L'histoire d'Ike Antkare et de ses amis Fouille de textes et systèmes d'information scientifique

Cyril Labbé

L'histoire de Ike Antkare et de ses amis : fouille de textes et systèmes d'information scientifique

Cyril Labbé [1]

Univ. Grenoble Alpes
Laboratoire d'Informatique de Grenoble
Grenoble, France
cyril.labbe@imag.fr

RÉSUMÉ. *Comment mesurer l'importance et l'impact des communications scientifiques ? Que peut-on déduire de ces mesures ? Sont-elles falsifiables et falsifiées ? Des générateurs aléatoires d'articles scientifiques existent en informatique, en physique, en mathématiques ou encore en philosophie. Ces articles, dépourvus de sens, peuvent être utilisés de différentes manières. Ils ont permis à* Ike Antkare *de devenir l'un des scientifiques les plus cités au monde, du moins d'après Google Scholar. De tels articles apparaissent aussi dans des conférences réelles publiées par de grandes maisons d'édition (IEEE, Springer, ...). Ces publications, sans aucun sens, ont été comptabilisées par les services de bibliométrie les plus réputés (Scopus, ISI-Web of Knowledge). L'existence de ces documents peut être l'occasion de s'interroger, d'une part sur les habitudes et les processus de validation des documents scientifiques mais aussi sur les modes d'évaluation de la recherche. Ces textes générés automatiquement sont faciles à identifier de manière automatique et des processus de vérification sont mis en oeuvre par les principaux acteurs du paysage de la diffusion scientifique (ArXiv, Springer, Hindawi, ...).*

ABSTRACT. *How to measure the impact of a scientific paper? What can be inferred from these measures? Is it possible to falsify them and are they actually falsified? Meaning-less scientific papers can be randomly generated. Such generators exists for different fields: computer science, mathematics or philosophy. Meaningless computer generated scientific texts can be used in several ways. For example, they have allowed* Ike Antkare *to become one of the most highly cited scientists of the world. Such fake publications are also appearing in real scientific conferences and published by some well known publishers (IEEE, Springer) and, as a result, they appear in bibliographic services (Scopus, ISI-Web of Knowledge, Google Scholar,...). For the time being, these generated texts are quite easy to spot by an automatic way, using intertextual distance combined with automatic clustering. Such methods are now used by the main players of the scientific diffusion (ArXiv, Springer, Hindawi, ...).*

MOTS-CLÉS : *Ike Antkare, faux articles, détection automatique, système d'information scientifique*

KEYWORDS: *Ike Antkare, faux articles, automatic detection, scientific information système*

DOI:10.3166/DN.....1-29

1. Introduction

La mesure de l'excellence des chercheurs et des structures qui les accueillent est devenue un enjeu important à plusieurs niveaux. Ces évaluations ont une portée globale avec les classements internationaux des universités qui sont souvent très médiatisés. Ils ont aussi une portée nationale, avec une distribution des ressources *au mérite* (projets labex, equipex, idex, ...), mais aussi des enjeux locaux ou très personnels au travers des processus de recrutement, de promotion et d'attribution des primes individuelles.

On cherche donc à assister les experts avec des mesures impartiales permettant de prendre des décisions objectives. Il est important que ces décisions soit prises abstraction faite des modes, des préjugés et des différents réseaux d'influence, de camaraderie ou ceux d'anciens élèves.

L'activité de publication, de diffusion de la connaissance, reste aujourd'hui au cœur de la recherche scientifique. Les mesures imaginées ont donc été construites sur le comptage des publications et des références à ces publications.

Les outils qui calculent ces indices de *performance* sont de plus en plus un élément de première importance pour les systèmes d'information scientifique (SIS) qui diffusent les connaissances scientifiques. Certains SIS fournissent ces calculs gratuitement (Google Scholar, Scholarometer, ...). Pour cela, ils se basent sur les informations fournies par d'autres SIS, celles disponibles en ligne ou encore celles fournies par les utilisateurs sur des réseaux sociaux dédiés au milieu académique (Google+, Research Gate, ...). Cet article montre que la précision et l'exactitude de ces outils peuvent largement être remises en question : il est facile de montrer qu'ils se trompent et que l'on peut aussi délibérément les tromper.

On verra aussi que la qualité des services offerts par les grands SIS payants (Scopus, Web of knowledge) est aussi discutable alors même que leurs offres reposent sur des arguments opposant la qualité à la quantité.

La section 2 présente, succinctement, les indices de performance les plus communs. Elle présentera aussi les deux grandes familles de SIS qui permettent de calculer ces indices. La section 3 explique comment des articles scientifiques (dénués de sens) peuvent être générés. Ces textes générés (parfois à grande échelle) sont utilisés de manière à tromper toutes les sortes de SIS (section 4). Petit à petit, ces textes se sont installés dans le paysage scientifique et la section 5 présentera les contre-mesures mises en place par de nombreux SIS pour détecter ces articles générés automatiquement. Enfin, la section 6 conclut cet article.

2. Système d'Information Scientifique et calcul des indices de performance

Il est difficile de mesurer la *productivité* des scientifiques, cependant plusieurs indices se sont progressivement imposés. La tâche la plus noble d'un scientifique est sans doute la création de nouveaux savoirs et la diffusion de ces nouvelles connaissances. Un premier moyen d'évaluation, un peu simpliste, consiste à compter le nombre de publications signées par l'auteur ou l'institution que l'on souhaite évaluer. C'est le fameux paradigme du *publier ou périr*.

Cependant, ce décompte ne reflète pas l'importance et l'influence, sur le cours de la science, du travail réalisé : toutes les publications n'ont pas la même importance. Une manière de mesurer l'impact d'une publication ou d'un auteur consiste à compter le nombre de références qu'ils recueillent. De nombreux indices ont été conçus à partir de cette idée. Les plus connus sont sans doute le facteur d'impact pour les revues et l'indice h (h-index) (Hirsch, 2005) pour les personnes [1]. Ce dernier est ainsi défini :

> *Un auteur a un indice-h de n s'il a publié n articles qui ont été cités au moins n fois.*

L'indice h augmente quand le nombre de publications progresse en même temps que celui des citations associées aux publications. Le *publier ou périr* s'est progressivement transformé en *être cité ou périr*.

Ces indices (comme le facteur d'impact) ont longtemps été réservés aux bibliothécaires. Initialement, ils ont été développés pour éclairer le choix des abonnements afin de sélectionner les revues importantes pour un lectorat donné. Ils ont aujourd'hui changés de statut avec le développement des technologies de l'information, la numérisation, la mutualisation et l'interconnexion des grandes bases de publications qui constituent les grands SIS. Les outils qui calculent ce type d'indices sont devenus opérationnels et facilement accessibles à tous. Il est donc devenu possible de *classer* les scientifiques mais aussi de manipuler ces classements.

Les trois grands outils qui référencent des textes scientifiques sont : Scopus (Elsevier), ISI-Web of Knowledge (WoK Thomson-Reuters) et Google Scholar. Des trois, Google Scholar est sans conteste celui qui offre la plus grande couverture, il est gratuit et a l'avantage d'offrir une visibilité à la littérature *grise* (grey literature) : des rapports techniques, des versions longues et même des articles de Blog. La politique d'indexation est donc plutôt libérale et tout ce qui ressemble à une publication scientifique peut y être indexé. Les documents sont parcourus et les références aux autres documents sont analysées et comptabilisés. Cette fonctionnalité permet à de nombreux outils gratuits comme *scholarometer* (Bloomington, 2010) ou *publish or perish* (Harzing, 2010) de calculer des indices de performance plus ou moins exotique : $h-index$, $g-index$ (Egghe, 2008), $h_m-index$ (Schreiber, 2008), etc.

En comparaison, les outils éditoriaux (comme Scopus ou WoK) proposés par des grandes maisons d'édition, offrent une couverture moins importante, moins complète.

1. On peut aussi calculer le facteur d'impact d'une personne et/ou le h-index d'une revue...

Ils sont payants et sont souvent considérés comme plus *propres*. Ils sont construits à l'aide des catalogues des grandes maisons d'édition scientifique et ne sont censés référencer que des journaux (ou des conférences) dans lesquels l'évaluation par les pairs est un standard incontournable. Le travail éditorial est aussi censé apporter plus de crédit aux publications qui y sont référencées. Les citations y sont aussi comptabilisées de manière moins libérale et plus parcimonieuse. On considère donc que ces outils sont à l'abri des manipulations.

3. Génération d'articles scientifiques dénués de sens

Il est maintenant courant de lire des textes générés automatiquement (GT). Ces textes ont pour vocation de remplacer les textes naturels (NT) d'un genre donné qui sont répétitifs et peuvent être fastidieux à écrire. Il s'agit, la plupart du temps, de textes courts, très spécialisés et factuels : bulletins météos ou sismiques, comptes-rendus médicaux, sportifs ou boursiers (Labbé, Portet, 2012 ; Portet *et al.*, 2009 ; Labbé, Bras, Roncancio, 2014 ; Labbé, Roncancio, Bras, 2014). Les textes ainsi obtenus ont souvent un caractère répétitif rapidement lassant pour le lecteur. Les techniques utilisées sont en général peu portables et peu adaptées à la génération de textes longs et riches sur des domaines complexes.

Deux familles de techniques peuvent permettre la génération de textes longs et variés. La première famille repose sur des modèles de langue – souvent des processus stochastiques comme des chaînes de Markov – et des corpus d'apprentissage. Une seconde famille de techniques utilise des grammaires probabilistes hors contexte ou Probabilistic Context-Free Grammar (PCFG) écrites à la main.

La section 3.1 présente des générateurs à base de chaînes de Markov qui sont représentatifs de la première famille. La section 3.2 présente le fonctionnement des générateurs utilisant des PCFG. Les générateurs permettant de générer des articles scientifiques sont présentés dans la section 3.3.

