an organisation. Numbers of NLP (Natural Language Processing) projects have therefore been developed to capitalise on the knowledge they contain. In addition, methods for solving technical problems exists, such as TRIZ, which are of considerable interest for the type of malfunction reported in incident or anomaly reports. As a result, a partnership has been formed between CNES, seeking to exploit these reports, and MeetSYS, a company specialised in using TRIZ to capitalise on expert knowledge. This thesis was seen as an opportunity to explore the use of NLP and corpus linguistics for the fine-grained extraction of information from the Ariane 5 anomaly reports in order to model a technical malfunction in the form of a vepole (formalism specific to the TRIZ method). This means being able to move from a raw, specialised and noisy text to a formalism independently designed. To this end, a multi-stage approach was put in place to get as close as possible to this formalism. One of its pillars is frame semantics, and the resulting FrameNet resource, which allows the identification and qualification of the textual elements of the problem. In this thesis, we explore several NLP and corpus linguistic approaches in the study of the reports, i.e. specialised and noisy texts, in order to identify the semantic structures that constitute the expression of technical malfunction. We thus combine various techniques such as Topic Modeling, word2vec and tool-based lexical analysis for corpus exploration, fine-tuning of neural models for automatic labelling, the use of LLMs for normalisation and automatic annotation, as well as syntactic analysis and pattern recognition for fine-grained analysis of language structures. A first corpus study allowed the identification of a nine-class typology of technical malfunction expression. It is based on lexical markers detection within the anomaly description. Using this typology allowed the annotation of specific lexical markers within the corpus. This enabled us to explore non-domain expert annotation on specialised data and, subsequently, to train a neural model based on transformers for anomaly reports automatic labelling. We also conducted a study for automatic normalisation of the reports to remove noise, before testing its impact on model training. Since this study did not show any improvement in the automatic labelling task, we are thus questioning the relevance of noisy data normalisation, particularly regarding the target task. We then focused our study on two class of the typology : Leakage of a liquid or a gas and Presence of an obstacle. For the former, we implemented several complementary corpus linguistic methods in order to bring out the frame of the leak in a technical environment. We were thus able to identify what elements make up the expression of leakage. For Presence of an obstacle, we explored the use of generative LLMs for automatic labeling. In this way, we were able to explore the capabilities and limitations of an LLM for Frame Semantic Role Labeling, as well as processing specialised and noisy text.