Veille d'image à partir d'un corpus journalistique paramétrable : le *Trophée Jules Verne*

Frédéric Pierron[1], Natalia Grabar[2], Grégory Pelletier[1]

[1]The Blast Machine, 66 rue Marceau, 93100 Montreuil (fpierron@theblastmachine.com, gpelletier@theblastmachine.com)

[2]Université Paris Descartes, Faculté de Médecine ; Inserm, U729 ; SPIM, Paris, 75006 France (natalia.grabar@spim.jussieu.fr)

Abstract

In this paper, we present the work which aims at the image watch. This work is realized on the basis of journalistic corpora (articles, radio and TV programs). We describe first the composition of the corpus and possibilities offered by the graphical interface to specify its composition and structuring. We present then some results acquired from processed corpora. We finish with conclusions and implications for the image watch.

Résumé

Dans cet article, nous présentons un travail qui est consacré à la veille d'image. La veille est effectuée à partir d'un corpus journalistique (articles de presse, émissions radio et télévisées). Nous décrivons d'abord le contenu et les possibilités de paramétrage du corpus de travail. Nous présentons ensuite quelques résultats acquis sur ce corpus et tirons des conclusions pour l'image de l'entreprise.

Mots-clés : Veille d'image, corpus journalistique, extraction d'information, analyse factorielle, facteur de vraisemblance

1. Introduction

La veille d'image concerne la recherche et le traitement de renseignements relatifs à l'image, et donc à la notoriété d'une entreprise ou d'une marque. La veille peut faire objet d'un intérêt constant de la part d'une entreprise, qui est alors attentive aux rumeurs, aux mécontentements, à ce qui se dit sur les forums de discussion ou sur les sites d'avis de consommateurs. La veille peut aussi être motivée par un événement ponctuel, ou plus ou moins ponctuel, comme par exemple une campagne de communication, un communiqué de presse, le sponsoring, ou autre. Dans ce dernier cas, la veille permet surtout de mesurer l'impact effectué par cet événement. Lorsqu'il s'agit du sponsoring des clubs sportifs, des sportifs individuels, des navigateurs ou autres, les entreprises s'intéressent de savoir si cette action a influencé leur image et si oui, de quelle manière. Les questions posées peuvent alors être de l'ordre très général et, en cas de résultats nuls ou négatifs, la décision pourra être prise d'arrêter le sponsoring :

> Parle-t-on du club subventionné ? Parle-t-on de nous à la même occasion ? Si on ne parle pas de nous, pourquoi ? Que dit-on du club alors ? Si on parle de nous, que dit-on ? L'image change-t-elle au fil du temps ? Si elle change, comment évolue-t-elle ?

Les questions posées peuvent aller plus loin et chercher à cerner un personnage donné et son impact journalistique : personne sponsorisée, encadrants du club, les joueurs, les proches du club, etc. Les réponses à ces questionnements peuvent peser ensuite sur les **décisions de** gestion du personnel, de recrutement, etc.

Dans ce travail, nous nous intéressons à analyser l'image du groupe CapGemini[1], sponsor principal du navigateur Olivier de Kersauzon et de son bateau Geronimo. L'analyse vise surtout à étudier l'impact de ce sponsoring, en particulier suite à la participation d'Olivier de Kersauzon au challenge Trophée Jules Verne.

Le Trophée Jules Verne est un challenge nautique qui récompense le tour du monde le plus rapide réalisé en équipage, sans escale et sans assistance. Le Trophée est né dans les oeuvres de Jules Verne et de l'idée de tenter vraiment de faire le tour du monde à la voile en moins de 80 jours. Cette idée est lancée en 1985 par un marin, Yves Le Cornec. En 1990, une dizaine de navigateurs se rassemblent à Paris afin de définir les règles du jeu. Le Trophée Jules Verne est ainsi l'unique récompense du navigateur qui aura amélioré le record du tour du monde à la voile. Il conserve le Trophée jusqu'à ce que son record soit battu, auquel cas le Trophée est transmis au nouveau recordman. Pour le parcours, il est nécessaire de couper la ligne de départ définie par une ligne imaginaire, reliant le phare de Créac'h sur l'Ile d'Ouessant et le phare du Cap Lizard ; faire le tour du monde en laissant à bâbord (à gauche) le Cap de Bonne Espérance, le Cap Leeuwin et le Cap Horn ; et recouper cette ligne en sens inverse. Tout navire propulsé par la seule force du vent et de l'équipage est autorisé. Le Trophée est ouvert à tout type de bateau sans restriction. Le 20 avril 1993, Bruno Peyron, à la tête de son équipe, boucle le premier tour du monde légendaire en 79 jours et devient ainsi le premier détenteur du Trophée Jules Verne. D'autres navigateurs la gagnent ensuite. Le 29 avril 2004 c'est Olivier de Kersauson et l'équipage du trimaran Geronimo, qui s'empare, pour la deuxième fois, du Trophée en bouclant le périple en 63 jours et 14 heures. Le 16 mars 2005, Bruno Peyron reprend son record en établissant un temps fabuleux de 50 jours, 16 heures et 20 min. sur le catamaran Orange II.

Les événements comme celui-ci sont largement couverts par la presse. C'est également la presse qui assure la création d'une image et la diffuse auprès du grand public. Pour la veille de l'image de la société CapGemini, nous nous concentrons ainsi sur l'analyse du discours journalistique. Nous cherchons alors à observer ce que disent les journalistes au sujet des événements et des personnages qui nous intéressent, mais aussi ce que disent les personnages sponsorisés ou impliqués eux-mêmes. Nous voulons en particulier démontrer que la manière dont on constitue et manipule les données textuelles permet de relever, avec des méthodes d'analyse constantes, différents types d'informations dans les corpus. Si la fiabilité des méthodes assure la validité des résultats, le paramétrage du matériel textuel fait ressortir les faits différents de ces corpus.