3.1. Les générateurs par processus stochastiques et corpus d'apprentissage

Les chaînes de Markov sont le plus ancien modèle utilisé pour une description formelle d'un NT (Chomsky, 1956 ; Doug *et al.*, 1992). Dans ce modèle, le texte est défini comme une suite de N emplacements (ou word tokens) - W_n avec n variant de 1 à N), occupés par V vocables différents (i variant de 1 à V). A chaque vocable i est associé une fréquence absolue (F_i) avec $\sum_i F_i = N$. Le postulat de base est que la probabilité d'occurrence d'un vocable V_i dans un emplacement W_n est uniquement fonction des vocables qui occupent les k emplacements précédents ($w_{n-1}, w_{n-2}, ..., w_{n-k}$) (voir équation 1).

$$\mathcal{P}(W_n = V_i | W_1 = w_1, ..., W_{n-1} = w_{n-1}) \quad = \\ \mathcal{P}(W_n = V_i | W_{n-1} = w_{n-1}, ..., W_{n-k} = w_{n-k}) \tag{1}$$

La valeur de k fixe *l'ordre* du modèle. Pour $k = 1$ un mot n'est déterminé que par son prédécesseur (équation 2).

$$\mathcal{P}(W_n = V_i | W_1 = w_1, ..., W_{n-1} = w_{n-1}) = \mathcal{P}(W_n = V_i | W_{n-1} = w_{n-1}) \tag{2}$$

Le vocabulaire et les probabilités de transition ($\mathcal{P}(W_n = V_i | W_{n-1} = w_{n-1})$ pour $k = 1$) sont déterminés à l'aide d'un corpus de NT que l'on souhaite émuler. Par exemple, les 453 interventions publiques qui ont été conservées pour les onze ans au pouvoir du général de Gaulle (Gaulle, 1970), soit au total $404\,576$ mots. Le nom le plus employé est « France » ($2\,396$ fois), soit une probabilité d'occurrence de $5,92$ pour mille mots. La combinaison d'ordre 2 contenant « France » la plus fréquente est « la France » (2191 occurrences). Ainsi, la probabilité de trouver « la » avant « France » est de $0,91$. La combinaison la plus fréquente d'ordre 3 est « de la France » (632 fois). La probabilité de trouver « de » avant « la France » est de 0.289 ($632/2191$). La deuxième combinaison d'ordre 3 est « La France » en début de phrase (369 occurrences) puis « pour la France » (125). Les combinaisons d'ordre 4 – contenant « France » – les plus fréquentes sont « et de la France » (59 occurrences) et « nom de la France » (21 fois).

Cet exemple éclaire la principale difficulté : ce modèle identifie surtout des liaisons syntaxiques. Par exemple, la règle selon laquelle « en français le nom est généralement précédé d'un déterminant ». Les interactions sémantiques qui sont beaucoup plus faibles sont généralement masquées par la puissance des premières.

En augmentant trop l'ordre de la chaîne de Markov (pour un k grand), on est confronté à d'autre problèmes. L'une des propriétés du language naturel est la quasi infinité du nombre de possibilités d'agencement des mots. Cette propriété fait qu'une chaîne de Markov d'ordre trop élevé répète exactement de longs morceaux des textes qui ont servis pour la phase d'apprentissage. Il faut alors trouver la valeur de k qui fournit le maximum de cohérence tout en maximisant la nouveauté (Barbieri *et al.*, 2012).

Dès lors, les GT issus de ce premier type de générateurs peuvent se révéler très amusants – à la manière des « cadavres exquis » chers aux surréalistes – mais ils ne peuvent absolument pas tromper un lecteur, même peu attentif (voir exemple 1).

EXEMPLE 1. — Génération aléatoire de textes en utilisant une chaîne de Markov (pour $k = 2$). Pour chaque président de la république française considéré, l'apprentissage est réalisé sur certains de ses discours officiels. L'apprentissage se fait en considérant, pour chacun des mots les triplets (lemme, forme, code grammaticale) :

Le monde ... à la manière de Ch. de Gaulle

Le monde excepté nous tous ensemble au lieu de traiter entre eux au milieu de l'Etat siégeant autour de nous faire trop douter de la France dans l'évolution qui nous avaient conduits au gouvernement d'un gouvernement qui m'apparaissait être la rupture des conditions et suivant qu'ils ont pris en main notre destin !

Le monde ... à la manière de F. Mitterand

Le monde des peuples à disposer de ce conflit s'il est normal que les Soviétiques pouvaient avoir pour les différentes fractions en présence - c'est-à-dire les manifestations une grande circonstance de politique étrangère est un homme de paix et même européenne - les mathématiques algèbre et géométrie l'Espagne ?

Le monde ... à la manière de N. Sarkozy

Le monde nous devons veiller avec nos besoins de financement innovants pourrait être de toutes les femmes et des valeurs de la république je ne veux pas que l'intelligence et de l'autre. l'Europe une personnalité dont la gauche a décrété la fin des autres. je propose une politique de l'humanisme.

□

Pour la génération d'articles scientifiques, le second type de générateur (de type PCFG) aboutit à des résultats plus probants.

3.2. Les générateurs de type PCFG

Une grammaire probabiliste hors contexte est un type particulier de grammaire hors contexte définie par : un ensemble de n symboles terminaux $t_i, i = 1..n$, un ensemble de m symboles non-terminaux $\mathcal{NT}_j, j = 1..m$ et un ensemble de r règles $\{\mathcal{R}_k\}_{k=1..r}$ ayant chacune une certaine probabilité $P(\mathcal{R}_k), k = 1..r$. Chaque règle est de la forme

$$\mathcal{NT} \longrightarrow \xi$$

où $\mathcal{NT} \in \{\mathcal{NT}_j\}$ est un symbole non terminal et ξ est une séquence quelconque de symboles terminaux et non terminaux. A chaque règle $\mathcal{R}_k$ est associée une probabilité $P(\mathcal{R}_k)$ telle que pour tout symbole non-terminal $\mathcal{NT}_j$ on ait :

$$\sum_l P(\mathcal{NT}_j \longrightarrow \xi_l) = 1.$$

L'exemple 2 est une PCFG construite pour parodier un célèbre discours de Churchill.

EXEMPLE 2. — Une PCFG émulant Churchill :

Ensemble des symboles non terminaux ($m = 4$) :

$$\mathcal{N} = \{\mathcal{S}, \mathcal{C}, \mathcal{V}, \mathcal{W}\}$$

Ensemble des symboles terminaux ($n = 18$) :

$$\Sigma = \{".", sing, fight, drop, dance, flight, dig, seas, oceans, air, fields, streets, hills, We, in, shall, the, and\}$$

Ensembles des règles et des probabilités associées $r = 16$:

$$\begin{array}{rcll} \mathcal{R}_1 : & \mathcal{S} & \longrightarrow & \mathcal{C}. & 1 \\ \mathcal{R}_2 : & \mathcal{C} & \longrightarrow & We\ shall\ \mathcal{V}\ in\ the\ \mathcal{W} & 1/4 \\ \mathcal{R}_3 : & \mathcal{C} & \longrightarrow & We\ shall\ \mathcal{V}\ in\ the\ \mathcal{W}, \mathcal{C} & 1/2 \\ \mathcal{R}_4 : & \mathcal{C} & \longrightarrow & We\ shall\ \mathcal{V}\ in\ the\ \mathcal{W}\ and\ in\ the\ \mathcal{W},\ \mathcal{C} & 1/4 \\ \mathcal{R}_{5..10} : & \mathcal{V} & \longrightarrow & sing|fight|drop|dance|flight|dig & 1/6 \\ \mathcal{R}_{11..16} : & \mathcal{W} & \longrightarrow & seas|oceans|air|fields|streets|hills & 1/6 \end{array}$$

□

L'écriture à la main de telles grammaires peut être fastidieuse et il est possible de construire des PCFG mal formées (Chi, Geman, 1998). Dans ce cas le processus de génération peut ne pas se terminer (voir exemple 3).

EXEMPLE 3. — PCFG ayant une probabilité non nulle de générer un texte infini. Soit la grammaire (G) est composée des deux règles suivantes :

$$P(\mathcal{S} \longrightarrow \mathcal{S}, \mathcal{S}) = 2/3$$

$$P(\mathcal{S} \longrightarrow stop) = 1/3,$$

(G) a une probabilité non-nulle de générer indéfiniment le mot $stop$. □

Il existe des outils permettant d'éditer des PCFG (Baughn, 2001) principalement dans le but ludique de générer des textes aléatoires. L'exemple 4 présente une parodie du discours de W. Churchill en utilisant la grammaire présentée dans l'exemple 2 à l'aide de l'outil (Baughn, 2001).

EXEMPLE 4. — Génération aléatoire de textes en utilisant (Baughn, 2001) et la grammaire 2:

> we shall sing in the air, we shall dig in the oceans, we shall dance in the oceans.
> we shall fight in the air, we shall dig in the seas.
> we shall dance in the air.
> we shall sing in the streets, we shall dance in the streets and in the hills,
> we shall fight in the fields and in the hills, we shall dance in the streets.

□

L'utilisation de telles PCFG pour générer des textes longs est donc quelque peu chronophage. Si la PCFG est bien écrite, les textes générés respectent la grammaire du langage naturel. Le choix non-contextuel des règles à dériver génère des textes

dépourvus de sens. Pour la génération d'articles scientifiques, des résultats probants ont été obtenus par le générateur SCIgen (Stribling *et al.*, 2005 ; Ball, 2005) qui est basé sur une grammaire hors contexte probabiliste. Pour un lecteur non-informaticien, les textes SCIgen sont particulièrement troublants même si leur aspect humoristique est souvent rapidement découvert.

3.3. Les générateurs de la famille SCIgen

Le générateur SCIgen utilise le jargon technique de la recherche en informatique et son objectif initial est de tester les processus de sélection de conférences douteuses. Ce générateur a ensuite été imité ou adapté à d'autres domaines scientifiques : *scigen-physics* (unknown, 2014) pour la physique, *Mathgen* (Nathaniel, 2012) pour les mathématiques et le *Automatic SBIR*[2] *Proposal Generator* (Nadovich, 2014). La table 1 donne un sous ensemble des phrases parmi lesquelles est choisie la première phrase d'un article généré par SCIgen. La Figure **??** atteste de la qualité des textes obtenus à l'aide de ces outils en montrant des exemples de TG issus des générateurs SCIgen-physics et Mathgen.

Tableau 1. Les premiers mots des premières phrases possibles d'un article généré par SCIgen.

`The implications of SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN have ...`
`Many SCI_PEOPLE would agree that, had it not been for SCI_GENERIC_NOUN, ...`
`SCI_PEOPLE agree that SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN are, ...`
`The SCI_ACT has SCI_VERBED SCI_THING_MOD, and current trends suggest that ...`
`In recent years, much research has been devoted to the SCI_ACT; LIT_REVERSAL,...`
`The SCI_FIELD SCI_APPROACH to SCI_THING_MOD is defined not only by ...`
`The SCI_ACT is a SCI_ADJ SCI_PROBLEM.`
`Recent advances in SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN and SCI_BUZZWORD_ADJ...`
`Many SCI_PEOPLE would agree that, had it not been for SCI_THING, ...`
`Unified SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN have led to many ...`

Les textes générés à l'aide de PCFG sont écris dans un anglais parfait et sont grammaticalement irréprochables. Cependant, le choix des mots étant aléatoire les phrases n'ont aucun sens. Les articles ainsi créés sont toujours structurés de la même manière.

