11e CONGRES INTERNATIONAL DE GENIE INDUSTRIEL – CIGI2015
Québec, Canada
26-28 octobre 2015

Étude comparative des méthodes d'analyse de similarité des défaillances de systèmes aéronautiques

ANOUAR HAKIM ELBADIRY[1], SAMUEL BASSETTO[2], MOHAMED-SALAH OUALI[3]…

[1] Candidat PhD Polytechnique
École Polytechnique, Montréal Canada
anouar-hakim.elbadiry@polymtl.ca

[2] PhD Polytechnique
École Polytechnique, Montréal Canada
samuel-jean.bassetto@polymtl.ca

[3] PhD Polytechnique
École Polytechnique, Montréal Canada
mohamed-salah.ouali@polymtl.ca

Résumé **- Dans l'industrie aéronautique, savoir connecter une observation de défaillance à des incidents opérationnels préalablement observés et analysés est une compétence recherchée non seulement pour accélérer le diagnostic de défaillances, mais aussi pour maîtriser les risques opérationnels liés aux systèmes aéronautiques et aux services proposés. Cet article propose une étude comparative des méthodes d'analyse de similarité des défaillances appliquées au domaine de l'aéronautique. Ces méthodes permettent de regrouper les incidents opérationnels consignés sous formes de textes par rapport à une taxonomie prédéfinie. L'étude compare la performance des méthodes à base de corpus et celles à base de connaissances selon que l'espace sémantique provient de domaines différents ou issue du domaine aéronautique. L'analyse de pertinence est réalisée sur un échantillon d'enregistrements de la base de données des accidents avioniques canadiens : CADORS. Les résultats de pertinence obtenus sont 10% meilleurs que les méthodes à base de connaissances, et 10% meilleurs que les méthodes à base de corpus appliquées dans des espaces sémantiques différents du corpus étudié.**

Abstract **- In the aviation industry, the ability to connect the observation of a failure to previously observed and analyzed operational incidents is a skill that is required not only to accelerate the diagnosis of failures but also to master the operational risks related to aviation systems and services. This article presents a comparative study of similarity analysis methods for failures applied to the domain of aviation. These methods enable users to group together operating incidents recorded in the form of texts in relation to a predefined taxonomy. The study compares the performance of corpus-based versus knowledge-based methods depending on whether the semantic space comes from different domains or specifically from aviation. The relevance analysis is done on a sample of records from the database for Canadian aviation accidents: CADORS. The relevance results obtained are 10% better than the knowledge-based methods and 10% better than the corpus-based methods applied in semantic spaces that differ from the corpus under study**

Mots clés **– Méthodes à base de connaissances, Méthodes à base de corpus, Similarité sémantique, Analyse de risque**
Keywords **- Knowledge-based method, Corpus-based method, Semantic similarity, Risk analysis**

1 INTRODUCTION

Au cours de la dernière décennie, les industries hautement technologiques ont démarré plusieurs projets de recherche afin de déceler les risques latents et émergents associés à leurs systèmes. Ces entreprises cherchent à se doter d'une politique proactive face aux risques industriels lors de l'exploitation de leurs systèmes dans divers contextes et conditions d'usage. Ces recherches ont mis le point sur l'utilité des analyses de risques dynamiques pour la maîtrise des risques opérationnels. L'idée principale de l'analyse de risques dynamiques se fonde sur le retour d'expérience à partir des informations recueillies des événements anormaux enregistrés pour actualiser les analyses de risques préalablement effectuées avant la mise en service du système concerné.

Cet article s'inscrit dans le champ de recherches sur le retour d'expérience dans le domaine de haute technologie en particulier les industriels de turboréacteurs. Ces industriels gèrent plusieurs processus de retour d'expériences tels que (développement, fabrication, test, installation, maintenance, etc.). Ces processus formels de remontée d'information fournissent des milliers d'enregistrements (incidents) par année. Chaque enregistrement comporte une partie codifiée et un champ libre permettant aux

experts de saisir, dans un langage libre, un texte relativement court (100 mots en moyenne) décrivant l'incident survenu. Chaque enregistrement fait l'objet d'une évaluation et éventuellement donne lieu à une action corrective. Ces enregistrements représentent une source d'informations sur les risques réels du système, et pouvaient à ce titre être utilisés efficacement pour les analyses préliminaires de risques des nouveaux systèmes ainsi que la mise à jour des analyses préliminaires de risques (APR) élaborées sous format AMDEC (Analyse des Modes de Défaillances, de leurs Effets et de leur Criticité) pour les systèmes existants. Malheureusement, ce retour d'information n'est pas capitalisé et utilisé à sa juste valeur.

Les analyses préliminaires de risques (APR), bien que visant l'exhaustivité, ne sont jamais complètes. L'expertise des analystes impliqués dans l'APR étant limitée, l'APR l'est aussi. Ces analyses sont faites de manière préliminaire ; elles se fondent sur les estimations d'experts qui peuvent être confortées ou remises en cause par les faits observés. Cette connexion entre l'APR d'un système et ses enregistrements d'événements est à la base des analyses de risques dynamiques. Cette problématique a été approchée par l'étude [Mili et al., 2008, Mili et al., 2009] dans le domaine des semi-conducteurs. Cependant, elle ne se traduit que rarement dans la réalité.

L'APR est une activité réalisée durant la phase de conception du système. Afin de relier cette activité avec les enregistrements d'observations d'opportunités d'améliorations survenant durant d'autres phases du cycle de vie du système, deux approches sont essentiellement employées. La première fait appel à une taxonomie prédéfinie afin de relier les enregistrements similaires et calculer, ainsi l'occurrence de leurs risques. Cette taxonomie permet aussi de relier les enregistrements et les APRs [Tumer et Stone, 2003, Mili et al., 2009]. La deuxième ne propose qu'un simple retour de fiches d'enregistrements aux concepteurs sans fournir de véritables liens entre les observations ni les analyses de risques. Cette dernière façon de procéder est jugée inefficace [Tulechki, 2011].

Dans un cas industriel, une taxonomie devient complexe à définir, utiliser et maintenir. Un système aéronautique tel un réacteur peut compter plusieurs milliers de sous-ensembles et des milliers de modes de défaillance. Alors qu'il est reconnu en ergonomie de logiciel dont la sélection de champs dans des listes de plus de 10 items devient fastidieuse. Il est donc compréhensible que la sélection des modes de défaillances dans une arborescence contenant des milliers de codes est extrêmement délicate. Les résultats escomptés sont alors erronés. Une autre limite de la codification est l'appauvrissement du contenu informationnel. Le fait de réduire un texte à un code prédéfini a pour effet de ne garder que les éléments saillants de l'événement au détriment de subtilités importantes présentes dans le texte.