Dans la suite de cet article, nous présentons d'abord le corpus journalistique qui nous a permis de faire les observations (sec. 2). La base de données où est stocké le corpus, et surtout son interface d'accès et d'export, permettent des fonctionnalités variées et intéressantes pour la compilation du et des corpus paramétrables selon les critères choisis : média, périodique, date, genre, auteur, etc. Nous présentons ensuite les méthodes d'analyse de corpus (sec. 3) et discutons quelques résultats obtenus (sec. 4). Nous terminons avec une conclusion (sec. 5).

[1] CapGemini est une société d'origine française spécialisée en conseil et en services informatiques. Pour plus d'information, voir le site *www.capgemini.com/*.

2. Matériel : un corpus journalistique paramétrable

Dans cette section, nous nous attachons à décrire le corpus journalistique *Trophée Jules Verne* à travers les points suivants que nous développerons :

1. Recrutement d'articles pertinents ;
2. Composition de la base d'articles ;
3. Interface de visualisation et d'export et ses différentes fonctionnalités ;

2.1. Recrutement d'articles pertinents

La détection d'articles pertinents est faite grâce à la participation, lors des étapes précédant à notre travail, de l'agence Argus Presse. Cette société se charge de choisir, dans de nombreux périodiques, à la radio et télévision, les articles et émissions qui concernent l'événement ciblé : le Trophée Jules Verne et le navigateur Olivier de Kersauzon. Ces articles et émissions peuvent donc provenir de différents médias. Lors du travail avec ce matériel, nous en avons distingué quatre : presse, radio, télévision et web. Les articles fournis par cette agence sont au format papier, tandis que les émissions radio et TV sont sous forme d'enregistrements sonores gravés sur des CD-ROM. Pour pouvoir exploiter l'ensemble de ces documents avec des outils d'accès au contenu, la première étape consiste à les numériser et à les mettre au même format. La numérisation est faite grâce à la scannérisation, à la saisie manuelle et au téléchargement des articles disponibles sur le web. Le matériel numérisé est ensuite stocké dans une base de données. Nous présentons d'abord le contenu de la base d'articles. Ensuite, nous décrivons les fonctionnalités de l'interface web à travers laquelle ce matériel est accessible.

2.2. Composition de la base d'articles

L'ensemble de ces articles, convertis sous forme textuelle, correspond à notre matériel de travail. Nous parlerons alors de *corpus*, car ce matériel semble satisfaire les critères définis dans (Habert et al., 1997) : c'est une collection de données langagières (1) sélectionnées selon leur thématique, qui est concentrée autour de la participation d'Olivier de Kersauzon au Trophée Jules Verne, (2) organisées explicitement selon les dimensions externes (sec. 2.2.1) et internes (sec. 2.2.2), que nous présentons ci-dessous. Une telle organisation rend les données paramétrables. (3) Et enfin, ces données nous servent d'échantillon du discours journalistique sur les événements que nous étudions.

Les données langagières de ce corpus sont donc caractérisées selon les dimensions internes et externes (Habert et al., 2001). Certains de critères qui composent ces dimensions proviennent des travaux antérieurs (Biber & Finegan, 1994 ; Sinclair, 1994 ; dub, 1999), d'autres ont été ajoutés spécifiquement dans le travail cité. Ces deux dimensions sont interliées, mais leurs critères apportent des informations différentes sur les données. Ainsi, la dimension externe est relative au contexte de production des articles. Elle englobe le ou les auteurs, le support, la date de création et/ou de parution, la taille de l'article en occurrences ou octets, le cadre de production, le mode de transmission, le type de destinataires, les objectifs de la parution, etc. Tandis que la dimension interne caractérise les articles plutôt par leur contenu. Elle concerne le niveau de style, la personnalisation, la technicité, etc. Les articles et émissions du corpus *Trophée Jules Verne* ne sont pas tous caractérisés pas l'ensemble de critères prévus. Ce sont essentiellement les critères de la dimension externe qui sont spécifiés. Mais d'autres critères, étant spécifiques au discours journalistique, peuvent être déduits assez facilement. Par exemple, l'objectif de la majorité d'articles consiste à informer, et non à persuader, enseigner

ou imposer. Ce sont les méthodes d'analyse des données textuelles qui nous permettent d'accéder aux éléments de connaissances relatifs aux critères internes.

2.2.1. Dimension externe du corpus

Les chiffres qui suivent caractérisent la composition du corpus *Trophée Jules Verne* :

– Le corpus est composé de 2 088 articles ;

– L'ensemble de ces articles correspond à environ 570 000 occurrences ;

– La date de production des articles s'étend entre mars 2003 et novembre 2004, la période qui correspond à l'agitation journalistique autour de la participation d'Olivier de Kersauzon au challenge Trophée Jules Verne ;

– Les articles sont caractérisés par leur genre journalistique : brève, reportage, interview, etc ;

– Les articles ont été écrits par 350 journalistes ;

– Les articles ont été produits par 191 supports : journaux, revues, émissions télévisées, sites web, etc. ;

– Les supports sont distribués en quatre médias : presse, télévision, radio et web.

Pour résumer, chaque article peut être caractérisé par des dimensions externes suivantes : une date, un journaliste, un genre, un support et un média. Comme nous le décrivons dans la section 2.3, chacune de ces caractéristiques constitue un critère lors de la sélection des articles et de la compilation de corpus. Il est ainsi possible d'extraire des articles parus entre tel et tel mois, des articles publiés par tel support ou tel journaliste, etc.