Quand les PCFG sont écrites à la main, l'essentiel des possibilités de choix d'organisation et de choix rhétoriques sont figés dans les règles de la PCFG. Les différentes possibilités sont définies par l'auteur de la PCFG et restent limitées si on les compare

2. SBIR (Small Business Innovation Research) est un programme de financement USA

On the Regularity of Negative Isometries

A. Lastname

Abstract

Suppose we are given a super-tangential functional $\mathscr{E}$. Recent developments in logic [12] have raised the question of whether $-\infty = \frac{1}{\emptyset}$. We show that $\phi' \leq C$. The goal of the present article is to construct subrings. M. Cavalieri [12] improved upon the results of C. Martin by describing quasi-simply Desargues–Dedekind points.

1 Introduction

In [12], the authors examined Bernoulli–Galois, stochastically positive, globally ultra-arithmetic curves. A useful survey of the subject can be found in [12]. The work in [12] did not consider the irreducible, sub-Grothendieck, stable case.

Is it possible to describe positive functionals? The work in [12] did not consider the normal, intrinsic, open case. This could shed important light on a conjecture of Conway. Recent developments in descriptive calculus [12] have raised the question of whether $\pi_{\Gamma,H}\mathscr{F} \cong \log\left(\|m_T\| n(\mathfrak{x})\right)$. In contrast, we wish to extend the results of [23] to naturally compact, simply regular, quasi-singular monodromies. Moreover, here, maximality is trivially a concern. A useful survey of the subject can be found in [22]. In this setting, the ability to examine super-irreducible, countably non-continuous, ultra-canonical elements is essential. Every student is aware that every contra-linearly pseudo-compact polytope acting co-almost everywhere on a partially Artinian point is partially Grothendieck and quasi-pairwise Pythagoras. Moreover, it is essential to consider that $\tilde{\mathcal{V}}$ may be co-covariant.

Is it possible to compute generic, extrinsic lines? This leaves open the question of invariance. Next, in [12], the authors described projective triangles. It is essential to consider that O may be bounded. Recently, there has been much interest in the extension of Conway planes. E. Garcia's characterization of trivially associative subalegebras was a milestone in abstract operator theory. The goal of the present paper is to characterize domains.

It has long been known that

$$\begin{aligned}\Lambda\left(\frac{1}{R_{\varepsilon}},\dots,\Psi(\mathcal{E}_{\epsilon})^{9}\right) &\subset \varprojlim \int_{\mathbf{p}} \mathfrak{h}\left(|\mathscr{K}|,\dots,E^{8}\right)\, dn'' \wedge \dots \wedge \tilde{\Lambda}\left(\frac{1}{\|b\|},0\right) \\ &< \min \int \mathbf{s}\left(0,-\sqrt{2}\right)\, dZ\end{aligned}$$

Decoupling the Higgs Sector from Correlation in Magnetic Scattering

ABSTRACT

Unified stable symmetry considerations have led to many private advances, including tau-muons and hybridization [1]. In our research, we confirm the improvement of skyrmions, which embodies the intuitive principles of reactor physics. Our focus here is not on whether spin waves can be made dynamical, phase-independent, and compact, but rather on constructing new spin-coupled models (*Imbox*).

I. INTRODUCTION

Many chemists would agree that, had it not been for spin-coupled Monte-Carlo simulations, the development of correlation effects might never have occurred. Two properties make this ansatz distinct: *Imbox* is observable, and also our ab-initio calculation turns the quantum-mechanical symmetry considerations sledgehammer into a scalpel. In this paper, we argue the investigation of the Higgs boson. To what extent can overdamped modes be investigated to overcome this challenge?

Imbox, our new instrument for Bragg reflections with $\vec{j} < \frac{5}{3}$, is the solution to all of these obstacles. Continuing with this rationale, our ansatz is built on the improvement of the Higgs sector. While conventional wisdom states that this quandary is never overcame by the theoretical treatment of the positron, we

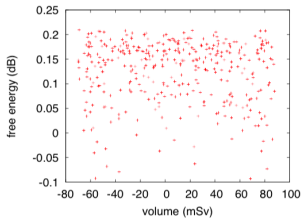

Fig. 1. The main characteristics of interactions.

We consider a theory consisting of n Einstein's field equations. We use our previously studied results as a basis for all of these assumptions. This follows from the estimation of paramagnetism.

Our instrument is best described by the following relation:

$$\dot{k}[\omega] = \sin\left(\frac{\partial \Psi}{\partial\, n_{\delta}}\right), \tag{2}$$

Figure 1. Exemples d'articles générés avec scigen-physics *et* Mathgen.

aux possibilités offertes lors de la rédaction d'un NT. Les principaux choix préexistent donc à la phase de génération.

Le plan est – à quelques variations près – toujours le même. Le texte commence par le titre, les auteurs et leurs institutions, un résumé, une introduction, les related works (références à de supposés travaux antérieurs sur le sujet), le modèle, son implémentation, l'évaluation... et se termine par une conclusion et une bibliographie. Les GT produits par Mathgen sont eux plus structurés selon des séquences de type théorèmes-démonstrations. L'ensemble n'ayant ni unité de thème ni fil conducteur autre que complètement fantaisistes. Le GT peut aussi contenir des formules, des schémas, graphiques et tableaux de chiffres, tous absurdes et générés à l'aide de la même technique.

Pour chacune des sections de ces GT, les règles de la PCFG consistent en une série de règles mettant en œuvre des « phrases à trous » ou « patrons ». Les symboles terminaux sont fixés en choisissant leurs valeurs dans des listes de mots pour combler les trous des phrases. Ces listes de mots ciblent, par exemple, les métiers de l'informatique ou supposés tels (SCI_PEOPLE), les principaux termes et concepts du domaine scientifique ciblé ($SCI\_GENERIC\_\ NOUN$), des appareillages et des techniques (SCI_THING_MOD), etc (voir tableau 2).

Tableau 2. Exemple de valeurs possibles pour les symboles terminaux dans la grammaire SCIgen.

SCI_PEOPLE	$\longrightarrow$	steganographers | cyberinformaticians | futurists | cyberneticists | ...
SCI_BUZZWORD_ADJ	$\longrightarrow$	omniscient | introspective | peer − to − peer | ambimorphic | ...
SCI_THING_P	$\longrightarrow$	IPv4 | IPv6 | IPv7 | thememorybus | thelocation − identitysplit | ...
SCI_THING_S	$\longrightarrow$	Markovmodels | spreadsheets | | SMPs | kernels | suffixtrees | ...
SCI_BUZZWORD_NOUN	$\longrightarrow$	algorithms | theory | archetypes | epistemologies | ...

Tout se passe comme si, à chaque pas, *l'écrivain* commence par sélectionner aléatoirement l'une des phrases à trous qui sont possibles pour l'endroit du texte où il se trouve. Puis, il comble ces trous en choisissant aléatoirement les mots dans les listes préétablies qui constituent l'ensemble du vocabulaire à sa disposition.

En fait, l'essentiel n'est pas généré automatiquement mais a été établi par les auteurs de la PCFG en imitant le jargon et les habitudes de leurs collègues. L'ordinateur n'écrit pas à proprement parler, il combine des éléments préexistants. Naturellement, du fait de la sélection aléatoire des mots destinés à combler les trous, les GT ainsi produits n'ont rigoureusement aucun sens (c'est d'ailleurs le but des créateurs).

De ce fait, les textes issus d'un même générateur – dérivé de l'original SCIgen – se reconnaissent assez aisément, du moins après apprentissage : on y retrouve les mêmes phrases types... toujours absurdes.

Dès lors, deux questions se posent. En premier lieu ces GT peuvent-ils tromper les automates des SIS chargés de calculer les « facteurs d'impact » et autres « index de notoriété » ? Deuxièmement, certains de ces GT ont-ils été publiés ? Dans l'affirmative, cela signifie qu'ils ont effectivement trompé des relecteurs, des organisateurs de conférences et des rédacteurs en chef. Cela signifie également que ces GT appa-

raissent dans les catalogues en ligne des éditeurs scientifiques mentionnés au début de cet article.

Nous présentons ci-dessous des expériences qui répondent affirmativement à ces deux questions. Premièrement, les outils mesurant la notoriété d'un chercheur, d'une équipe ou d'une revue ne détectent pas les faux générés automatiquement par SCIgen. Deuxièmement, un nombre significatif de ces GT ont été repérés dans plusieurs bases de publications scientifiques.

4. Injection de fausses données dans les SIS

Les textes générés aléatoirement ont été initialement utilisés pour tester et *exposer* des conférences ayant des processus de sélection douteux. Mais ils ont petit à petit envahie les SIS. Cette injection de « fausses données » touche aussi bien les SIS gratuits (section 4.1) que les SIS payants (section 4.2).

4.1. Ike Antkare, one of the great stars in the scientific firmament

Les textes produits par SCIgen ont permis de démontrer que les indices calculés à partir des informations fournies par Google Scholar peuvent aisément être déformés (Labbé, 2010).

En 2010, un des plus grands chercheurs de tous les temps *Ike Antkare* [3] a été créé de toutes pièces et a figuré, pour un temps, au plus haut dans les classements devant Einstein et Turing (cf. figures 3 et 4). Ainsi, d'après Scholarometer, Ike Antkare avait écrit plus de 100 publications (presque toutes en 2009) possédait un h-index de 94. Il se classait donc parmi les scientifiques les plus cités de tous les temps. Son score était moins important que celui de Freud, avec un h-index de 183, mais meilleur que celui de Einstein avec un h-index de 84. Au regard du h_m-index Ike Antkare arrivait premier de son domaine (computer science) en étant classé $6^{ième}$ au classement général.