Il serait pertinent de permettre d'une part aux opérationnels sur le terrain de décrire les problèmes rencontrés dans des champs de texte libre. D'autre part, le rapprochement entre les enregistrements passés et les actuels devrait être réalisé sur une recherche en plein texte. Un outil plus proche du langage naturel serait mieux utilisé, permettant aux experts d'explorer les collections de rapports en fonction des particularités de leur contenu textuel. Le rapprochement d'une observation décrite en langage naturel d'une autre observation ou d'un risque également explicité par un texte relève de l'analyse de similarité entre deux textes. C'est à cela que s'intéresse cet article.

La performance de l'analyse de similarité sémantique textuelle est limitée par deux phénomènes : la polysémie et la synonymie. Les méthodes d'analyse sémantique des mots sont introduites pour minimiser l'impact de ces deux phénomènes. Deux catégories de méthodes d'analyse peuvent être distinguées : 1) celles qui sont à base de corpus et 2) celles qui sont à base de connaissances. Les méthodes à base de corpus permettent de calculer la similarité entre les différents mots du corpus en utilisant un espace sémantique très large. Alors que les méthodes à base de connaissance utilisent des ressources lexicales pour calculer la similarité entre les mots.

Dans les études réalisées auparavant sur les méthodes à base de corpus [Mihalcea et al., 2006, Shrestha, 2011], l'espace sémantique utilisé diffère du domaine analysé. Plusieurs travaux de recherche ont analysé l'impact du choix de l'espace sémantique sur la pertinence des résultats de similarité obtenus, sans proposer une solution permettant d'améliorer la pertinence des résultats [Wicaksana et Wahyudi, 2011].

L'objectif de cet article est d'évaluer les performances des méthodes à base de corpus [LSA] [Landauer et al., 1998] et [NPMI] [Turney, 2001, Bouma, 2009] appliquées à un domaine et un espace sémantique identiques. Nous avons choisi les enregistrements d'incidents/accidents CADORS [Tulechki et Tanguy, 2013], comme domaine d'analyse. Ces rapports sont enregistrés sous formes de textes courts de 20 à 200 mots écrits en anglais avec un langage technique très proche du langage utilisé par les industriels œuvrant dans domaine de l'aéronautique à travers le monde.

Nous avons choisi quatre (4) méthodes d'analyse de la similarité entre les mots : [LSA] , [NPMI], [W&P] [Wu et Palmer, 1994] et [LIN] [Lin, 1998]. Ces méthodes seront combinées avec la méthode générique de représentation vectorielle (GVSM) [Wong et al., 1985] afin de calculer la similarité entre les enregistrements de la base de données (CADORS). La performance des résultats obtenus est assurée par : les résultats pertinents retrouvés, les résultats pertinents non retrouvés et les résultats non pertinents retrouvés. Un résultat est considéré pertinent s'il coïncide avec les jugements des experts.

À la suite de cette introduction, la section 2) présente une revue des méthodes d'analyse de la similarité sémantique. Les travaux réalisés sur l'analyse des risques à travers l'exploitation des rapports d'incidents, et sur la comparaison de différentes méthodes d'analyse de la similarité sémantique de mots. La section 3 évalue les performances des quatre méthodes considérées. L'impact du choix de l'espace sémantique et de sa taille sur les performances des méthodes à base de corpus sont évalués. Un exemple est présenté pour illustrer les différentes étapes de l'expérience. La section 5 présente une conclusion et les recommandations quant à l'usage des méthodes de calcul de similarité dans l'estimation dynamique des risques dans le domaine aéronautique.

2 REVUE DE LITTERATURE

Dans le domaine des semi-conducteurs, Mili *et al.* [Mili et al., 2008, Mili et al., 2009] ont utilisé une taxonomie préétablie afin de connecter les différents événements opérationnels et les relier à leurs analyses de risques préliminaires. Le système de codification est difficilement adapté à l'industrie aéronautique caractérisée par une technique imbriquée et complexe. L'étude récente de Tulechki (2011) réalisée sur les rapports d'incidents

aériens a mis le point sur les faiblesses du système de codification. La codification utilisée représente la réalité à un instant donné, alors que la réalité est en perpétuelle évolution, ce qui contredit les principes des analyses de risques dynamiques. De plus, la représentation d'un événement par un code réduit la richesse contenue dans les enregistrements, et ignore plusieurs informations qui, tout en étant présentes dans le texte original, ne trouvent pas leur place dans la codification choisie. Ce travail de recherche a soulevé l'importance de l'analyse sémantique pour la connexion des événements similaires. Tulechki [Tulechki, 2011] propose d'utiliser les techniques de traitement automatique du langage naturel (TALN) pour étudier la similarité entre des enregistrements d'incidents et d'accidents. Cette approche permet, selon leurs auteurs, de remédier aux faiblesses du système de codification adopté généralement par les industries.

Le modèle sémantique vectoriel (VSM) [Salton et al., 1975] est considéré parmi les premiers modèles permettant de représenter un document sous forme d'un sac à mots. Dès lors, un second document sera présenté également comme un ensemble de mots et il sera possible de représenter une matrice *(T)* avec en ligne les textes (T_i), en colonne les mots (M_j). Le terme de cette matrice à l'emplacement *(i,j)* vaut l'occurrence du mot M_j dans le texte T_i. Il est alors possible d'effectuer un calcul de distances entre les différentes lignes de la matrice, autrement dit, entre les différents textes. La matrice de similarité (S) du modèle VSM peut s'écrire comme le produit scalaire de la matrice texte T et sa transposé T^t : $S = T * T^t$ (Éq1).

Cependant, le modèle VSM présente principalement trois limitations :

- La première est que l'espace vectoriel formé par les termes est considéré comme orthogonal. Cette hypothèse est forte, car le calcul de la similarité sémantique fait face à l'ambiguïté lexicale entre les termes. Un mot peut prendre plusieurs sens dépendamment du contexte (polysémie), et un sens peut être exprimé par plusieurs termes (synonymie). Cette problématique affecte la pertinence des résultats obtenus. Afin de remédier à cette limitation, un prétraitement dit de correction est appliqué. Il généralise le modèle VSM (GVSM) afin qu'il puisse tenir compte des corrélations dans le calcul de la distance entre les deux textes [Wong et al., 1985]. La matrice de similarité (S) du modèle GVSM peut s'écrire comme le produit scalaire de la matrice texte (T), la matrice de similarité entre les différents mots (A) et le transposé T^t : $S = T * A * T^t$ (Éq2).
- La deuxième limitation est que plusieurs termes possèdent un caractère générique (par exemple : de, la, je, vous, nous, la ponctuation, etc. Ceci nécessite une étude préalable pour éliminer les termes génériques afin d'améliorer la pertinence de l'analyse de la similarité entre les différents textes.
- La troisième limitation est que le calcul de similarité dans le modèle VSM se fonde sur l'occurrence des termes dans les textes. Plus les textes sont longs, plus la probabilité de trouver des termes similaires est forte. Ce phénomène rend difficile la comparaison de textes longs avec la méthode VSM.