Le tableau 1 présente la distribution des 2 088 articles, en termes du volume textuel, temporel ou des supports ou entre les médias. Le web, étant statistiquement insignifiant dans cet ensemble, ne figure pas dans le tableau. Par ailleurs, de nombreux articles du web sont en réalité des articles publiés en presse. D'après les chiffres du tableau 1, la presse est majoritaire dans le discours journalistique du corpus *Trophée Jules Verne* : le nombre d'articles publiés par la presse, 1 310, représente 63 % de la totalité. Cette partie du corpus est donc largement supérieure au nombre d'émissions TV et radio, 548 et 230 respectivement. Ce déséquilibre devient même plus important si l'on regarde le pourcentage du nombre d'occurrences (colonne *% occ.*), du volume temporel (colonne *% temps*) ou du nombre de supports impliqués (colonne *% supp.*).

Média	nbArt	% art.	% occ.	% temps	% supp.	durée m./art.
presse	1 310	63	71	70	73	15 mn 46 s
télévision	548	26	20	20	19	11 mn 17 s
radio	230	11	9	9	8	10 mn 55 s

TAB. 1 : Distribution des articles du corpus Trophée Jules Verne.

Le fait d'avoir ramené tout le matériel au format unique, c'est-à-dire le format textuel, a permis d'établir une relation entre la durée temporelle des émissions radio et télévisées et le nombre d'occurrences des articles de journaux, revues, etc. La taille temporelle des articles constitue en soi une information intéressante. Elle indique en particulier le temps que les lecteurs consacrent à la lecture d'un article, et donc à l'impact médiatique de cet article. La dernière colonne, *durée m./art.*, montre que la durée moyenne des articles et des émissions

n'est pas la même. Ce sont encore les articles de la presse auxquels on consacre le plus de temps. Ensuite viennent la télévision et la radio. La durée moyenne aussi élevée montre un réel intérêt des médias pour le Trophée Jules Verne et, dans le cas de ce corpus, en particulier à un de ses navigateurs, Olivier de Kersauzon. Une moyenne aussi élevée augmente également la possibilité de voir apparaître les mentions des sponsors des navigateurs.

2.2.2. Dimension interne du corpus

Parmi les dimensions internes, nous savons que tous les articles sont caractérisés par un même sujet : le Trophée Jules Verne et Olivier de Kersauzon. La technicité des articles dépend des journalistes, selon qu'ils connaissent bien le monde de la navigation ou non. L'interaction avec le public, la personnalisation, etc. dépendent des médias et des genres. Par exemple, dans les émissions radio ou télévisées, les journalistes peuvent s'adresser à leur public plus directement que ne le font les auteurs d'articles de presse. Mais ces critères doivent faire objet d'une étude spécifique.

Par ailleurs, un travail supplémentaire a été fait afin de délimiter dans les articles le discours indirect, c'est-à-dire les citations. Les journalistes recourent ainsi fréquemment aux citations. Ils font parler leurs personnages et utilisent leurs paroles dans les situations suivantes :

– Pour confirmer leurs propres opinions, déjà exprimées ou à venir, avec les paroles de la personne questionnée ;

– Pour faire parler un spécialiste (statisticiens, médecins, PDG, etc.) et faire dire des choses qui dépassent leurs compétences ;

– Pour faire exprimer une opinion qu'ils n'osent pas dire par eux-mêmes.

Cette couche supplémentaire de description permet de distinguer les citations du reste de texte et, éventuellement, étudier la cohérence du discours journalistique et du discours des personnes interviewées.

2.3. Interface d'interrogation et d'export des articles

La figure 1 présente l'interface d'interrogation et d'export des articles. Cette interface comporte deux parties principales :

– La partie supérieure est prévue pour la formulation des requêtes, ce qui permet de sélectionner les articles et ensuite structurer le corpus ;

– La partie inférieure permet de visualiser et parcourir les articles.

Lors de la formulation des requêtes, trois types de fonctions sont prévus :

– *Application de filtres* selon (1) la date, (2) les supports (qui peuvent être de type *presse*, *TV*, *radio* ou *web*), (3) le contenu lexical spécifique (mots-clés à apparaître dans le texte des articles ou dans les titres seulement), (4) d'autres limitations, et finalement (5) la sélection de citations uniquement. Il est également possible de limiter la recherche d'auteurs ou de supports ayant publié au moins *n* articles ;

– *Regroupement des articles* selon les supports, médias, genres, auteurs ou mois ;

– *Formatage de l'ensemble des articles sélectionnés* en fonction du format d'entrée de quelques outils déjà prévus : Hyperbase, Cordial, Lexico et un dernier format qui donne l'ensemble des critères externes sur chaque article (média, support, genre, mois et auteur).

Les résultats de la sélection des articles sont présentés alors dans la partie inférieure de l'interface. Cette option peut être enlevée en cochant sur *Ne pas afficher les résultats*, ce qui rend l'interrogation de la base plus rapide. Par ailleurs, un fichier avec l'ensemble des articles, prêt à télécharger, apparaît dans la partie supérieure droite de la fenêtre (fichier *Hyperbase-group-by-genre-.zip*). Un nom par défaut est donné à ce fichier, mais ce nom peut être modifié par l'utilisateur.

Notons qu'il est également possible de visualiser les articles en ligne. L'interface de visualisation, en plus d'accès aux articles par auteur, support, etc., propose également des présentations graphiques qui montrent la distribution des articles selon les critères habituels.

À partir des articles de la base et grâce aux interfaces utilisateur il est ainsi possible de paramétrer des corpus en choisissant un critère donné. Par exemple, en focalisant sur les périodes et les mois, les articles sont organisés selon l'ordre chronologique de leur parution. Un tel regroupement des articles permettra d'observer les variations et les constantes thématiques ou autres tout au long de la période. Le regroupement des articles en fonction des supports permettra de dégager, par exemple, les groupes de cohésion lexicale des supports et, donc, de sujets abordés et de manières dont les journalistes en parlent. Le fait de pouvoir créer les corpus aux formats d'entrée de quelques outils de lexicométrie et de Traitement Automatique de Langues diminue par ailleurs le travail de préparation de données.