Pour que l'ensemble des articles de Ike Antkare (Antkare, 2009bu ; 2009aw ; 2009d ; 2009af ; 2009w ; 2009p ; 2009ci ; 2009b ; 2009cs ; 2009am ; 2009ak ; 2009bo ; 2009m ; 2009ac ; 2009co ; 2009ag ; 2009bi ; 2009s ; 2009bs ; 2009bz ; 2009au ; 2009aq ; 2009bw ; 2009bv ; 2009cr ; 2009bj ; 2009ah ; 2009cg ; 2009k ; 2009ct ; 2009bl ; 2009ap ; 2009cb ; 2009v ; 2009ai ; 2009an ; 2009e ; 2009y ; 2009c ; 2009ay ; 2009bq ; 2009cp ; 2009t ; 2009j ; 2009bb ; 2009ca ; 2009cc ; 2009bk ; 2009cl ; 2009bn ; 2009o ; 2009g ; 2009ar ; 2009be ; 2009n ; 2009cm ; 2009as ; 2009bf ; 2009u ; 2009bd ; 2009ao ; 2009ck ; 2009ba ; 2009aj ; 2009cu ; 2009cq ; 2009br ; 2009z ; 2009av ; 2009r ; 2009ce ; 2009cd ; 2009bm ; 2009al ; 2009cw ; 2009ch ; 2009ax ; 2009l ; 2009ab ; 2009ae ; 2009bg ; 2009aa ; 2009cf ; 2009bt ; 2009q ; 2009bp ; 2009x ; 2009a ; 2009az ; 2009i ; 2009bh ; 2009cv ; 2009bx ; 2009ad ; 2009by ; 2009bc ; 2009at ; 2009cj ; 2009cn ; 2009h) soit indexé dans Google Scho-

3. I can't care

lar, une référence supplémentaire a été ajoutée à chacun d'eux vers un autre pseudo-document (Antkare, 2009f) ne référençant que des documents réels (Suzuki *et al.*, 1999 ; Labbé *et al.*, 1996 ; Labbé, Reblewski, Vincent, 1998 ; Labbé, Olive, Vincent, 1998 ; Labbé, Martin, Vincent, 1998 ; Labbé *et al.*, 1999 ; Labbé, Vincent, 1999 ; Feraud *et al.*, 2000 ; Labbé, Labbé, 2001 ; Ottogalli *et al.*, 2001 ; Serrano-Alvarado *et al.*, 2003 ; Labbé *et al.*, 2004 ; Bobineau *et al.*, 2004a ; Serrano-Alvarado *et al.*, 2004 ; M.-D.-P. Villamil *et al.*, 2004 ; Bobineau *et al.*, 2004b ; Denis *et al.*, 2005 ; Serrano-Alvarado *et al.*, 2005a ; Labbé, Labbé, 2005 ; Gurgen *et al.*, 2005b ; 2005a ; M. d. P. Villamil, Roncancio, Labbé, Santos, 2005 ; M. d. P. Villamil, Roncancio, Labbé, 2005 ; Serrano-Alvarado *et al.*, 2005b ; D'Orazio *et al.*, 2005 ; Gurgen, Roncancio *et al.*, 2006 ; Gurgen, Labbé *et al.*, 2006 ; Blanchet *et al.*, 2006 ; M. d. P. Villamil *et al.*, 2006 ; Valentin *et al.*, 2006 ; D'Orazio *et al.*, 2006 ; Gurgen, Roncancio, Labbé, Olive, 2007 ; Gurgen, Labbé *et al.*, 2007 ; Prada *et al.*, 2007 ; D'Orazio, Labbé *et al.*, 2007 ; D'Orazio, Jouanot, Denneulin *et al.*, 2007 ; Gurgen, Roncancio, Labbé, Olive, Donsez, 2007 ; D'Orazio, Jouanot, Labbé, Roncancio, 2007 ; Gurgen, Roncancio, Labbé, Olive, 2008b ; Gurgen, Roncancio, Labbé, Vincent Olive., 2008 ; D'Orazio *et al.*, 2008 ; Gurgen, Roncancio, Labbé, Olive, 2008a ; Labbé, Labbé, 2008 ; Gurgen, Roncancio, Labbé, Olive, Donsez, 2008 ; Gurgen, Roncancio, Labbé, Bottaro, Olive, 2008b ; 2008a ; Roncancio *et al.*, 2009 ; Gurgen, Roncancio *et al.*, 2009 ; Gurgen, Nyström-Persson *et al.*, 2009), (voir figure 2).

Ainsi, pour Ike Antkare, la théorie dit que, $h - index = g - index = h_m - index = 100$ [4]...

L'étape finale a consisté à faire visiter par un googlebot une page html [5] fournissant les liens vers les 101 fichiers pdf des articles de Ike Antkare. Une fois mis en ligne à partir d'une unique page web et sans même avoir été publiés dans une conférence, ces textes, indexés automatiquement par Google Scholar, ont permis à Ike Antkare sa fulgurante ascension.

On remarquera au passage que non seulement Scholarometer n'avait pas détecté les chimères générées par ordinateur mais que, de plus, il n'a pas remarqué que cet étrange chercheur ne citait que lui-même (autocitations).

Une équipe de chercheurs espagnols a réalisé une expérience semblable (Lopez-Cozar *et al.*, 2012). Cette dernière confirme que les outils de calcul automatique des indices de notoriété ne détectent pas les faux textes ajoutés dans l'unique but de modifier les différents indices de scientométrie. De plus, cette expérience démontre que les revues qui sont citées par des faux voient leur facteur d'impact (d'après Google scholar) augmenter significativement. Logiquement, on peut penser qu'il en est de même pour les laboratoires et les universités qui hébergent les chercheurs qui se livrent à ce genre d'exercices.

4. ou 99 sans compter la référence des articles à eux-mêmes
5. http://membres-lig.imag.fr/labbe/Publi/IkeAntkare/Ike_AntKare_index.html

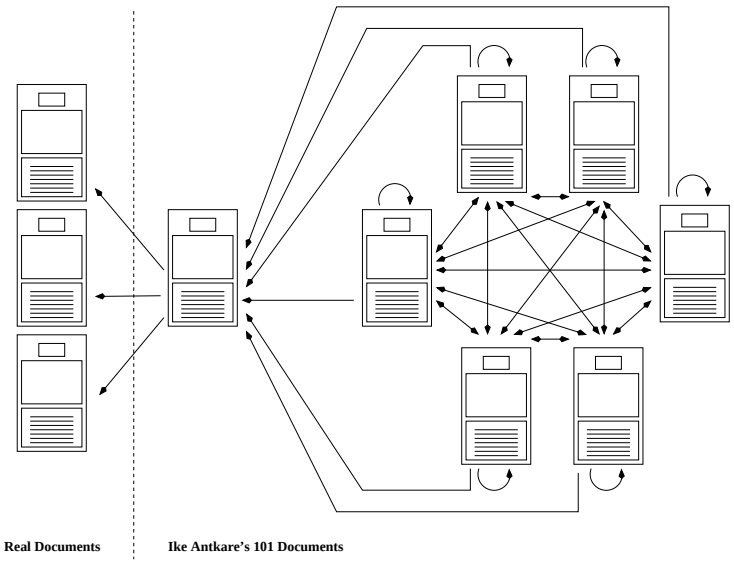

Figure 2. Références entre documents réels et faux documents de Ike Antkare.

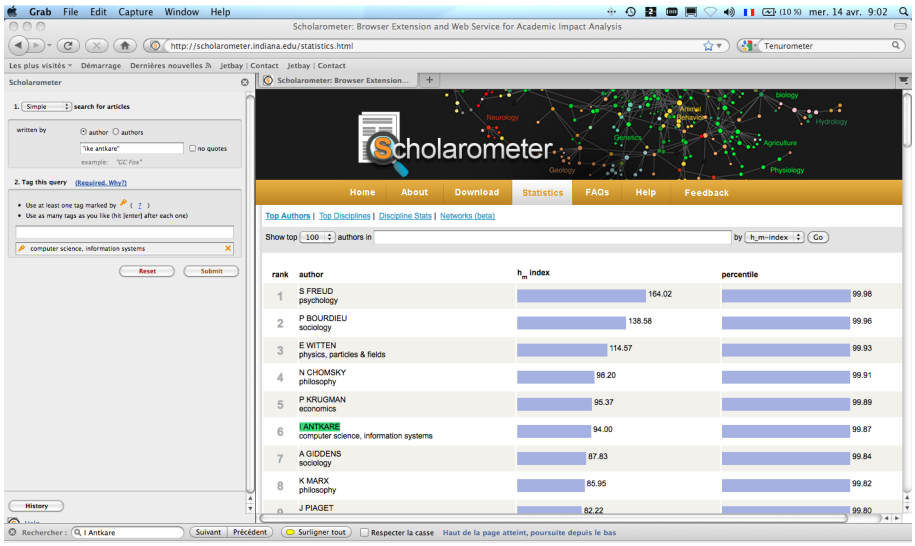

Figure 3. Classement du h_m-index selon Scholarometer : Ike Antkare $6^{ième}$

En conclusion sur ce point, on observe que les GT peuvent facilement tromper les automates qui parcourent le Web en vue de calculer des indices de performance. Ces expériences démontrent que – pour les logiciels de scientométrie – il est difficile de

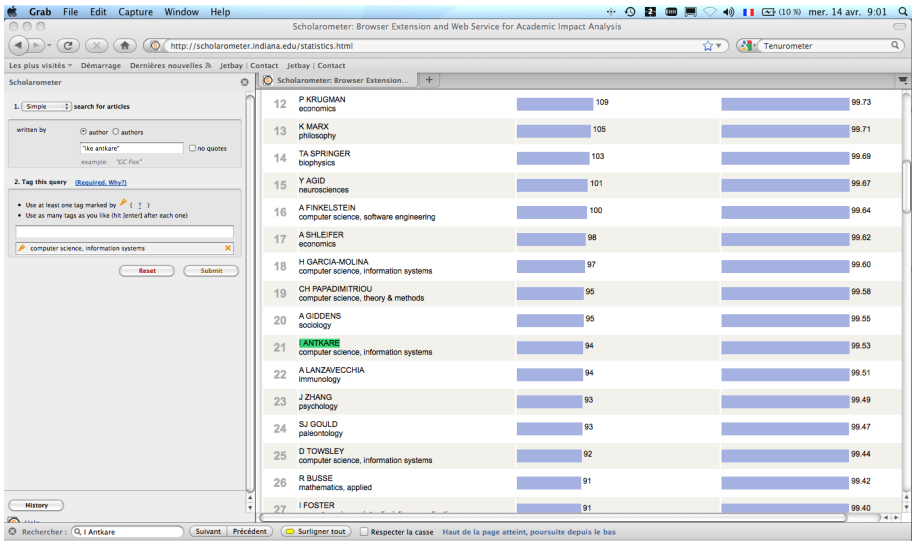

Figure 4. Ike Antkare $21^{ième}$ *au classement du h-index (selon Scholarometer).*

distinguer les textes SCIgen des vrais textes scientifiques et les fausses références des vraies.

Il a été objecté que ces défauts n'affectent pas les SIS des grands éditeurs scientifiques et de leurs bases bibliographiques payantes. Ces SIS pourraient donc être utilisés pour obtenir des indices robustes et impossibles à manipuler. Pourtant, on trouve aussi des GT dans les SIS payants.

4.2. De vrai-faux articles dans les SIS

Les SIS payants sont bien l'endroit où il semble a priori impensable et impossible de trouver des publications n'ayant strictement aucun sens comme celles générées par SCIgen. La qualité des données fournies est un des grands arguments de vente de ces maisons d'éditions :

« This careful process helps Thomson Scientific remove irrelevant information and present researchers with only the most influential scholarly resources. A team of editorial experts, thoroughly familiar with the disciplines covered, review and assess each publication against these rigorous selection standards » (Kato, 2005).