Les études [Tulechki et Tanguy, 2012, Tulechki et Tanguy, 2013] ont utilisé les méthodes d'analyse de la similarité afin de connecter les incidents aéronautiques. La méthode dite de second ordre est proposée en utilisant un espace vectoriel à base de documents (pivots) au lieu de mots pour représenter les enregistrements permettant ainsi d'améliorer la densité des matrices de représentations. Il sera donc possible de représenter une matrice avec en ligne les textes (Ti), en colonne les textes pivots (Pj). Le terme de cette matrice à l'emplacement (i,j) vaut la valeur de la similarité entre le texte (Ti) et le texte pivot (Pi) calculée en utilisant l'une des méthodes de calcul de la similarité textuelle telle que VSM ou GVSM. Ces études soulignent l'utilité de l'analyse sémantique pour connecter les événements opérationnels et maîtriser ainsi les risques industriels récurrents et émergents. Elles permettent de faire face aux lacunes du système de codification utilisé. Cependant, ces études ne spécifient pas les méthodes adoptées pour analyser la similarité textuelle entre le texte Ti et le texte pivot Pj. L'apport en pertinence de l'utilisation de cette méthode par rapport aux méthodes ordinaires (VSM et GVSM) n'était pas chiffré.

Le modèle généralisé de la représentation vectorielle GVSM fait appel aux méthodes d'analyse de la similarité entre les mots pour calculer la matrice (A) de la formule (Éq2). Le terme de cette matrice à l'emplacement (i,j) vaut la similarité ($\mathrm{Sim}(w1, w2)$) entre les deux mots w1 et w2. Les résultats de la similarité entre les différents textes dépendent largement du calcul de cette matrice. Le calcul de la similarité entre termes repose sur deux approches différentes. La première est purement statistique et dépend entièrement du corpus choisi comme un espace sémantique (méthodes à base de corpus). Alors que la deuxième approche est fondée sur la connaissance lexicale du langage humain (méthodes à base de connaissance).

2.1 Méthodes à base de corpus

Les méthodes à base de corpus permettent de calculer la similarité entre les mots en utilisant un espace sémantique large afin que les résultats soient significatifs et pertinents. Le calcul de similarité entre les mots issus de l'ensemble des enregistrements est lié à leurs cooccurrences dans les différents textes du corpus. Nous citons deux méthodes : l'Analyse Sémantique Latente [LSA] et la méthode normalisée [NPMI] basée sur la méthode Point d'Information Mutuelle et récupération de l'information [PMI-IR].

- La méthode [LSA] [Landauer et al., 1998] permet de calculer la similarité sémantique entre les mots. Chaque mot est représenté par un vecteur dans un espace de documents d'un espace sémantique très large. La similarité sémantique entre les mots dépend de leurs cooccurrences dans les documents du corpus choisi. Cette méthode utilise la théorie de décomposition en valeurs singulières (SVD) qui permet de décomposer la matrice de cooccurrences mots-documents (T) au produit de trois matrices $T = U * P * V^T$. (U) est la matrice des vecteurs propres de ($T^T * T$). (V) est la matrice des vecteurs propres de ($T * T^T$). (P) est la matrice diagonale contenant les valeurs singulières de la matrice (T). Une approximation de la matrice (T) est introduite en négligeant les valeurs singulières faibles. La méthode [LSA] permet ainsi de réduire la dimension de l'espace vectoriel afin d'accélérer les calculs de la similarité.
- La méthode [PMI-IR] permet de calculer la similarité sémantique entre deux mots en fonction de leurs cooccurrences dans les documents de l'espace sémantique choisi. C'est une approche similaire à celle de la méthode [LSA], sauf que les méthodes de calcul utilisées sont différentes. Cette méthode est largement utilisée par les moteurs de recherche internet comme « Yahoo ». Turney (2001) a proposé quatre requêtes pour calculer cette similarité.

Nous optons pour la première proposition pour sa simplicité à mettre en œuvre. Cette méthode a été normalisée pour devenir [NPMI] [Bouma, 2009] afin que les valeurs de similarité soient comprises entre -1 et 1. La similarité entre deux mots w1 et w2 en utilisant la méthode [NPMI] se fonde sur le calcul des probabilités d'occurrences des mots dans le corpus choisi comme espace sémantique, la formule s'écrit :

$$\mathrm{Sim}(w1,w2) = \left(\log\left(\mathrm{p}(w1,w2) \Big/ \mathrm{p}(w1) * \mathrm{p}(w2)\right)\right) * \left(-\frac{1}{\log(\mathrm{p}(w1,w2))}\right) \quad (\text{Éq.3})$$

2.2 Méthodes à base de connaissances

Plusieurs méthodes à base de connaissances étaient développées pour permettre aux utilisateurs de calculer la similarité entre mots avec certitude. L'utilisation de ces méthodes a besoin de ressources d'inventaire lexical comme les dictionnaires. Par ailleurs, d'autres ressources lexicales structurées sont développées depuis les années 2000 pour faciliter la programmation du calcul des distances sémantiques entre mots comme WordNet [Fellbaum, 1998]. Par contre, l'affectation du sens correct d'un mot dans un contexte donné reste une tâche délicate à automatiser. Les logiciels développés permettent d'estimer le sens d'un mot mais les erreurs restent présentes limitant ainsi la pertinence de ces outils. Trois approches existent : la similarité à base de traits, la similarité à base de distance taxonomique et la similarité à base de contenu informationnel.