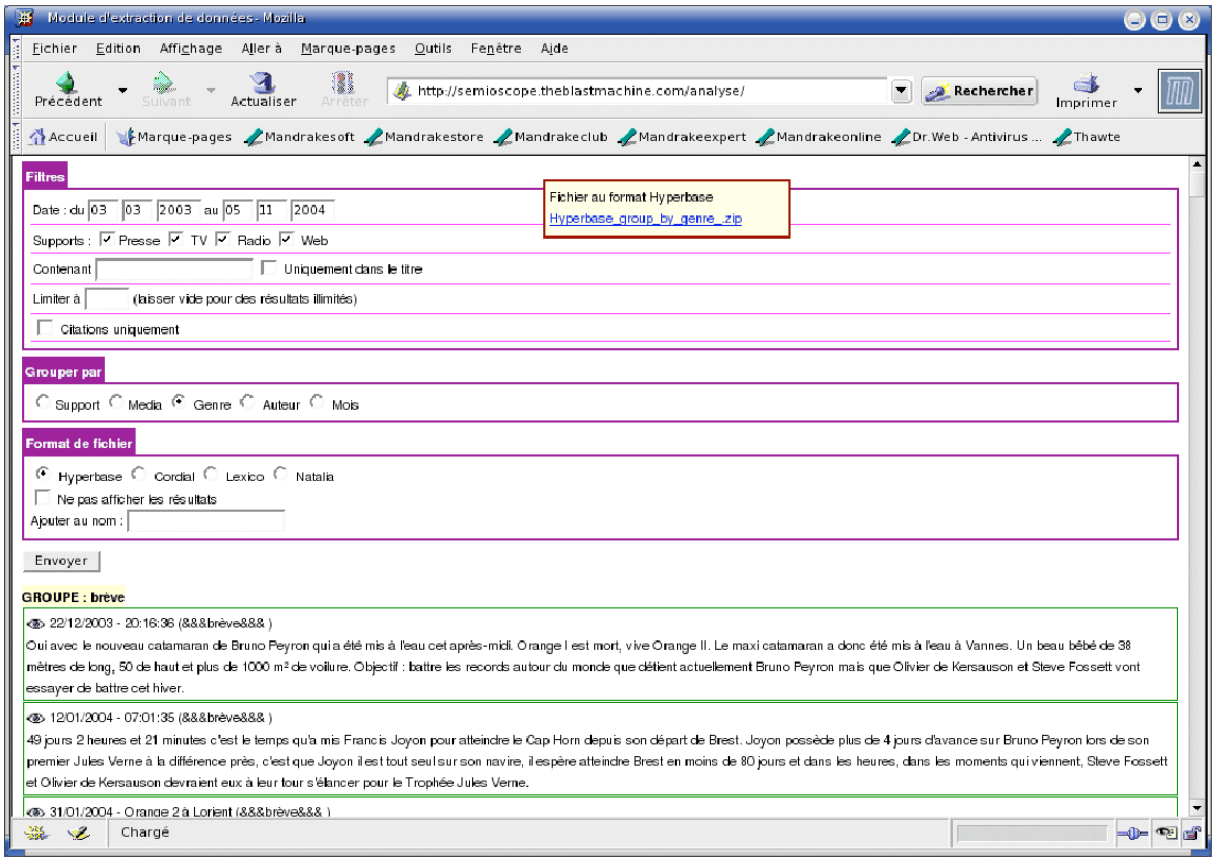

FIG. 1 : Interface d'interrogation et d'export.

3. Méthodes

Pour obtenir les résultats présentés dans la section suivante, nous utilisons les méthodes de fouille de textes et de lexicométrie. Les deux nous permettent de dégager et d'extraire l'information recherchée de l'ensemble des articles analysés.

Nous utilisons ainsi un outil de l'analyse lexicométrique Hyperbase (Brunet, 1988), essentiellement son module de l'analyse factorielle. L'analyse factorielle permet entre autre de calculer les spécificités lexicales d'un texte par rapport au corpus de référence Frantext ou entre les différentes parties de ce texte. Ce module effectue pour ceci l'analyse de correspondance. Les distorsions de fréquences des mots, dues à la taille des données linguistiques, sont atténuées grâce aux calculs de leurs écarts réduits. Ceci permet d'avoir des données pondérées et donc plus facilement comparables entre les corpus ou les parties d'un même corpus[2] . Hyperbase a été principalement appliqué pour l'analyse de corpus littéraires, par exemple l'oeuvre de Balzac (Brunet, 2003a), de Rabelais (Brunet, 2003b), de Rimbaud (Brunet, 2004), et d'autres auteurs français. De manière générale, il a permis de dégager les tendances de l'évolution du vocabulaire de la langue française à travers les romans et nouvelles (Brunet, 1999). Mais son application s'est étendue à d'autres types de textes, par exemple les textes idéologiques (Valette & Grabar, 2004), et à d'autres langues, comme le portugais (Maciel & Brunet, 2000). La fonctionnalité d'Hyperbase qui permet de travailler sur les fins de mot, rend aisée l'analyse des unités morphologiques. Par exemple, en formulant une demande sous la forme *-ique,* on filtrera tous les lexèmes qui finissent en *-ique* et on pourra analyser ensuite ce lexique de même que sa distribution à travers le temps. (Brunet, 1981) analyse ainsi différents affixes du français de 1789 à nos jours et montre les zones d'émergence de certains d'entre eux. En appliquant cet outil aux textes journalistiques, nous cherchons à découvrir, comme dans les travaux cités, les spécificités lexicales propres à ces corpus et à leurs sous-parties. Nous faisons l'hypothèse que ces spécificités feront ressortir les caractéristiques de personnages, journalistes, etc., en fonction du paramétrage des corpus.