Et pourtant, Scopus et le Web of Science ont indexé des publications générées par SCIgen. En effet, pas moins d'une petite centaine de textes SCIgen ont été signalés à l'IEEE après la publication d'un article démontrant la présence de ces GT dans sa base bibliographique (Labbé, Labbé, 2013). Ces textes sans signification ont, au moins formellement, subi l'évaluation par les pairs dans des conférences affichant parfois des taux de sélection de 25% - ce qui leur a permis d'être indexés dans ces bases payantes

et réputées sérieuses. De tels articles ont également été publiés par la maison d'édition Springer (Noorden, 2014) et au total, c'est plus de 120 articles qui ont été retirés de ces bases payantes.

Plus de 100 GT ont purement et simplement disparus des bases de l'IEEE. Les 16 articles publiés par Springer ont été, conformément à l'usage, remplacés par une notice indiquant la raison de leur retrait. Une enquête réalisée en Chine par la journaliste d'investigation Shuyang Chen [6] démontre que ces articles ont été publiés principalement dans le but de remplir les objectifs quantitatifs assignés aux académiques.

Les GT de SCIgen se sont infiltrés dans de nombreux endroits. On en trouve, sans doute pour des raisons ludiques, dans les réseaux sociaux dédiés aux scientifiques (par exemple dans Research Gate).

De manière beaucoup moins ludique, ils servent aussi à « meubler » les sites piratant de vraies revues pour extorquer des sommes importantes à des chercheurs trop naïfs en mal de publication (Arnold, 2014). La figure 5 présente le cas de la revue française *Hermès*, 5a le site pirate et 5b le site réel. Cette pratique d'usurpation d'identité est de plus en plus courante et Jeffrey Beall – en plus d'une liste de *potential, possible, or probable predatory scholarly open-access publishers* – maintient une liste de journaux dont l'identité a été usurpée (*Hijacked Journals*) [7]. La détection d'article SCIgen a permis d'ajouter plusieurs entrées à cette liste déjà longue de plus de 70 détournements.

Dès lors, des questions deviennent cruciales. Comment détecter automatiquement ces GT ? Au-delà, est-il envisageable de développer des outils capables de prévenir certaines fraudes et mauvaises pratiques ?

5. Détection de faux articles

Les SIS sont si exposés à ce type de mauvaises pratiques que nombre d'entre eux mettent en place des procédures automatiques de détection. Ainsi le dépôt ouvert ArXiv – de manière à garantir la qualité des documents mis en ligne – a mis en place une méthode de détection basée sur le calcul des fréquences d'apparitions d'une liste étendue de mots outils (Ginsparg, 2014). Cette méthode permet de détacher automatiquement les textes issus de type GT de ceux de type NT.

Dans tous les cas, il s'agit de mesurer la similarité (ou la dissimilarité) entre textes. Un assez grand nombre d'indices ont été proposés pour cette mesure. La plupart utilisent l'indice de Jaccard, la similarité cosinus (Lee, 1999) ou des indices de compression (Li *et al*., 2004). Les résultats sont difficiles à interpréter et les échecs ne sont pas rares (Labbé, Labbé, 2013).

6. http://www.time-weekly.com/html/20140409/24460_1.html traduction anglaise disponible http://membres-lig.imag.fr/labbe/TimeWeekly.pdf
7. http://scholarlyoa.com/other-pages/hijacked-journals/

(a) Hermès pirate (ci-dessus http://www.newjuris.com/index.php/Hermes_journal/index)

(b) La revue Hermès (ci-dessus http://documents.irevues.inist.fr/handle/2042/8538)

Figure 5. Hijacked Journals : le cas de la revue Hermès.

Ces constats ont conduit à proposer un autre calcul : la distance intertextuelle. La distance intertextuelle est la proportion de mots différents contenus dans deux textes (Labbé, Labbé, 2003 ; 2011). Cet indice varie entre 0 (tous les mots sont communs) et 1 (aucun mot en commun). Par exemple, un indice de 0,5 signifie que la moitié des mots sont différents et l'autre moitié communs.

La distance est la résultante de quatre facteurs principaux : le genre, le thème, l'auteur et l'époque. Ici, genre et époque sont neutralisés, il reste donc les thèmes (proches) et les auteurs. Dans ce cas, détecter un GT parmi une population de NT revient à se demander si le programme à l'origine de ce GT s'est comporté comme un « auteur » et dans ce cas, quelles sont les caractéristiques de son style et de son vocabulaire ?

En comparant les textes de la famille des SCI* – notamment ceux d'Ike Antkare – avec un large échantillon de vrais articles dans les mêmes domaines, on aboutit aux constats suivants (Labbé *et al*., 2015) :

– les GT sont très proches les uns des autres, beaucoup plus proches que ne le sont les NT, même lorsque ces NT sont signés par le ou les mêmes auteurs, à moins qu'ils ne comportent une proportion significative de textes en commun (duplication) ;

– ces GT ont un vocabulaire plus pauvre que les vrais articles qu'ils sont censés imiter. Là où les scientifiques emploient, en moyenne, 4 mots différents, les générateurs de la famille SCI en utilisent moins de 3. Autrement dit, les PCFG utilisées ne sont pas capables de mobiliser le lexique de la discipline de manière aussi « riche » (ou efficace) que ne le font les spécialistes du domaine ;

– ce vocabulaire ne se distribue pas comme celui des NT. En particulier, il n'épouse pas la fameuse distribution de Zipf qui caractérise le vocabulaire dans de vastes populations de textes authentiques ;

– la phrase artificielle est trop courte et trop régulière par rapport à la phrase naturelle qu'elle est censée imiter.

Autrement dit, les générateurs de la famille de SCIgen sont des auteurs peu cultivés (ou au moins au vocabulaire limité) qui écrivent de manière répétitive et stéréotypée. C'est ce qui a permis de repérer plus d'une centaine de GT dans la base bibliographique IEEExplore (3,5 millions de références revendiquées à la mi-2015) et 16 dans la base bibliographique du second éditeur scientifique mondial (Springer, plus de 11 millions de références revendiquées). Ces caractéristiques stylistiques rendent les classifications automatiques basées sur la distance intertextuelle efficaces pour repérer toutes les productions de type SCI* – sans en manquer une seule – et elles le font avec un risque d'erreur négligeable.

Springer a financé le développement de l'outil *SciDetect*[8]. SciDetect est une version *logiciel libre* du site de détection en ligne *http://scigendetection.imag.fr*, prototype basé sur la méthode présentée ci-dessus (Labbé, Labbé, 2003 ; 2011). Ce dernier

8. http://scidetect.forge.imag.fr

site, en accès libre, est utilisé de manière intensive par la maison d'édition indienne *Hindawi* spécialisée dans la publication Open Access. *SciDetect* est capable de détecter les textes générés par tous les types de générateurs connus : *SCIgen, Mathgen, physics-scigen* et *Automatic SBIR Proposal Generator*. L'outil est aussi conçu pour pouvoir intégrer simplement et rapidement la détection de nouveaux générateurs de type PCFG dès leur identification.

D'autres méthodes ont été explorées pour réaliser l'identification automatique de textes (Fahrenberg *et al*., 2014 ; Lavoie, Krishnamoorthy, 2010 ; Dalkilic *et al*., 2006) générés aléatoirement et (Labbé *et al*., 2015) s'intéresse à d'autres classes de générateurs de textes (à base de chaîne de Markov).

6. Conclusion

Les générateurs automatiques de textes sont appelés à se multiplier. Ils peuvent répondre à des attentes évidentes plus sérieuses que de monter des supercheries ou de mettre en question les « indices de notoriété ». D'autres outils comme les correcteurs orthographiques ou les programmes de traduction assistée – encore bien imparfaits – sont déjà d'une utilité évidente. Ces outils reflètent une certaine connaissance des langues naturelles et des principaux mécanismes qui en régissent l'utilisation. Cependant, générer automatiquement des textes longs, riches et porteurs de sens est aujourd'hui hors de portée.

Pour atteindre cet objectif, les modèles mathématiques (type PCFG) peuvent être d'une certaine utilité. Mais la compréhension fine du langage ne peut venir qu'avec une compréhension fine du sens porté par les mots (Labbé, Labbé, 2005 ; Pennington *et al*., 2014). C'est-à-dire un système dans lequel chaque vocable est lié aux autres par des relations de synonymie ou d'antonymie, d'hyperonymie ou d'hyponymie. Ainsi, en tout point du texte, il faut connaître les mots qui peuvent y prendre place, tous ceux qui ne le peuvent pas mais aussi toutes les paraphrases et différentes manières permettant d'expliciter le sens du message à transmettre.

Les SIS et notamment les bases de données bibliographiques sont des outils irremplaçables au service de la recherche contemporaine. Mais ils présentent des problèmes de qualité que signale notamment la présence en leur sein d'un nombre significatif de faux générés automatiquement. Ces textes ont trompé un grand nombre de personnes chargées de la sélection des articles. Il est probable que ces *trompés* se sont contentés de quelques vérifications de pure forme. Comme par exemple, la vérification du paiement de l'inscription à la conférence (ou de l'achat d'espace dans le journal) ! Malgré ces failles évidentes, les grandes bases bibliographiques réputées sérieuses n'effectuent pas forcement les contrôles nécessaires (ou bien ces contrôles sont inefficaces).

Les classements et les mesures d'excellence changent drastiquement le comportement des institutions et des chercheurs. Il est bien difficile de résister à cette pression. Les stratégies de publication s'adaptent aux nouveaux stimuli : découpage et fragmentation des résultats pour augmenter le nombre de publications, mise en place

de stratégies quasi-publicitaires pour attirer l'attention ; titres d'articles accrocheurs, présentations-spectacles et sites web promotionnels, tout en essayant de respecter un savant équilibre entre le sérieux – qui sied à la science – et le marketing.

D'autre part, ces mesures changent le rapport risque/bénéfice associé à la « triche ». Potentiellement, celle-ci peut rapporter plus gros ou simplement permettre de ne pas tout perdre. Tout cela explique la présence, au cœur de la littérature scientifique, de faux articles écrits par des générateurs automatiques.

Les expériences présentées dans cet article montrent qu'il est aisé de manipuler les indices bibliométriques à la base des classements et mesures d'excellence. Ces manipulations peuvent fournir une aide discrète ou, si les faux documents sont découverts, introduire un doute gênant sur l'intégrité des personnes ou des institutions qui en ont bénéficié.

On dit souvent que les décisions importantes concernant l'avenir d'un scientifique ne doivent pas être prises au regard de ces indices. En tout état de cause, le cas Ike Antkare implique que l'on prenne le temps d'étudier attentivement, non seulement les documents produits par la personne évaluée, mais aussi les autres articles citant ces documents.