2.2.1 Similarité à base de traits

Son origine revient souvent dans la littérature à l'étude de Tversky (1977). Ses études considèrent la similarité entre deux termes comme l'ensemble des traits communs pondérés dont on retire les traits spécifiques à chaque terme. Par la suite Lesk (1986) a proposé une nouvelle formule très simple pour le calcul de la similarité en définissant les traits comme l'ensemble des mots de la définition d'un sens. La formule proposée par Lesk est: $\mathrm{Sim}(x,y) = \mathrm{D}(x) \cap \mathrm{D}(y)$, avec D(x) est l'ensemble des mots qui constituent la définition du mot (x) dans la base lexicale choisie. La pertinence de cette méthode est tributaire des mots présents dans cette ressource lexicale. Plus que la définition du mot est longue plus qu'il est trouvé similaire à plusieurs mots. Afin de remédier à ce problème, Wilks et Stevenson (1998) ont proposé de pondérer chaque mot de la définition par l'ensemble de ses sens exprimés. Pour cibler les mots liés au contexte d'étude, Navigli (2009) a restreint le calcul de similarité des termes objets de la requête aux mots spécifiques au contexte de l'analyse. De cette manière la dispersion de l'analyse a été réduite. Cette proposition nécessite une grande supervision humaine et présente une grande difficulté pour définir les traits des termes spécifiques du domaine analysé. Dans une étude récente, le dictionnaire standard utilisé par Lesk (1986) était remplacé dans l'étude de Banerjee et Pedersen (2002) par la ressource structurée WordNet qui grâce à sa hiérarchie donne, de meilleurs résultats. Pirró et Euzenat (2010) ont proposé très récemment une nouvelle formule permettant de cibler les traits communs ou d'asymétrie entre deux termes selon les valeurs données aux paramètres de la formule normalisée. Cependant, ce concept de traits demeure vague et présentent beaucoup de difficultés dans la pratique.

2.2.2 Similarité à base de distance taxonomique :

Cette approche est traduite par le calcul du nombre des arcs séparant les termes. Rada et al. (1989) ont basé le calcul de similarité entre deux concepts c1 et c2 sur l'identification du nombre minimal d'arcs reliant les deux concepts. Cette approche est souvent utilisée dans l'analyse de la similarité entre les configurations de gènes provoquant certaines maladies graves, caractérisées par des données numériques. Cette méthode avantage les termes les plus proches de la racine par rapport aux termes génériques. Pour remédier à cette limitation, Wu et Palmer (1994) ont proposé une nouvelle formule qui tient compte de la position du plus petit ancêtre commun par rapport à la racine. La formule de calcul de la similarité de la méthode [W&P] s'écrit : $\mathrm{Sim}_{w\&p}(c1,c2) = \frac{2*N3}{N1+N2+2*N3}$ (Éq 4), Avec N1 est le nombre d'arcs entre le concept c1 et le plus petit ancêtre commun de c1 et c2. N2 est le nombre d'arcs entre le concept c2 et le plus petit ancêtre commun de c1 et c2. N3 est le nombre d'arcs entre la racine et le plus petit ancêtre commun de c1 et c2. Par la suite, Leacock et Chodorow (1998) ont choisi de normaliser la formule de Rada et al. (1989) en utilisant la profondeur totale par rapport à la taxonomie. Ainsi, la similarité entre deux concepts est calculée en utilisant la longueur du chemin le plus court et le chemin le plus long entre le concept racine et le concept le plus bas dans la taxonomie.

2.2.3 Similarité à base de contenu informationnel :

Ces approches sont similaires à celles qui sont fondées sur les arcs (à base de taxonomie), sauf que la distance entre les concepts n'est pas utilisée pour le calcul de la similarité. Cette propriété lui donne avantage lorsque le calcul de la distance taxonomique manque de fiabilité. La pertinence de ces approches dépend largement de la structure et de la hiérarchisation des concepts dans la ressource lexicale utilisée. Le contenu informationnel (IC) d'un concept (C) est donné par la formule $\mathrm{IC}(C) = -\log(\mathrm{P}(C))$. Resnik [Resnik, 1995, Resnik, 2011] a exclu le contenu informationnel des concepts objet de l'analyse de la similarité. Il a utilisé seulement le contenu informatif de l'ensemble des concepts qui subsume les deux concepts c1 et c2 pour mesurer la similarité. Jiang et Conrath (1997) jugent que l'intégration du contenu informatif lié aux deux concepts analysés améliore la pertinence des résultats. Ils proposent une nouvelle formule de calcul de la similarité. Une année plut tard Lin (1998) a proposé une formule similaire qui prouve son efficacité au niveau de la pertinence de l'analyse de la similarité entre les mots [Slimani et al., 2007]. La formule de similarité de [LIN] entre deux concepts c1 et c2 s'écrit : $\mathrm{Sim}(c1,c2) = \frac{2*\log(\mathrm{P}(C))}{\log(\mathrm{P}(c1))+\log(\mathrm{P}(c2))}$ (Éq 5).

Afin de comparer la pertinence des méthodes de calcul de similarité entre les mots, plusieurs études sont réalisées sur le sujet. Mihalcea et al. (2006) ont comparé la pertinence de huit méthodes : Deux méthodes à base de corpus : [PMI-IR] et [LSA], et six méthodes à base de connaissance : [J&C], [L&C], [Lesk], [LIN], [W&P]et [Resnik]. Les indicateurs de pertinence utilisés sont : La précision P (**P** = $\frac{a}{a+b}$) qui permet de calculer le ratio des résultats pertinents retrouvés (a) par rapport aux résultats non pertinents retrouvés (b). Le rappel (**R** = $\frac{a}{a+c}$) qui permet de calculer le ratio des résultats pertinents retrouvés (a) par rapport aux résultats pertinents non retrouvés (c). F-mesure (**F** = $\frac{2*P*R}{R+P}$) qui permet de combiner et de pondérer les deux indicateurs P et R. En effet, ce critère permet de juger la pertinence des différentes méthodes d'analyse de la similarité entre mots. Les résultats

obtenus montrent que la méthode [PMI-IR] est la plus pertinente par rapport aux autres méthodes. Bien que la différence soit très mince au niveau des indicateurs de comparaison (P, R et F), cette conclusion est très encourageante puisque les méthodes à base de corpus sont plus faciles à mettre en œuvre et ne nécessitent aucune supervision humaine. Dans cette étude, Le corpus « British National Corpus » (BNC) était utilisé comme un espace sémantique pour le calcul en utilisant la méthode [LSA]. Alors que les résultats de la méthode [PMI-IR] étaient calculés en utilisant l'ancien moteur de recherche AltaVista. L'analyse de pertinence est faite sur les textes du Microsoft Research Paraphrase (MRPC). Dans cette étude les résultats des deux méthodes [LSA] et [PMI-IR] n'étaient pas fondés sur le même espace sémantique. Li et al. (2010) ont réalisé une étude comparative sur les méthodes d'analyse sémantique à base de connaissance en utilisant WordNet. Ils ont conclu que [L&C] donne des résultats meilleurs que [W&P], et que la méthode [J&C] est plus efficace que la méthode de [LIN]. Par la suite, Wicaksana et Wahyudi (2011) ont montré que les méthodes à base de connaissances donnent de meilleurs résultats par rapport à méthodes à base de corpus ([LSA] dans ce cas). La méthode [W&P] à base de distance taxonomique est considérée meilleure que la méthode à base de contenu informationnel [J&C]. Dans les deux cas WordNet était utilisé comme une ressource lexicale pour supporter le calcul de la similarité à base de connaissances. Cette étude était appliquée dans trois domaines différents (Transport, Livre et Affaire). Deux espaces sémantiques étaient utilisés pour le calcul de la méthode [LSA] : General Reading et Encyclopedia. Dans cette étude, les résultats de pertinence obtenus pour les trois domaines sont différents. Nous constatons ainsi l'impact du choix du domaine analysé et de l'espace sémantique choisi sur les résultats de calcul de la similarité. Cependant, toutes ces recherches n'ont pas proposé une solution pour le choix de l'espace sémantique par rapport au domaine analysé.