Nous établissons également les associations entre les mots des corpus avec le calcul du facteur de vraisemblance. Cet algorithme est souvent utilisé avec les données textuelles pour détecter les associations entre les mots ou syntagmes et pour calculer les voisinages de mots, leurs cooccurrences, les collocations, etc. (Manning & Schütze, 1999). Le facteur de vraisemblance met en concurrence deux hypothèses : deux mots apparaissent séparément dans le texte et sont donc indépendants l'un de l'autre ; deux mots apparaissent souvent ensemble et il existe une dépendance entre eux. Il a été ainsi appliqué aux corpus médicaux pour la détection de lexèmes liés morphologiquement et sémantiquement (Zweigenbaum & Grabar, 2003) et aux corpus biologiques pour la détection de relations entre les gènes et leurs fonctions (Grabar et al., 2005). Dans ce travail, nous l'appliquons pour faire ressortir les spécificités des personnages auxquels nous portons intérêt. Nous l'appliquons aux mots qui se trouvent dans une même phrase ou dans des phrases voisines, soit une fenêtre de 2*30 mots.

[2]Pour plus d'information sur le fonctionnement de cet outil, voir par exemple le site *http ://ancilla.unice.fr/ brunet/pub/hyperbase.html.*

4. Résultats de l'exploration de corpus

4.1. Les marins

La période pendant laquelle Olivier de Kersauzon s'est lancé dans le challenge Trophée Jules Verne, trois autres marins et leurs équipes ont participé : Steve Fosset, Bruno Peyron et Francis Joyon. Ce dernier a navigué en solitaire. La figure 2 montre le positionnement de ces quatre marins et l'émergence de thématiques autour d'eux. Elle est générée avec Hyperbase à partir du corpus d'articles groupés par mois. Elle montre que le positionnement de chaque marin est différent sur la totalité de la période couverte par le corpus. La discussion qui suit est concentrée autour de cette constatation, elle est également nourrie par les résultats du calcul d'associations avec le facteur de vraisemblance.

Avant de parler des navigateurs, nous faisons une remarque sur la nature du Trophée Jules Verne. Le Trophée Jules Verne n'est pas une course ni une compétition. C'est un challenge. Le Trophée a pour principe de mettre face à face un équipage et un record à battre. Par contre, les médias préfèrent de présenter un face à face entre deux équipages. Cette comparaison est d'autant plus aisée que, durant la période étudiée, plusieurs marins se lancent à battre le record du tour du monde à la voile. Dans ce cas, les repères du Trophée Jules Verne servent seulement comme unité de comparaison entre les deux équipages qui ne partent pas en même temps mais qui doivent passer par les mêmes étapes. De manière générale, plus il y a de possibilités de comparer les performances de deux navigateurs, plus il y a d'animation journalistique. Il en est ainsi pour la comparaison entre Olivier de Kersauzon et Bruno Peyron. Mais, lorsque Steve Fosset bat le record, les médias (français) continuent de s'intéresser à Olivier de Kersauzon pour deux raisons :

– Les journalistes accréditent rapidement la thèse selon laquelle Steve Fosset, même s'il a battu le record du monde, n'a pas droit au Trophée Jules Verne car il n'a pas respecté les règles du Trophée : (1) il a refusé de payer les droits de participation qui sont de 30 000 euros et (2) il n'utilise pas la ligne de départ/arrivée notifiée dans les règles du Trophée. Cette situation permet donc de relancer la couverture médiatique jusqu'à l'arrivée d'Olivier de Kersauzon ;

– À ce moment, Olivier de Kersauzon est le seul navigateur encore en course à pouvoir remporter le Trophée Jules Verne.

Les navigateurs sont donc le filtre par lequel le Trophée Jules Verne est présenté par les journalistes. Il va de soi qu'un tel challenge requiert de ses participants de nombreuses qualités et performances. Ceux qui y participent sont certainement les meilleurs. Par contre, il est intéressant de remarquer que, inconsciemment, les journalistes s'attachent à différencier les navigateurs et à leur assigner un archétype puissant et discriminant. Chaque navigateur de cette période est donc perçu d'une manière différente et représente sans doute une facette du « marin idéal ». Dans ce qui suit, nous présenterons les caractéristiques récurrentes des navigateurs.

Steve Fosset est ainsi présenté comme un *aventurier milliardaire*. Son bateau est Cheyenne. Fosset incarne le mythe de l'homme riche, donc libre. Extravagant, car riche, il n'a plus rien à perdre, puisqu'il a tout. Il cherche les extrêmes parce que l'argent lui donne accès à tout, y compris à l'impossible. La notion d'*aventure* est très présente dans ce corpus, par exemple rien qu'à travers ces quelques mots :

aventure, aventurier, aventures, aventureux, s'aventurer, aventuriers, aventurière

Notons que cette notion est particulièrement associée à Steve Fosset. Il est par ailleurs américain et incarne le business.

Bruno Peyron est un *skipper*. Son bateau est Orange. C'est un homme d'action et le technicien de l'eau et de la mer. Lorsqu'il parle, il parle technique, il mesure, il évalue et il explique. Il est armé d'un bateau high-tech, ce qui l'associe à des professionnels. Il a un regard rationnel, sec, voire froid. On lui associe essentiellement des verbes d'action et des mots liés au repérage temporel. Il est souvent identifié comme Peyron le baulien, par opposition à Kersauzon, le Brestois.

Francis Joyon est le *navigateur et homme solitaire*. Son bateau est Idec. C'est un homme parmi les hommes, un homme de mérite et finalement un personnage « touchant ». Il est le *roi* et ce grade est dû à sa peine. Il est aussi désigné comme le *bricoleur* du milieu. À son arrivée, plusieurs journalistes s'étonnent qu'un tel exploit ne soit salué que par quelques dizaines de personnes. Mais Francis Joyon *ne cherche pas les honneurs*, il incarne une forme de simplicité humaine.