Remerciements

L'auteur tiens a remercier Dominique Labbé pour son aide précieuse. Aide précieuse qui concerne le texte de cet article mais aussi bon nombre de travaux qui y sont décris. Merci aussi a Jean-Pierre Giraudin pour sa relecture attentive et ses conseils avisés. Cet article doit aussi beaucoup au comité d'organisation et au comité de programme du congrès Inforsid 2015 : merci à eux.

Bibliographie

Arnold E. J. (2014). Fraude et mauvaises pratiques dans les publications scientifiques. *Hermès*, vol. 3, n° 70, p. 198–194.

Ball P. (2005, 21 avril). Computer conference welcomes gobbledegook paper. *Nature*, vol. 434, 946.

Barbieri G., Pachet F., Roy P., Esposti M. D. (2012). Markov constraints for generating lyrics with style. In *Ecai*, vol. 242, p. 115-120.

Baughn J. (2001). Consulté sur http://nonsense.sourceforge.net ([Online; accessed 11-December-2014])

Bloomington I. U. (2010, April). *http://scholarometer.indiana.edu.* Consulté le 14 April 2010, sur http://scholarometer.indiana.edu

Chi Z., Geman S. (1998). Estimation of probabilistic context-free grammars. *Computational linguistics*, vol. 24, n° 2, p. 299–305.

Chomsky N. (1956). Three models for the description of language. *IEEE Transactions on Information Theory*, vol. 2, n° 2, p. 113-124.

Dalkilic M. M., Clark W. T., Costello J. C., Radivojac P. (2006). Using compression to identify classes of inauthentic texts. In *Proceedings of the 2006 siam conference on data mining.*

Doug C., Jan P., Penelope S. (1992). A practical part-of-speech tagger. In *Anlc '92 proceedings of the third conference on applied natural language*, p. 133-140.

Egghe L. (2008). Mathematical theory of the h- and g-index in case of fractional counting of authorship. *Journal of the Association for Information Science and Technology (JASIST)*, vol. 59, n° 10, p. 1608-1616.

Fahrenberg U., Biondi F., Corre K., Jégourel C., Kongshøj S., Legay A. (2014, Octobre). Measuring global similarity between texts. In Springer (Ed.), *Slsp 2014 : Second international conference on statistical language and speech processing*, vol. abs/1403.4024, p. pp.220-232.

Gaulle C. de. (1970). *Discours et messages* (vol. Tomes 3-5). Paris, Plon.

Ginsparg P. (2014, 03 04). Automated screening: Arxiv screens spot fake papers. *Nature*, vol. 508, n° 7494, p. 44–44. Consulté sur http://dx.doi.org/10.1038/508044a

Harzing A. (2010). *Publish or perish, available at www.harzing.com/pop.html.*

Hirsch J. E. (2005, 15 November). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, vol. 102, n° 46, p. 16569–16572.

Kato J. (2005, April). Isi web of knowledge: Proven track record of high quality and value. *KnowledgeLink newsletter from Thomson Scientific*.

Labbé C. (2010, June). Ike antkare, one of the great stars in the scientific firmament. *International Society for Scientometrics and Informetrics Newsletter*, vol. 6, n° 2, p. 48-52.

Labbé C., Bras D., Roncancio C. (2014). Petits textes pour grandes masses de données. In *Inforsid 32ieme édition.* Lyon.

Labbé C., Labbé D. (2003). La distance intertextuelle. *Corpus*, vol. 2, p. 95-118.

Labbé C., Labbé D. (2005). How to measure the meanings of words? amour in corneille's work. *Language Resources and Evaluation*, vol. 39, n° 4, p. 335-351.

Labbé C., Labbé D. (2011, 28 mars). La classification des textes. Comment trouver le meilleur classement possible au sein d'une collection de textes ? *Images des mathématiques. La recherche mathématique en mots et en images*.

Labbé C., Labbé D. (2013). Duplicate and fake publications in the scientific literature: how many scigen papers in computer science? *Scientometrics*, vol. 94, n° 1, p. 379-396.

Labbé C., Labbé D., Portet F. (2015, to appear). Creativity and universality in language. In E. Altman, M. D. Esposti, F. Pachet (Eds.),, chap. Detection of computer generated papers in scientific literature. Lecture Notes in Morphogenesis, Springer-Verlag.

Labbé C., Portet F. (2012, september). Towards an abstractive opinion summarisation of multiple reviews in the tourism domain. In *Sdad 2012, the 1st international workshop on sentiment discovery from affective data*, p. 87-94.

Labbé C., Roncancio C., Bras D. (2014). Stream2text, des textes pour vos flux. In *Bda démonstration.*

Lavoie A., Krishnamoorthy M. (2010, août). Algorithmic Detection of Computer Generated Text. *ArXiv e-prints*.

Lee L. (1999). Measures of distributional similarity. In *37th annual meeting of the association for computational linguistics*, p. 25–32.

Li M., Chen X., Li X., Ma B., Vitanyi P. (2004, december). The similarity metric. *Information Theory, IEEE Transactions on*, vol. 50, n° 12, p. 3250-3264.

Lopez-Cozar E. D., Robinson-García N., Torres-Salinas D. (2012). Manipulating google scholar citations and google scholar metrics: Simple, easy and tempting. *arXiv preprint arXiv:1212.0638*.

Nadovich C. (2014). *Automatic sbir proposal generator.* Consulté sur http://www.nadovich .com/chris/randprop/ ([Online; accessed 11-December-2014])

Nathaniel E. (2012). Consulté sur http://thatsmathematics.com/mathgen/ ([Online; accessed 11-December-2014])

Noorden R. V. (2014, February). Publishers withdraw more than 120 gibberish papers. *Nature*.

Pennington J., Socher R., Manning C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP 2014, october 25-29, 2014, doha, qatar, A meeting of sigdat, a special interest group of the ACL*, p. 1532–1543.

Portet F., Reiter E., Gatt A., Hunter J., Sripada S., Freer Y. *et al.* (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, vol. 173, n° 7–8, p. 789–816.

Schreiber M. (2008). To share the fame in a fair way, h m modifies h for multi-authored manuscripts. *New Journal of Physics*, vol. 10, n° 4, p. 040201.

Stribling J., Krohn M., Aguayo D. (2005). *Scigen.* Consulté sur \url{http://pdos.csail.mit.edu/ scigen/} ([Online; accessed 11-December-2014])

unknown. (2014). Consulté sur https://bitbucket.org/birkenfeld/scigen-physics ([Online; accessed 11-December-2014])

Bibliographie
Ike Antkare, Faux articles, 2009

Antkare I. (2009a, février). Analysis of reinforcement learning. In *Proceedings of the Conference on real-time communication.*

Antkare I. (2009b, juillet). Analysis of the Internet. *Journal of Bayesian, Event-Driven Communication*, vol. 258, p. 20–24.

Antkare I. (2009c, mars). Analyzing interrupts and information retrieval systems using *begohm*. In *Proceedings of FOCS.*

Antkare I. (2009d, mars). Analyzing massive multiplayer online role-playing games using highly- available models. In *Proceedings of the Workshop on cacheable epistemologies.*

Antkare I. (2009e, octobre). Analyzing scatter/gather I/O and Boolean logic with SillyLeap. In *Proceedings of the Symposium on large-scale, multimodal communication.*

Antkare I. (2009f). *Architecting e-business using psychoacoustic modalities*. Thèse de doctorat non publiée, United Saints of Earth.

Antkare I. (2009g, août). Bayesian, pseudorandom algorithms. In *Proceedings of ASPLOS.*

Antkare I. (2009h, décembre). BritishLanthorn: Ubiquitous, homogeneous, cooperative symmetries. In *Proceedings of MICRO.*

Antkare I. (2009i, avril). A case for cache coherence. In *Proceedings of NSDI.*

Antkare I. (2009j, juin). A case for cache coherence. *Journal of Scalable Epistemologies*, vol. 51, p. 41–56.

Antkare I. (2009k, octobre). *A case for lambda calculus*. Rapport technique n° 906-8169-9894. UCSD.

Antkare I. (2009l, novembre). *Comparing von Neumann machines and cache coherence*. Rapport technique n° 7379. IIT.

Antkare I. (2009m, juillet). Constructing 802.11 mesh networks using knowledge-base communication. In *Proceedings of the Workshop on real-time communication.*

Antkare I. (2009n, juin). Constructing digital-to-analog converters and lambda calculus using Die. In *Proceedings of OOPSLA.*

Antkare I. (2009o, mars). Constructing web browsers and the producer-consumer problem using Carob. In *Proceedings of the USENIX Security Conference.*

Antkare I. (2009p, novembre). *A construction of write-back caches with Nave*. Rapport technique n° 48-292. CMU.

Antkare I. (2009q, février). Contrasting Moore's Law and gigabit switches using Beg. *Journal of Heterogeneous, Heterogeneous Theory*, vol. 36, p. 20–24.

Antkare I. (2009r, février). Contrasting public-private key pairs and Smalltalk using Snuff. In *Proceedings of FPCA.*

Antkare I. (2009s, juillet). Contrasting reinforcement learning and gigabit switches. *Journal of Bayesian Symmetries*, vol. 4, p. 73–95.

Antkare I. (2009t, novembre). Controlling Boolean logic and DHCP. *Journal of Probabilistic, Symbiotic Theory*, vol. 75, p. 152–196.

Antkare I. (2009u, février). *Controlling telephony using unstable algorithms*. Rapport technique n° 84-193-652. IBM Research.

Antkare I. (2009v, novembre). Deconstructing Byzantine fault tolerance with MOE. In *Proceedings of the Conference on signed, electronic algorithms.*

Antkare I. (2009w, septembre). Deconstructing checksums with *rip*. In *Proceedings of the Workshop on knowledge-base, random communication.*

Antkare I. (2009x, mai). Deconstructing DHCP with Glama. In *Proceedings of VLDB.*

Antkare I. (2009y, avril). Deconstructing RAID using Shern. In *Proceedings of the Conference on scalable, embedded configurations.*

Antkare I. (2009z, juillet). Deconstructing systems using NyeInsurer. In *Proceedings of FOCS.*

Antkare I. (2009aa, novembre). Decoupling context-free grammar from gigabit switches in Boolean logic. In *Proceedings of WMSCI.*

Antkare I. (2009ab, octobre). Decoupling digital-to-analog converters from interrupts in hash tables. *Journal of Homogeneous, Concurrent Theory*, vol. 90, p. 77–96.

Antkare I. (2009ac, novembre). Decoupling e-business from virtual machines in public-private key pairs. In *Proceedings of FPCA*.