À travers ces études, nous constatons que les résultats obtenus divergent et manquent de consistance. Elles restent tributaires de trois facteurs essentiels :

- La nature et la taille de l'espace sémantique choisi pour les méthodes à base de corpus. La pertinence de ces méthodes s'améliore en choisissant un espace sémantique large et proche du domaine analysé.
- Le domaine objet de l'analyse de la similarité sémantique textuelle. Les résultats se distinguent d'un domaine à un autre.
- L'outil adopté pour mesurer la distance sémantique entre mots à base de connaissances. Les outils sont divers et dotés de méthodologies différentes. La plupart d'entre eux utilisent la base lexicale WordNet pour mesurer la distance entre les mots. Dans ce travail de recherche, nous avons choisi l'outil WS4J [Wittek et al., 2015] basé sur WordNet.

3 EXPERIENCES ET RESULTATS

Dans la littérature, le corpus utilisé comme un espace sémantique pour les méthodes à base de corpus diffère du domaine analysé. Nous estimons que le choix d'un espace sémantique proche du domaine analysé donnera des résultats plus pertinents. Pour cela, nous utilisons dans cette section la base de données des accidents avioniques canadiens : CADORS. Ce corpus sera le domaine d'étude pour effectuer les analyses de similarité textuelle entre les différents enregistrements. Il sera utilisé aussi, comme un espace sémantique pour calculer la similarité entre les mots en utilisant les méthodes à base de corpus [LSA] et [NPMI]. Les résultats obtenus sont comparés aux résultats des mêmes méthodes appliquées dans des espaces sémantiques différents du domaine analysé. Ils sont aussi comparés aux résultats obtenus en utilisant les méthodes à base de connaissances [W&P] et [LIN]. Nous répondons à travers cette étude à la question du choix de l'espace sémantique adapté aux analyses de la similarité des incidents aéronautiques consignés sous formes de textes courts et dans un langage très technique.

Trois expériences sont effectuées afin de prouver l'avantage du choix des méthodes à base de corpus dans un espace sémantique identique au domaine analysé.

- La première permet de comparer les méthodes à base de corpus [LSA] et [NPMI] appliquées dans un espace sémantique identique au domaine analysé avec les méthodes à base de connaissances [W&P] et [LIN].
- La deuxième expérience permet de comparer la méthode à base de corpus [LSA] appliquée dans un espace sémantique identique au domaine analysé (incidents aéronautiques) avec la même méthode appliquée dans deux autres espaces sémantiques différents.
- La troisième expérience permet de comparer les méthodes à base de corpus [LSA] et [NPMI] appliquées dans un espace sémantique identique au domaine analysé avec trois tailles différentes.

3.1 Description des expériences

Nous avons extrait 40 000 incidents à partir de la base de données CADORS. Nous avons opté pour des enregistrements d'incidents/accidents survenus sur les moteurs d'avions fabriqués par notre partenaire industriel. La taille d'échantillon pour l'analyse des performances des différentes méthodes est 381 rapports parmi les 40 000 rapports afin, d'assurer un niveau de confiance de 95% et un intervalle de confiance de 5%. Une catégorisation manuelle était réalisée à l'aide de deux experts séparément pour identifier parmi les 40 000 rapports ceux qui sont jugés similaires aux 381 rapports. Les deux experts avaient les mêmes jugements sur 91% des rapports analysés. Ces jugements nous permettent de construire la matrice de référence S_{Ref} à partir de laquelle toutes les analyses de pertinence sont réalisées.

Dans un premier temps, un traitement de données est effectué pour fiabiliser les calculs de la similarité. Les textes extraits sont indexés par des numéros uniques pour faciliter les analyses. Par la suite, les textes sont téléchargés pour traitement à l'aide du logiciel Wordstat [Pollach, 2010]. Les textes analysés sont écrits dans la langue anglaise. Un traitement de lemmatisation est réalisé à l'aide de l'application Wordstat pour ramener les différents mots à leurs termes racines. Cette opération permet de fusionner l'occurrence de plusieurs mots de même racine. Le logiciel Wordstat permet d'identifier les mots non reconnus par le dictionnaire. La correction de ces mots doit être faite manuellement afin de fiabiliser le calcul d'occurrence des mots dans les textes. Dans le domaine d'aéronautique, plusieurs abréviations sont utilisées pour exprimer une fonction, un processus ou un service. Ces mots doivent être maintenus même s'ils ne sont pas reconnus par le dictionnaire. Cette tâche nécessite un travail manuel pour différencier les erreurs d'écriture par rapport aux abréviations. Les mots génériques doivent être supprimés aussi du calcul de la similarité. L'application Wordstat est dotée d'une base de données incluant tous les mots génériques. Cette liste peut être modifiée en ajoutant ou supprimant des mots.

L'outil Wordstat peut être paramétré pour exclure les mots qui se répètent à une fréquence élevée. Dans notre cas, nous avons exclu tous les mots qui existent dans plus que 80% de textes. Par exemple, le mot "Ranway" est un mot fréquent dans les textes extraits, sa présence n'améliore pas le calcul de la similarité. Les abréviations sont parfois écrites avec leurs significations. Nous avons opté pour supprimer les significations et garder les abréviations (exemple : flight service station (FSS)).

Une fois la liste des mots est définie, nous procédons à la première expérience par le calcul des matrices de la similarité entre mots : A_{LSA} , A_{NMPI}, $A_{W\&P}$ et A_{LIN}, en utilisant les quatre méthodes [LSA] , NMPI, [W&P] et [LIN]. Les deux matrices A_{LSA} et A_{NMPI} sont carrées de dimensions (N1) (voir Tab-3). Les deux matrices $A_{W\&P}$ et A_{LIN} sont carrées de dimensions (N2) (voir Tab-3).