Olivier de Kersauzon est le marin par excellence, appelé *Amiral*. Il navigue sur Geronimo, noté Géronimo dans les contextes avec Steve Fosset. C'est un *digne héritier d'Éric Tabarly*. Kersauzon est un breton et il appartient au patrimoine français. C'est le navigateur, l'homme qui comprend le vent, la voile, l'homme qui essaie de dominer la mer. Même si ses propos regorgent de données techniques, sa principale inquiétude est liée à une bonne compréhension des phénomènes naturels qu'il accepte pour mieux en tirer parti. En outre, il est soumis au chronométrage et au temps qui presse : il est poursuivi par les *heures*, *minutes* voire *secondes*. C'est aussi un *aristo*, au sens originel : une personne qui se distingue naturellement du reste des humains. Ayant déjà été le détendeur du Trophée Jules Verne, il est représenté par les journalistes comme l'emblème français. C'est lui qui est opposé à Steve Fosset, entre autre sur l'axe *français* / *américain*. Le parler d'Olivier de Kersauzon est très simple, il est concentré dans la partie droite supérieure de la figure. Il forme une nébulosité à lui tout seul. C'est un parler modeste car constitué de mots très courants et peu recherchés, il est accessible au grand public :

– des pronoms et possessifs : *on, ça, c', notre, quelques* ;

– des argumentatifs : *parce que, donc, quand, c'est-à-dire, comme*, etc.

Même si c'est un aristo, il est très proche du public.

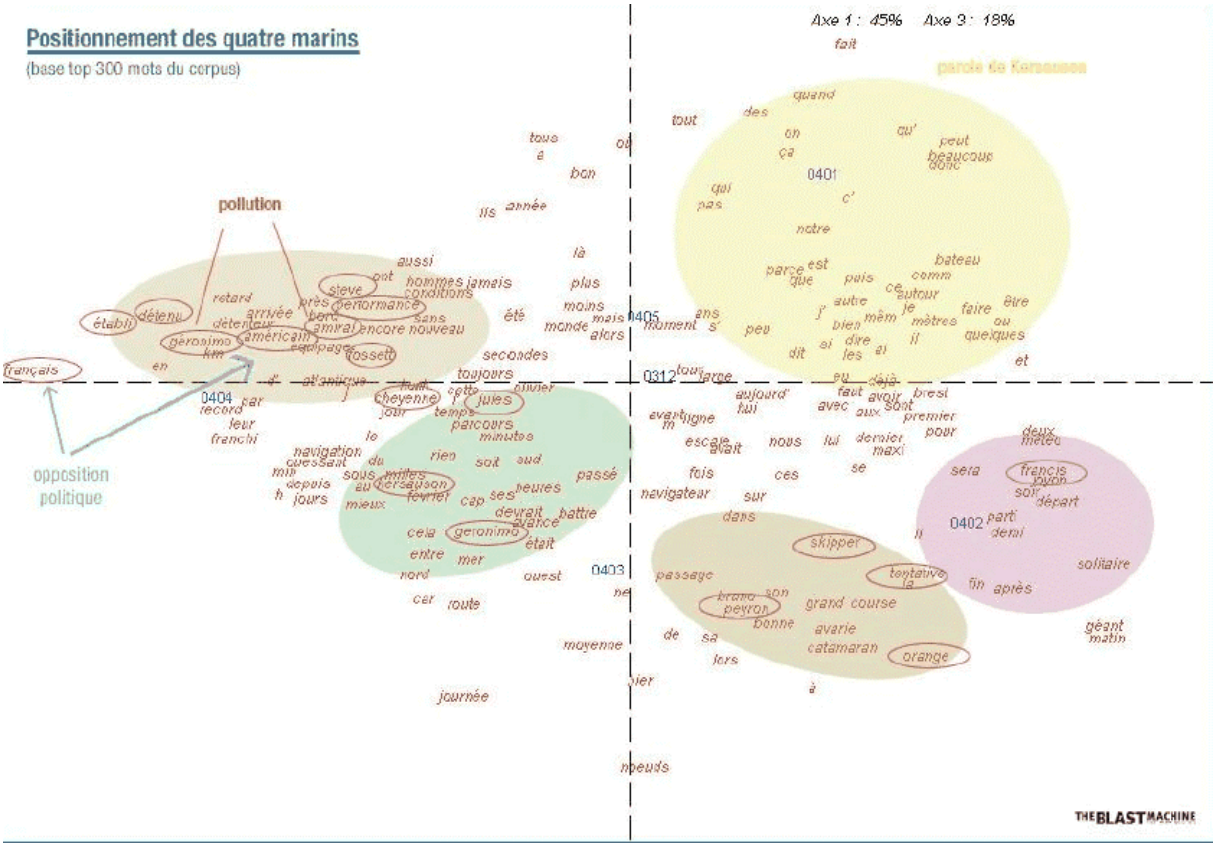

FIG. 2 : Positionnement des quatre marins.

Comme nous l'avons indiqué, la figure 2 a été générée à partir du corpus d'articles groupés par mois. Le fait d'avoir travaillé avec un corpus chronologique, nous a donné les clés pour dégager les personnalités des navigateurs, mais un tel groupement a montré également que chaque mois présente une spécificité lexicale et, surtout, que chaque marin se trouve au centre d'attention des journalistes pendant une période donnée :

– Les préparations de départ d'Olivier de Kersauzon, pendant *0312*, se trouvent au milieu de la figure ;

– En *0401* on le fait beaucoup parler Olivier de Kersauzon ;

– *0402* est centré sur Francis Joyon qui a terminé le tour du monde en solitaire en 72 jours environ ;

– *0403* est une période transitoire entre Bruno Peyron et Olivier de Kersauzon : une nouvelle avarie d'Orange, les icebergs et la tempête des 50è Hurlants de Kersauzon, ...

– *0404* est consacré à Steve Fosset, qui bat le record du monde, l'ayant réduit à 58 jours. C'est également à la fin du mois que Kersauzon finit le périple, en 63 jours, et reprend le Trophée car Fosset n'a pas respecté les règles ;

– *0405* rejoint le départ : le tour est fait, la course est finie.