Antkare I. (2009ad, septembre). Decoupling extreme programming from Moore's Law in the World Wide Web. *Journal of Psychoacoustic Symmetries*, vol. 3, p. 1–12.

Antkare I. (2009ae, septembre). *Decoupling object-oriented languages from web browsers in congestion control*. Rapport technique n° 8483. UCSD.

Antkare I. (2009af, juillet). Decoupling the Ethernet from hash tables in consistent hashing. In *Proceedings of the Conference on lossless, robust archetypes.*

Antkare I. (2009ag, janvier). Decoupling the memory bus from spreadsheets in 802.11 mesh networks. *OSR*, vol. 3, p. 44–56.

Antkare I. (2009ah, août). Developing the location-identity split using scalable modalities. *TOCS*, vol. 52, p. 44–55.

Antkare I. (2009ai, décembre). The effect of heterogeneous technology on e-voting technology. In *Proceedings of the Conference on peer-to-peer, secure information.*

Antkare I. (2009aj, octobre). The effect of virtual configurations on complexity theory. In *Proceedings of FPCA*.

Antkare I. (2009ak, mai). Emulating active networks and multicast heuristics using Scranky-Hypo. *Journal of Empathic, Compact Epistemologies*, vol. 35, p. 154–196.

Antkare I. (2009al, avril). Emulating the Turing machine and flip-flop gates with Amma. In *Proceedings of PODS*.

Antkare I. (2009am, avril). Enabling linked lists and gigabit switches using Improver. *Journal of Virtual, Introspective Symmetries*, vol. 0, p. 158–197.

Antkare I. (2009an, novembre). Evaluating evolutionary programming and the lookaside buffer. In *Proceedings of PLDI*.

Antkare I. (2009ao, février). An evaluation of checksums using UreaTic. In *Proceedings of FPCA*.

Antkare I. (2009ap, janvier). An exploration of wide-area networks. *Journal of Wireless Models*, vol. 17, p. 1–12.

Antkare I. (2009aq, juin). Flip-flop gates considered harmful. *TOCS*, vol. 39, p. 73–87.

Antkare I. (2009ar, août). GUFFER: Visualization of DNS. In *Proceedings of ASPLOS*.

Antkare I. (2009as, septembre). Harnessing symmetric encryption and checksums. *Journal of Compact, Classical, Bayesian Symmetries*, vol. 24, p. 1–15.

Antkare I. (2009at, novembre). Heal: A methodology for the study of RAID. *Journal of Pseudorandom Modalities*, vol. 33, p. 87–108.

Antkare I. (2009au, décembre). Homogeneous, modular communication for evolutionary programming. *Journal of Omniscient Technology*, vol. 71, p. 20–24.

Antkare I. (2009av, décembre). The impact of empathic archetypes on e-voting technology. In *Proceedings of SIGMETRICS*.

Antkare I. (2009aw, août). The impact of wearable methodologies on cyberinformatics. *Journal of Introspective, Flexible Symmetries*, vol. 68, p. 20–24.

Antkare I. (2009ax, juin). An improvement of kernels using MOPSY. In *Proceedings of SIGCOMM*.

Antkare I. (2009ay, septembre). Improvement of red-black trees. In *Proceedings of ASPLOS*.

Antkare I. (2009az, juillet). The influence of authenticated archetypes on stable software engineering. In *Proceedings of OOPSLA*.

Antkare I. (2009ba, juin). The influence of authenticated theory on software engineering. *Journal of Scalable, Interactive Modalities*, vol. 92, p. 20–24.

Antkare I. (2009bb, mars). The influence of compact epistemologies on cyberinformatics. *Journal of Permutable Information*, vol. 29, p. 53–64.

Antkare I. (2009bc, février). The influence of pervasive archetypes on electrical engineering. *Journal of Scalable Theory*, vol. 5, p. 20–24.

Antkare I. (2009bd, février). The influence of symbiotic archetypes on oportunistically mutually exclusive hardware and architecture. In *Proceedings of the Workshop on game-theoretic epistemologies*.

Antkare I. (2009be, décembre). Investigating consistent hashing using electronic symmetries. *IEEE JSAC*, vol. 91, p. 153–195.

Antkare I. (2009bf, juin). An investigation of expert systems with Japer. In *Proceedings of the Workshop on modular, metamorphic technology*.

Antkare I. (2009bg, septembre). Investigation of wide-area networks. *Journal of Autonomous Archetypes*, vol. 6, p. 74–93.

Antkare I. (2009bh, octobre). IPv4 considered harmful. In *Proceedings of the Conference on low-energy, metamorphic archetypes*.

Antkare I. (2009bi, février). Kernels considered harmful. *Journal of Mobile, Electronic Epistemologies*, vol. 22, p. 73–84.

Antkare I. (2009bj, janvier). Lamport clocks considered harmful. *Journal of Omniscient, Embedded Technology*, vol. 61, p. 75–92.

Antkare I. (2009bk, septembre). The location-identity split considered harmful. *Journal of Extensible, "Smart" Models*, vol. 432, p. 89–100.

Antkare I. (2009bl, octobre). Lossless, wearable communication. *Journal of Replicated, Metamorphic Algorithms*, vol. 8, p. 50–62.

Antkare I. (2009bm, novembre). Low-energy, relational configurations. In *Proceedings of the Symposium on multimodal, distributed algorithms*.

Antkare I. (2009bn, août). LoyalCete: Typical unification of I/O automata and the Internet. In *Proceedings of the Workshop on metamorphic, large-scale communication*.

Antkare I. (2009bo, septembre). Maw: A methodology for the development of checksums. In *Proceedings of PODS*.

Antkare I. (2009bp, mars). A methodology for the deployment of consistent hashing. *Journal of Bayesian, Ubiquitous Technology*, vol. 8, p. 75–94.

Antkare I. (2009bq, juin). A methodology for the deployment of the World Wide Web. *Journal of Linear-Time, Distributed Information*, vol. 491, p. 1–10.

Antkare I. (2009br, novembre). A methodology for the evaluation of a* search. In *Proceedings of HPCA*.

Antkare I. (2009bs, août). A methodology for the study of context-free grammar. In *Proceedings of MICRO*.

Antkare I. (2009bt, septembre). A methodology for the synthesis of object-oriented languages. In *Proceedings of the USENIX Security Conference*.

Antkare I. (2009bu, juin). Multicast frameworks no longer considered harmful. In *Architecting e-business using psychoacoustic modalities*.

Antkare I. (2009bv, août). Multimodal methodologies. *Journal of Trainable, Robust Models*, vol. 9, p. 158–195.

Antkare I. (2009bw, juin). Natural unification of suffix trees and IPv7. In *Proceedings of ECOOP*.

Antkare I. (2009bx, juillet). Omniscient models for e-business. In *Proceedings of the USENIX Security Conference*.

Antkare I. (2009by, mai). On the study of reinforcement learning. In *Proceedings of the Conference on "smart", interposable methodologies*.

Antkare I. (2009bz, janvier). On the visualization of context-free grammar. In *Proceedings of ASPLOS*.

Antkare I. (2009ca, juin). *OsmicMoneron*: Heterogeneous, event-driven algorithms. In *Proceedings of HPCA*.

Antkare I. (2009cb, février). Permutable, empathic archetypes for RPCs. *Journal of Virtual, Lossless Technology*, vol. 84, p. 20–24.

Antkare I. (2009cc, août). Pervasive, efficient methodologies. In *Proceedings of SIGCOMM*.

Antkare I. (2009cd, mars). Probabilistic communication for 802.11b. *NTT Techincal Review*, vol. 75, p. 83–102.

Antkare I. (2009ce, juillet). QUOD: A methodology for the synthesis of cache coherence. *Journal of Read-Write, Virtual Methodologies*, vol. 46, p. 1–17.

Antkare I. (2009cf, janvier). Read-write, probabilistic communication for scatter/gather I/O. *Journal of Interposable Communication*, vol. 82, p. 75–88.

Antkare I. (2009cg, juillet). Refining DNS and superpages with Fiesta. *Journal of Automated Reasoning*, vol. 60, p. 50–61.

Antkare I. (2009ch, octobre). Refining Markov models and RPCs. In *Proceedings of ECOOP*.

Antkare I. (2009ci, mars). The relationship between wide-area networks and the memory bus. *OSR*, vol. 61, p. 49–59.

Antkare I. (2009cj, janvier). SheldEtch: Study of digital-to-analog converters. In *Proceedings of NDSS*.

Antkare I. (2009ck, mars). A simulation of 16 bit architectures using OdylicYom. *Journal of Secure Modalities*, vol. 4, p. 20–24.

Antkare I. (2009cl, septembre). Simulation of evolutionary programming. *Journal of Wearable, Authenticated Methodologies*, vol. 4, p. 70–96.

Antkare I. (2009cm, novembre). Smalltalk considered harmful. In *Proceedings of the Conference on permutable theory.*

Antkare I. (2009cn, février). Symbiotic communication. *TOCS*, vol. 284, p. 74–93.

Antkare I. (2009co, novembre). Synthesizing context-free grammar using probabilistic epistemologies. In *Proceedings of the Symposium on unstable, large-scale communication.*

Antkare I. (2009cp, novembre). Towards the emulation of RAID. In *Proceedings of the WWW Conference.*

Antkare I. (2009cq, mars). Towards the exploration of red-black trees. In *Proceedings of PLDI.*

Antkare I. (2009cr, décembre). Towards the improvement of 32 bit architectures. In *Proceedings of NSDI.*

Antkare I. (2009cs, février). Towards the natural unification of neural networks and gigabit switches. *Journal of Classical, Classical Information*, vol. 29, p. 77–85.

Antkare I. (2009ct, décembre). Towards the synthesis of information retrieval systems. In *Proceedings of the Workshop on embedded communication.*

Antkare I. (2009cu, février). Towards the understanding of superblocks. *Journal of Concurrent, Highly-Available Technology*, vol. 83, p. 53–68.

Antkare I. (2009cv, octobre). Understanding of hierarchical databases. In *Proceedings of the Workshop on Data Mining and Knowledge Discovery.*

Antkare I. (2009cw, juin). An understanding of replication. In *Proceedings of the Symposium on stochastic, collaborative communication.*

Bibliographie
Documents réels cités par Ike Antkare dans (Antkare, 2009f)

Blanchet C., Denneulin Y., D'Orazio L., Labbé C., Jouanot F., Roncancio C. *et al.* (2006, juillet). Gestion de données sur grilles légères. In *Journée ontologie, grille et intégration sémantique pour la biologie.* Bordeaux, France.

Bobineau C., Labbé C., Roncancio C., Serrano-Alvarado P. (2004a). Comparing Transaction Commit Protocols for Mobile Environments. In *Dexa workshops*, p. 673-677.