Les matrices A_{LSA} , A_{NMPI} sont calculées en utilisant le corpus des 40.000 enregistrements comme un espace sémantique. L'outil SCILAB est utilisé pour effectuer le calcul mathématique. Les deux matrices $A_{W\&P}$ et A_{LIN} sont calculées en utilisant l'outil WS4J avec la base lexicale WordNet.

Les matrices de cooccurrence des termes dans les différents textes sont calculées en utilisant l'outil Wordstat. Les deux matrices T_{LSA} , T_{NMPI} sont identiques de dimension 40000*N1. Les deux matrices $T_{W\&P}$, T_{LIN} sont identiques en contenue et sont de dimensions 40000*N2.

Ces matrices sont intégrées par la suite au modèle GVSM (Éq2) pour calculer les matrices de de la similarité globale $S_{Global/LSA}$, $S_{Global/NMPI}$, $S_{Global\ /\ w\&p}$ et $S_{Global/LIN}$. Ce sont des matrices carrées de dimensions 40000 par 40000. Des sous-matrices de dimensions 381*40000 sont extraites à partir de ces matrices S_{Sample/LSA_40000}, $S_{Sample/NMPI\_40000}$, $S_{Sample\ /\ w\&p}$ et $S_{Sample/LIN}$. Ces matrices sont transformées par la suite à des matrices booléennes SB_{Sample/LSA_40000}, $SB_{Sample/NMPI\_40000}$, $SB_{Sample\ /\ w\&p}$ et $SB_{Sample/LIN}$, dont les valeurs supérieures ou égales à 0.6 sont arrondies à 1 et les valeurs inférieures à 0.6 sont ramenées à zéro [Godoy et Amandi, 2006]. Un travail d'analyse de sensibilité autour de la valeur 0.6 sera fait dans des travaux ultérieurs. Ces matrices sont comparées à la matrice de référence construite à partir des jugements des experts S_{Ref}. Le résultat de cette comparaison permet de calculer les paramètres a, b et c, et par la suite, calculer les indicateurs de pertinences P, R et F.

Pour la deuxième expérience, nous avons utilisé l'outil en ligne http://[LSA] .colorado.edu/ pour calculer la similarité entre les mots identifiés (N1) en utilisant la méthode [LSA]. Nous avons utilisé les deux espaces sémantiques : General_Reading_up_to_03rd_Grade et Cognit. De la même manière décrite dans la première expérience, les résultats finaux seront trois matrices booléiennes SB_{Sample/LSA_40000}, SB_{Sample/LSA_GR}, et SB_{Sample/LSA_Cognit}. Ces matrices sont comparées à la matrice de référence construite à partir des jugements des experts S_{Ref}. Le résultat de cette comparaison permet de calculer paramètres a, b et c, et par la suite, calculer les indicateurs de pertinences P, R et F.

Pour la troisième expérience, nous avons calculé les matrices de similarité entre les mots en utilisant les méthodes à base de corpus dans un espace sémantique identique au domaine analysé. Nous avons choisi trois tailles différentes 40000, 30000 et 20000. De la même manière décrite dans la première expérience, les résultats finaux seront trois matrices booléiennes SB_{Sample/LSA_40000}, SB_{Sample/LSA_30000}, et SB_{Sample/LSA_20000}. Ces matrices sont comparées à la matrice de référence construite à partir des jugements des experts S_{Ref}. Le résultat de cette comparaison permet de calculer paramètres a, b et c, et par la suite, calculer les indicateurs de pertinences P, R et F.

3.2 Exemple

Nous avons choisi un exemple parmi l'échantillon étudié pour décrire d'une manière simple les trois expériences de ce travail. Nous avons choisi quatre enregistrements indexés E1, E2, E3 et E4 (Tableau 1).

Tableau 1. Textes choisis pour illustrer l'expérience

Events	Textes
E1	During the initial climb, the no. 2 hydraulic system completely failed when it was time to retract the landing gear. The crew declared an emergency. The maintenance team found a leak in a hydraulic line attached to the hydraulic pump. The no. 2 hydraulic system line and pump were replaced
E2	According to information from the Aircraft Maintenance Division, the No. 2 hydraulic system line and pump were replaced. Declared an emergency when it experienced a hydraulic failure at landing gear retracting.
E3	Immediately after takeoff, an Air Inuit reported a pressurization problem and that it was unable to climb to 17 000 ft. It did not declare an emergency and continued to its destination at an altitude of 8 000 ft.
E4	A Porter Airline aborted takeoff runway 32 due to a bird strike. Runway inspected and snowy owl recovered. Some departures and arrivals were delayed due to the inspection of the Runway

Par la suite, nous avons défini la liste des mots nécessaires. Cette liste est obtenue après l'élimination des chiffres, termes génériques et mots avec fréquences élevées. Ensuite, un traitement de lemmatisation est fait sur ces mots en utilisant l'outil Wordstat. Nous avons calculé par la suite la matrice de similarité entre mots, Pour illustrer la différence entre les différentes méthodes utilisées pour calculer la similarité entre les deux mots (Landing, Gear). La valeur de la similarité Sim (Gear, Landing) varie entre 0.1 et 0.88 en fonction de la méthode utilisée.

Une fois les matrices de similarité entre les mots sont calculées en utilisant les différentes méthodes, elles sont intégrées au modèle GVSM (Éq2) avec les matrices de cooccurrences pour calculer la similarité entre les enregistrements E1, E2, E3 et E4. Les résultantes sont transformées à des matrices booléiennes en comparant chaque élément des matrices de similarités à la valeur-limite 0.6). Par la suite, chaque élément (i,j) de la matrice de la similarité est comparé à l'élément (i,j) de la matrice du jugement des experts S_{Ref}. Dans cet exemple, les experts jugent que les seuls enregistrements similaires sont E1 et E2. Ce jugement sera utilisé pour le calcul des critères de pertinence afin de pouvoir comparer les différentes méthodes.

3.3 Synthèse des résultats et discussion

Le tableau (Tableau 2) donne les résultats de pertinence en utilisant la méthode GVSM combinée avec les deux méthodes [LSA], [NPMI] appliqués dans un corpus d'incidents aéronautiques de dimension 40000, et avec les deux méthodes [W&P] et [LIN]. Selon le critère de la précision (P), les méthodes à base de corpus donnent des résultats nettement supérieurs par rapport aux méthodes à base de connaissances. Ce résultat peut être expliqué par deux éléments.

Le premier élément est lié au langage utilisé dans le domaine analysé. Malgré les différents canaux d'émission, Les textes sont généralement écrits de la même manière en utilisant des mots identiques. Les combinaisons standards utilisés dans le domaine aéronautique (ex : Fuel leakage) permettent aux méthodes à base de corpus d'identifier plus de résultats pertinents que les méthodes à base de connaissance. Ces aspects permettent de réduire les effets des deux phénomènes : la polysémie et la synonymie. Cependant, les mots utilisés pour l'analyse de la similarité à base de connaissances se trouvent proches dans la hiérarchie de la structure lexicale WordNet. En effet, certains enregistrements non pertinents sont retrouvés comme des enregistrements similaires.