4.2. Les groupes d'auteurs et de supports

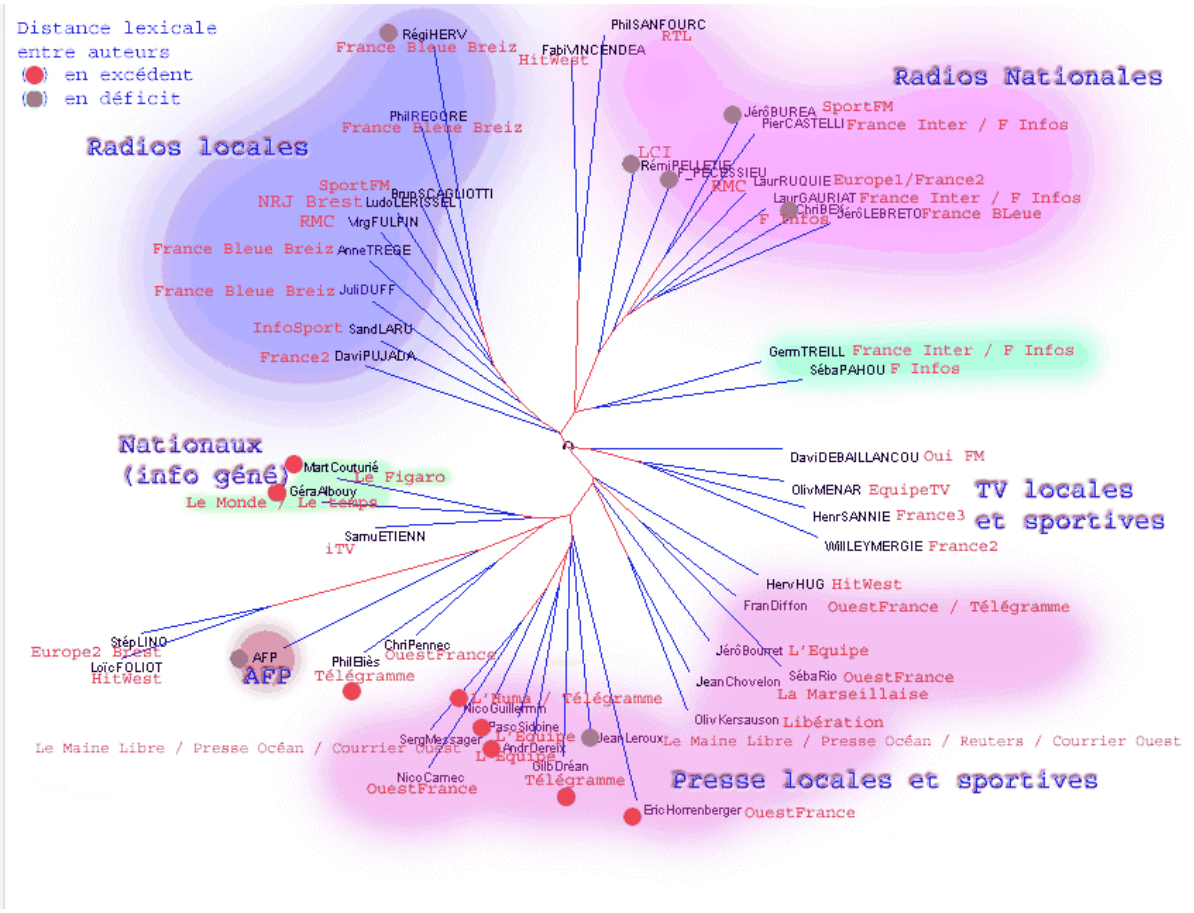

FIG. 3 Émergence de groupes d'auteurs et de supports.

Cette partie du travail a été effectuée à partir du corpus d'articles groupés par auteur. Il apparaît ainsi que les spécificités des auteurs s'établissent essentiellement en fonction du support pour lequel ils écrivent. La répartition s'effectue entre autres sur la base du vocabulaire. Les journalistes de journaux et revues s'autorisent ainsi plus de termes « difficiles » et techniques, tandis que les journalistes de la TV et de la radio ont tendance à utiliser un langage simplifié à l'oral.

Nous voyons que la figure 3 montre l'émergence de groupes de journalistes sur la base de la cohésion lexicale de leurs articles. Cette vision se dégage de l'analyse factorielle effectuée par Hyperbase. On voit que chaque grand groupe de journalistes correspond en réalité à un groupe de supports, où nous distinguons nettement les journalistes qui écrivent et les journalistes qui racontent :

– Le groupe des radios locales comporte ainsi, entre autres, France Bleu Breiz, NRJ Brest et RMC ;

– Le groupe des radios nationales comporte France Inter, France Info, LCI et Europe 1 ;

– Le groupe de télévisions locales et sportives France 3, France 2, Équipe TV, Oui FM ;

– Le groupe de la presse locale et sportive est très étendu, il couvre ainsi L'Équipe, Courrier Ouest, L'Humanité, Libération, la Marseillaise, Ouest France, Télégramme, etc. ;

– Le groupe de la presse nationale généraliste regroupe Le Figaro et Le Monde ;

– Et enfin, AFP forme un groupe isolé à lui tout seul.

La différence entre l'écrit et l'oral reste forte. Dans ce travail, elle émerge seulement sur la base de critères lexicaux. Mais elle est renforcée parallèlement par d'autres critères, comme par exemple ceux exploités par Biber (1994) : formes et modes verbaux, pronoms, etc.