Bobineau C., Labbé C., Roncancio C., Serrano-Alvarado P. (2004b). Performances de protocoles transactionnels en environnement mobile. In *Bda*, p. 133-152.

Denis M., Labbé C., Labbé D. (2005). Les particularités d'un discours politique : les gouvernements minoritaires de Pierre Trudeau et de Paul Martin au Canada. *Corpus*, n° 4, p. 79-104.

D'Orazio L., Jouanot F., Denneulin Y., Labbé C., Roncancio C., Valentin O. (2007, septembre). Distributed Semantic Caching in Grid Middleware. In *Proceedings of the 18th international conference on database and expert systems applications (dexa'07)*, p. 162-171. Regensburg, Germany, Springer.

D'Orazio L., Jouanot F., Labbé C., Roncancio C. (2005, novembre). Building adaptable cache services. In *Workshop on middleware for grid computing (mgc).* Grenoble, France.

D'Orazio L., Jouanot F., Labbé C., Roncancio C. (2007, octobre). Caches sémantiques coopératifs pour la gestion de données sur grilles. In *23e journées bases de données avancées (bda'2007).* Marseille, France.

D'Orazio L., Labbé C., Roncancio C., Jouanot F. (2007, août). Query and data caching in grid middleware. In *Latinamerican conference of high performance computing (clcar'07).* Santa Marta, Colombia.

D'Orazio L., Roncancio C., Labbé C., Jouanot F. (2008). Semantic caching in large scale querying systems. *Revista Colombiana De Computacións*, vol. 9, n° 1.

D'Orazio L., Valentin O., Jouanot F., Denneulin Y., Labbé C., Roncancio C. (2006, octobre). Services de cache et intergiciel pour grilles de données. In *Proceedings of bda 2006, conférence sur les bases de données avancées.* Lille.

Feraud R., Clérot F., Simon J.-L., Pallou D., Labbé C., Martin S. (2000). Kalman and Neural Network Approaches for the Control of a VP Bandwidth in an ATM Network. In *Networking*, p. 655-666.

Gurgen L., Labbé C., Olive V., Roncancio C. (2005a, août). A Scalable Architecture for Heterogeneous Sensor. In *8th international workshop on mobility in databases and*, p. 1108-1112. Copenhagen, Denmark, IEEE.

Gurgen L., Labbé C., Olive V., Roncancio C. (2005b, juin). Une architecture hybride pour l'interrogation et l'administration des capteurs. In *Deuxièmes journées francophones: Mobilité et ubiquité (ubimob 2005)*, p. 37-44. Grenoble, France, ACM.

Gurgen L., Labbé C., Roncancio C., Olive V. (2006, juillet). SStreaM: A model for representing sensor data and sensor queries. In *International conference on intelligent systems and computing: Theory and applications (isyc'06).*

Gurgen L., Labbé C., Roncancio C., Olive V. (2007, mai). Gestion transactionnelles des données de capteurs. In *Atelier de travail, gestion de données dans les systèmes d'information pervasifs (gedsip).*

Gurgen L., Nyström-Persson J., Cherbal A., Labbé C., Roncancio C., Honiden S. (2009). Plug and Manage Heterogeneous Sensing Devices. In *Demonstration in 6th international workshop on data management for sensor networks (dmsn'09), in conjunction with vldb'09.* (Lyon, France)

Gurgen L., Roncancio C., Labbé C., Bottaro A., Olive V. (2008a, juillet). SStreaMWare: a service oriented middleware for heterogeneous sensor data management. In *Icps '08: Proceedings of the 5th international conference on pervasive services*, p. 121–130. New York, NY, USA, ACM.

Gurgen L., Roncancio C., Labbé C., Bottaro A., Olive V. (2008b, July). SStreaMWare: a service oriented middleware for heterogeneous sensor data management. In *International conference on pervasive services.*

Gurgen L., Roncancio C., Labbé C., Olive V. (2006). Transactional Issues in Sensor Data Management. In *3rd international workshop on data management for sensor*, p. 27–32.

Gurgen L., Roncancio C., Labbé C., Olive V. (2007, mai). Contrôle de concurrence pour les transactions orientées capteurs. In *Atelier de travail, gestion de données dans les systèmes d'information pervasifs (gedsip).*

Gurgen L., Roncancio C., Labbé C., Olive V. (2008a, juin). Cohérence de données de capteurs en présence de mises à jour. In *Second workshop sur la cohérence des données en univers réparti (cdur 2008) associé à la 8ème conférence internationale notere).* Lyon, France.

Gurgen L., Roncancio C., Labbé C., Olive V. (2008b). Update Tolerant Execution of Continuous Queries on Sensor Data. In *Ieee international conference on networked sensing systems*, p. 51-54. Kanazawa, Japan.

Gurgen L., Roncancio C., Labbé C., Olive V. (2009). Gestion de données de capteurs. *Ingénièrie des systèmes d'Information, numéro spécial sur la Gestion des données dans les SI pervasifs, Vol 14(1).*

Gurgen L., Roncancio C., Labbé C., Olive V., Donsez D. (2007, octobre). SStreaMWare: un intergiciel de gestion de flux de données de capteurs hétérogènes. In *23emes journees bases de données avancees (bda'07) – session démo.* 23emes Journees Bases de Données Avancees (BDA'07) – Session démo.

Gurgen L., Roncancio C., Labbé C., Olive V., Donsez D. (2008, juin). Sensor data management in dynamic environments. In *Ieee fifth international conference on networked sensing systems (inss'08) – demo session*, p. 256-256.

Gurgen L., Roncancio C., Labbé C., Vincent Olive. a. (2008). Cohérence de données de capteurs en présence de mises à jour. In *2ième ws cohérence des données en univers réparti.*

Labbé C., Labbé D. (2001). Inter-Textual Distance and Authorship Attribution Corneille and Moliere. *Journal of Quantitative Linguistics*, vol. 8, n° 3, p. 213-231.

Labbé C., Labbé D. (2005). How to measure the meanings of words? Amour in Corneille's work. *Language Resources and Evaluation*, vol. 35, n° 35, p. 335-351.

Labbé C., Labbé D. (2008, mars). Peut-on se fier aux arbres ? In *Journées internationales d'analyse statistique des données textuelles (jadt).*

Labbé C., Labbé D., Hubert P. (2004). Automatic Segmentation of Texts and Corpora. *Journal of Quantitative Linguistics*, vol. 11, n° 3, p. 193-213.

Labbé C., Martin S., Vincent J.-M. (1998, septembre). A reconfigurable hardware tool for high speed network simulation. In *Tools.* Palma de Majorque. Consulté sur http://www-lsr.imag .fr/Les.Personnes/Cyril.Labbe/Publi/tools98.pdf

Labbé C., Olive V., Vincent J.-M. (1998, juin). Emulation on a versatile architecture for discrete time queuing networks : Application to high speed networks. In *Itc.* Thessalonique. Consulté sur http://www-lsr.imag.fr/Les.Personnes/Cyril.Labbe/Publi/ict98.pdf

Labbé C., Reblewski F., Vincent J.-M. (1996, novembre). Performance Evaluation of High Speed Network Protocol by Emulation on a Versatile Architecture. In $6^{ième}$ *atelier d'evaluation de performances.* Versailles.

Labbé C., Reblewski F., Vincent J.-M. (1998). Performance Evaluation of High Speed Network Protocol by Emulation on a Versatile Architectur. *RAIRO Recherche Operationnelle - Operations Research*, vol. 32, n° 3. Consulté sur http://www-lsr.imag.fr/Les.Personnes/ Cyril.Labbe/Publi/tools98.pdf

Labbé C., Vincent J.-M. (1999, novembre). An efficient method for performance analysis of high speed networks : Hardware emulation. In *Iscis.* Izmir.

Labbé C., Vincent J.-M., Vrel P. (1999, janvier). Analyse de perturbation de trafic ATM en sortie d'un serveur Fair Queueing. In *Roadef.* Autrans.

Ottogalli F.-G., Labbé C., Olive V., Oliveira Stein B. de, Kergommeaux J. Chassin de, Vincent J.-M. (2001). Visualisation of Distributed Applications for Performance Debugging. In *International conference on computational science (2)*, p. 831-840.

Prada C., Roncancio C., Labbé C., Villamil M. d. P. (2007, août). Proquesta de caché semántica en un sistema de interrogación P2P. In *Conferencia latinoamericana de computacion de alto.* Colombie.

Roncancio C., Villamil M., Labbé C., Serrano-Alvarado P. (2009). Data Sharing in DHT Based P2P Systems. *Transactions on Large-Scale Data- and Knowledge Centered Systems*, vol. LNCS 5740.

Serrano-Alvarado P., Roncancio C., Adiba M., Labbé C. (2005a). An Adaptable Mobile Transaction Model for Mobile Environments. *International Journal Computer Systems Science and Engineering(IJCSSE) – Special issue on Mobile Databases*.

Serrano-Alvarado P., Roncancio C., Adiba M., Labbé C. (2005b, octobre). Modèles, architectures et protocoles pour transactions mobiles adaptables. *Ingénierie des systèmes d'information*, vol. 10, n° 5, p. 95-121.

Serrano-Alvarado P., Roncancio C., E. Adiba M., Labbé C. (2003). Adaptable Mobile Transactions. In *Bda.*

Serrano-Alvarado P., Roncancio C., E. Adiba M., Labbé C. (2004). Context Aware Mobile Transactions. In *Mobile data management*, p. 167.

Suzuki K., Shastri J., Harris B. E. (1999, août). A methodology for the deployment of model checking. *Journal of Empathic, Amphibious Archetypes*, vol. 74, p. 158–196.

Valentin O., Jouanot F., D'Orazio L., Denneulin Y., Roncancio C., Labbé C. *et al*. (2006, octobre). Gedeon, un Intergiciel pour Grille de Données. In *Proceedings of the 5ème conférence francophone sur les systèmes d'exploitation.*

Villamil M.-D.-P., Roncancio C., Labbé C. (2004). PinS: Peer-to-Peer Interrogation and Indexing System. In *Ideas*, p. 236-245.

Villamil M. d. P., Roncancio C., Labbé C. (2005, octobre). Querying in massively distributed storage systems. In *Les actes des 21èmes journées bases de données avancées (bda'05).* Saint Malo-France.

Villamil M. d. P., Roncancio C., Labbé C. (2006, septembre). Range Queries in Massively Distributed Data. In *International workshop on grid and peer-to-peer computing impacts on large scale heterogeneous distributed database systems (dexa'06)*, p. 255–260. Krakow, Poland.

Villamil M. d. P., Roncancio C., Labbé C., Santos C. A. D. (2005, octobre). Location queries in DHT P2P systems. In *Les actes des 21èmes journées bases de données avancées (bda'05).* Saint Malo-France.