Tableau 2 : Résultats de calcul des cirières P, R et F pour l'échantillon

Approches	Méthode	P	R	F
À base de corpus	NPMI	74.1%	88.7%	80.6%
	LSA	73.5%	87.2%	79.7%
À base de connaissance	W&P	58.0%	94.7%	72.0%
	LIN	56.2%	93.2%	70.1%

Le deuxième élément est lié aux abréviations, le domaine de l'aéronautique utilise des abréviations qui ne peuvent être analysées qu'avec les méthodes à base de corpus (par exemple IFR : Instrument Flight Rules). Ces abréviations n'ont aucune signification lexicale. Les méthodes à base de corpus ont l'avantage d'analyser la similarité textuelle en tenant compte des abréviations spécifiques du domaine. Le ratio des abréviations est de 15% (Tableau 4) par rapport au total des mots utilisés dans le calcul de la similarité. L'élimination de ces mots du calcul de similarité en utilisant les méthodes à base de connaissance réduit la dimension des mots et favorise ainsi les méthodes à base de corpus.

Selon le critère du rappel (R), il est évident que les méthodes à base de connaissances permettent de retrouver plus de résultats pertinents par rapport aux méthodes à base de corpus puisque les mots utilisés se trouvent proches dans la hiérarchie lexicale.

Le critère de la F-mesure donne avantage aux méthodes à base de corpus par rapport aux méthodes à base de connaissances. Les méthodes à base de corpus appliquées dans le domaine aéronautique en l'utilisant aussi comme un espace sémantique sont 10% meilleures que les méthodes à base de connaissances [W&P] et [LIN]. La différence entre les deux méthodes [LSA] et [NPMI] est négligeable.

Le tableau Tab-7 donne les résultats de pertinence des méthodes à base de corpus appliquées dans le domaine aéronautique en utilisant trois (3) espaces sémantiques différents: Cognition, Lectures générales et incidents aéronautiques. La pertinence est 10% meilleure si l'espace sémantique choisi est similaire au domaine analysé. Les résultats sont plus pertinents si le corpus de l'espace sémantique choisi converge vers le corpus du domaine analysé. Le fait d'opter pour un espace sémantique similaire au domaine analysé réduit les effets de la polysémie puisque les mots ont généralement la même signification. En plus les enregistrements sont écrits en utilisant les mêmes mots techniques ce qui réduit aussi l'effet de la synonymie.

Tableau 3 : Résultats de calcul des cirières P, R et F pour les trois espaces sémantiques appliqués à [LSA]

Domaine	Espace sémantique	P	R	F
Incidents aéronautiques	Cognition	56%	95%	70.5%
Incidents aéronautiques	General_Reading_up_to_03rd_Grade	55%	95%	69.7%
Incidents aéronautiques	Incidents aéronautiques	73.5%	87.2%	79.8%

Nous avons choisi trois tailles différentes du corpus analysé pour mesurer l'effet sur les résultats de la similarité en utilisant les méthodes à base de corpus [LSA] et [NPMI]. Le nombre des nouveaux mots identifiés a augmenté de 12% alors que le nombre des rapports a doublé de 20.000 à 40.000 (Tableau 4). Cela est dû au langage technique très limité utilisé pour rédiger les rapports d'incidents aéronautiques. Le nombre total des mots utilisés pour le calcul de la similarité sémantique à base de corpus a une tendance vers une stabilité à une taille maximale. Nous constatons aussi que les résultats se dégradent graduellement si la taille de l'espace sémantique diminue (Tab-9). Quoique la pertinence se soit dégradée juste de 4% en changeant la taille de l'espace sémantique entre 40000 et 20000, le choix d'un corpus large reste souhaitable pour améliorer la pertinence des résultats.

Tableau 4 : Nombre de mots selon la taille de l'espace sémantique

Tailles du corpus	Mots	Abréviation	Total
20000	N6=7240	1245	N5=8485
30000	N4=7952	1315	N3=9267
40000	N2=8240	1389	N1=9629

Tableau-9 : Résultats de calcul des critères P, R et F pour [LSA] et [NPMI] en fonction de la taille du corpus

Paramétres	Tailles du corpus	NPMI	LSA
P	20000	69.77%	69.23%
	30000	72.00%	71.30%
	40000	74.13%	73.52%

R	20000	85.71%	83.33%
	30000	87.05%	86.57%
	40000	88.47%	87.20%
F	20000	76.92%	75.63%
	30000	78.81%	78.20%
	40000	80.67%	79.78%

3.4 Application

La base de données de 40.000 événements contient plusieurs incidents avec différents niveaux de criticité. Quotidiennement, les avions ratent des atterrissages, annulent des décollages, ou heurtent des oiseaux. Les techniques de traitement sémantiques telles que la méthode proposée ci-haut, permettent le calcul de la fréquence des incidents d'une manière non supervisée et instantanément. Les incidents fréquents ou émergents sont facilement détectables pour prendre les mesures nécessaires. Une catégorisation peut être faite plus tard par type d'aéronefs, de moteurs, de régions ou de compagnies d'aviation.

Sur les figures ci-dessus, la méthode proposée est appliquée sur deux types d'incidents différents. Le premier incident est le problème des systèmes d'atterrissages d'avions, qui est considéré comme le problème le plus constaté sur les données sélectionnées (0,5%). Le deuxième incident est le problème des passagers malades, ce qui est très rare, mais peut causer l'annulation de vols, ou un atterrissage d'urgence (0,015%).

Sur la figure2, nous constatons que la plupart des événements ont moins de 0,07, tandis que pour le second figure1, la plupart des événements ont un taux égal à zéro. Ceci peut être expliqué par le fait, que le deuxième incident traite un problème technique ; plusieurs termes techniques peuvent être utilisés dans d'autres incidents non semblables.

Un FMEA dynamique sera proposé dans le prochain travail. Le calcul de l'occurrence d'un nouvel incident enregistré est effectué en utilisant la méthode NPMI appliqué dans le même corpus, et combinée avec la méthode GVSM.