Par contre, nous remarquons aussi que la presse locale et la presse sportive se confondent. Ceci, semble-t-il, est entériné depuis longtemps par les professionnels : les journaux sportifs sont une presse locale dédiée à un espace virtuel dénommé *sport*, mais le traitement est identique à celui de la presse locale. On retrouve par exemple des équivalences entre les couples comme :

maire / *député*, *entraîneur* / *directeur d'équipe*

policier / *arbitre*, *acteur local* / *joueur*

5. Conclusion

Dans cet article, nous avons présenté un travail de veille d'image d'une entreprise. Cette veille est effectuée grâce à l'analyse du corpus journalistique ciblé, *Trophée Jules Verne*. Nous avons mis en avant deux aspects de ce travail : constitution du corpus d'étude et son exploitation. Le corpus peut ainsi être paramétré selon différents critères, appartenant essentiellement à la dimension externe (auteur, date, support, média et genre). Mais l'accès par mots-clés, de même que la recherche de citations sont également disponibles. Notons qu'il a été prévu de rendre le corpus *Trophée Jules Verne* disponible pour la recherche et que l'accès à ce corpus est possible à travers l'interface décrite dans la sec. 2.3.

Le corpus a été analysé avec un outil de lexicométrie et un algorithme de calcul de vraisemblance. Les deux nous permettent d'accéder aux informations recherchées et de caractériser les personnages de navigateurs de même que les journalistes et supports. Grâce aux méthodes appliquées, nous dégageons l'image médiatique des marins impliqués dans le challenge le Trophée Jules Verne au début de l'année 2004 : Olivier de Kersauzon, Bruno Peyron, Francis Joyon et Steve Fosset. Nous avons vu que chaque marin joue un rôle spécifique et précis dans le challenge. Nous dégageons également les groupes de journalistes et de supports qui se consacrent à cette thématique. Là aussi, il existe des spécificités lexicales discriminantes, en particulier entre les journalistes qui écrivent et ceux qui racontent.

Quel est donc l'impact médiatique pour le groupe CapGemini suite au sponsoring d'Olivier de Kersauzon ? Et quelles sont les conclusions tirées pour ce groupe, l'intéressé de l'étude ? Notre étude a montré que l'image médiatique d'Olivier de Kersauzon est très favorable et positive : c'est un marin compétent qui connaît la navigation et la mer, il sait gérer les médias et son image. En plus, suite à cette participation, il remporte le Trophée. L'impact que ce sponsoring peut avoir sur le groupe est donc positif. Par contre, le nom du groupe est très peu associé au navigateur ou à son bateau. La différence est flagrante par rapport à d'autres sponsors, comme Orange ou Idec, qui ont donné leurs noms aux bateaux subventionnés. Cette homonymie, sponsor/bateau favorise l'émergence du sponsor dans les articles. Et comme ce n'est pas le cas pour CapGemini, il se trouve être largement déficitaire. La présence de CapGemini est donc plutôt implicite ou d'ordre iconique. D'autres conclusions portent sur l'émergence de journalistes passionnés de la navigation et sympathisant à Olivier de Kersauzon. Ces journalistes peuvent être ciblés par la suite afin d'assurer la couverture d'un événement sportif ou autre en relation avec ce navigateur.

Références

dub (1999). *The Dublin Core Element Set Version 1.1*. Technical report, Dublin Core Metadata Initiative. Disponible à *http://purl.org/dc/documents*.

Biber, D. (1994). Representativeness in corpus design. *Linguistica Computazionale*, IX-X, Current Issues in Computational Linguistics, in honor of Don Walker : 377--408.

Biber, D. & Finegan, E. (1994). Intra-textual variation within medical research articles. *Corpus-based research into language*, 12, : 201-222.

Brunet, E. (1981). *Les suffixes*. In *Le vocabulaire français de 1789 à nos jours. D'après les données du Trésor de la langue française*, Librairie Slatkine : 415-493.

Brunet, E. (1988). Une mesure de la distance intertextuelle : la connexion lexicale. *Revue d'Informatique et Statistique dans les Sciences Humaines*, 1-4 : 81-116.

Brunet, E. (1999). *Aperçu statistique sur l'évolution du vocabulaire français*.In *Nouvelle histoire de la langue française*, Éditions du Seuil : 675-627.

Brunet, E. (2003a). *Lexicométrie balzacienne*. In *À la recherche des Illusions perdues*, Nizet : 29-48.

Brunet, E. (2003b). *Nouvelles méthodes statistiques. L'exemple de Rabelais*. In *Ancien et moyen français sur le Web*, Les éditions DAVID : 33-54.

Brunet, E. (2004). *Statistiques rimbaldiennes*. In Les *littératures de l'Europe unie*, Cesenatico, Université de Bologne.

Grabar, N., Sillam, M., Jaulent, M.-C., Lefebvre, C., Henrion, E. & Néri, C. (2005). From likelihoodness between words to the finding of functional profile for ortholog genes. In *RANLP 2005 WS on Text Mining Research, Practice and Opportunities*, Borovets, Bulgaria.

Habert, B., Grabar, N., Jacquemart, P. & Zweigenbaum, P. (2001). Building a text corpus for representing the variety of medical language. In *Corpus Linguistics*, Lancaster.

Habert, B., Nazarenko, A. & Salem, A. (1997). *Les linguistiques de corpus*. Paris, Armand Colin/Masson.

Maciel, C. & Brunet, E. (2000). De FRANTEXT à PORTEXT. In *Desafios da Lusofonia*, Nice.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA, MIT Press.

Sinclair, J. (1994). *EAGLES. Corpus typology*. Technical report, EAG-CWG-IR-2. Disponible à *http://www.ilc.pi.cnr.it/EAGLES96/*. Visité le 02/03/2003.

Valette, M. & Grabar, N. (2004). Caractérisation de textes à contenus idéologiques : statistique textuelle ou extraction de syntagme ? L'exemple du projet PRINCIP. In *Journées de traitement automatique des données textuelles (JADT)*, Liège, Belgique.

Zweigenbaum, P. & Grabar, N. (2003). Corpus-based associations provide additional morphological variants to medical terminologies. In *American Medical Informatics Association (AMIA)*.