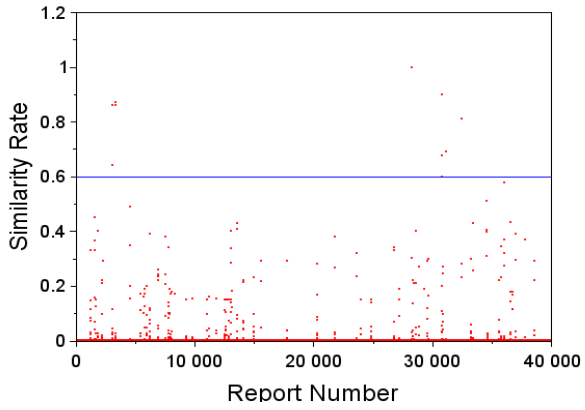

Figure 1. Resultats de similarité du probléme : landing gear GVSM (NPMI_40000)

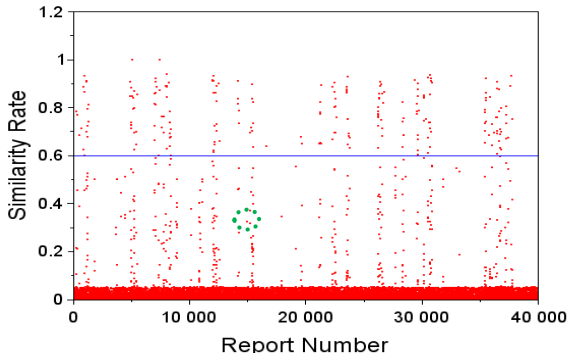

Figure 2. Resultats de similarité du probléme : sick passenger GVSM (NPMI_40000)

4 CONCLUSION

Cet article a permis de mettre en évidence la pertinence des méthodes à base de corpus ([LSA] et [NPMI]) appliquées dans le domaine d'incidents aéronautiques en l'utilisant aussi comme un espace sémantique. Les résultats sont 10% meilleurs que les méthodes à base de connaissances. L'utilisation des abréviations dans le domaine d'aéronautique (15% des mots) donne avantage aux méthodes à base de corpus. Les résultats sont aussi 10% meilleurs par rapport aux méthodes à base de corpus en utilisant des espaces sémantiques différents du domaine analysé. Cela explique que notre proposition permet de réduire les effets de la polysémie et la synonymie sur les calculs de la similarité. Les résultats de pertinence s'améliorent de 4% si la taille de l'espace sémantique augmente de 20000 à 40000 enregistrements. Cette amélioration s'avère faible par rapport à une taille doublée. Cela est dû au langage technique limité utilisé dans les textes des incidents. La taille de l'espace sémantique reste un élément-clé pour améliorer la pertinence des résultats de la similarité. Les résultats de ce travail seront utilisés prochainement dans une modélisation dynamique de risque afin d'actualiser les analyses de risques préliminaires en connectant les événements et les reliant à leurs analyses de risques préétablies.

5 REFERENCES

Banerjee, S., et Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. *Computational linguistics and intelligent text processing*, Springer**:** 136-145.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*: 31-40.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference*.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database, *MIT Press*.

Godoy, D., et Amandi, A. (2006). Modeling user interests by conceptual clustering. *Information Systems* 31(4): 247-265.

Jiang, J. J., et Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Landauer, T. K., Foltz, P. W., et Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes* 25(2-3): 259-284.

Leacock, C., et Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database* 49(2): 265-283.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of the 5th annual international conference on Systems documentation*, ACM.

Li, H., Tian, Y., Ye, B., et Cai, Q. (2010). Comparison of current semantic similarity methods in wordnet. *Computer Application and System Modeling (ICCASM), 2010 International Conference on*, IEEE.

Lin, D. (1998). An information-theoretic definition of similarity. *ICML*.

Mihalcea, R., Corley, C., et Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *AAAI*.

Mili, A., Bassetto, S., siadat, A., et Tollenaere, M. (2009). Dynamic risk management unveil productivity improvements. *Journal of Loss Prevention in the Process Industries*: 25-34.

Mili, A., Siadat, A., Hubac, S., et Bassetto, S. (2008). Dynamic management of detected factory events and estimated risks using FMECA. *2008 IEEE International Conference on Management of Innovation & Technology (ICMIT 2008), 21-24 Sept. 2008*, Piscataway, NJ, USA, IEEE.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)* 41(2): 10.

Pirró, G., et Euzenat, J. (2010). A feature and information theoretic framework for semantic similarity and relatedness. *The Semantic Web–ISWC 2010*, Springer**:** 615-630.

Pollach, I. (2010). Software Review: WordStat 5.0. *Organizational Research Methods*.

Rada, R., Mili, H., Bicknell, E., et Blettner, M. (1989). Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on* 19(1): 17-30.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Resnik, P. (2011). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *arXiv preprint arXiv:1105.5444*.

Salton, G., Wong, A., et Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM* 18(11): 613-620.

Shrestha, P. (2011). Corpus-based methods for short text similarity. *Rencontre des Étudiants Chercheurs en Informatique pour le Traitement automatique des Langues* 2(1).

Slimani, T., BenYaghlane, B., et Mellouli, K. (2007). Une extension de mesure de similarité entre les concepts d'une ontologie. *the Proceedings of SETIT*: 1-10.

Tulechki, N. (2011). Des outils de TAL en support aux experts de sûreté industrielle pour l'exploitation de bases de données de retour d'expérience. *Actes des 13èmes Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2011)*.

Tulechki, N., et Tanguy, L. (2012). Effacement de dimensions de similarité textuelle pour l'exploration de collections de rapports d'incidents aéronautiques. *Actes de la conférence annuelle du Traitement Automatique des Langues Naturelles (TALN) 2012*.

Tulechki, N., et Tanguy, L. (2013). Similarité de second ordre pour l'exploration de bases textuelles multilingues. *Actes de la 20e conférence du Traitement Automatique du Langage Naturel (TALN)*.

Tumer, I. Y., et Stone, R. B. (2003). Mapping function to failure mode during component development. *Research in Engineering Design* 14(1): 25-33.

Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL.

Tversky, A. (1977). Features of similarity. *Psychological Review* 84(4): 327-352.

Wicaksana, I., et Wahyudi, B. (2011). Comparison Latent Semantic and WordNet Approach for Semantic Similarity Calculation. *arXiv preprint arXiv:1105.1406*.

Wilks, Y., et Stevenson, M. (1998). Word sense disambiguation using optimised combinations of knowledge sources. *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, Association for Computational Linguistics.

Wittek, P., Darányi, S., Kontopoulos, E., Moysiadis, T., et Kompatsiaris, I. (2015). Monitoring Term Drift Based on Semantic Consistency in an Evolving Vector Field. *arXiv preprint arXiv:1502.01753*.

Wong, S. M., Ziarko, W., et Wong, P. C. (1985). Generalized vector spaces model in information retrieval. *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM.

Wu, Z., et Palmer, M. (1994). Verbs semantics and lexical selection. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics.