Université de Neuchâtel
Institut de Mathématiques
Institut d'Informatique

SEMINAIRE MATHEMATIQUES ET SOCIETE

Mercredi 14 novembre 2012

L'ordinateur peut-il écrire ?

Cyril Labbé
Laboratoire d'Informatique de Grenoble - Université Joseph Fourier
(cyril.labbe@imag.fr)

Dominique Labbé
Laboratoire PACTE (CNRS - Institut d'Etudes Politiques de Grenoble)
(dominique.labbe@iep.grenoble.fr)

Résumé
Il existe des "générateurs" d'articles scientifiques (notamment en informatique, physique, mathématique ou philosophie). La conférence explique comment marchent ces générateurs et présente une méthode qui permet de détecter les faux articles réalisés avec eux. Pour l'instant, ces articles sont assez faciles à identifier car ils sont dépourvus de sens et surtout parce qu'ils n'ont pas la diversité de vocabulaire et de syntaxe des humains. Pourtant, plusieurs fausses publications se sont faufilées dans des congrès et dans les grandes bases de données bibliographiques payantes qui servent à évaluer les chercheurs. Enfin, il existe plusieurs voies pour améliorer ces générateurs qui seront peut-être, dans le futur, des auxiliaires utiles pour la rédaction des textes.

Abstract
There are some "generators" of scientific papers (including computer science, physics, mathematics or philosophy). We explain how these generators work and we present a method detecting generated texts. Until now, these false papers are fairly easy to identify because they are meaningless and mostly because they do not have the diversity of human vocabulary and syntax. However, several generated papers are appearing in scientific conferences and in the major bibliographic databases such as Scopus or WoK. Finally, there are several ways to improve these generators which may provide in the future, a useful help for text writing.

Newspeak was founded on the English language as we now know it, though many Newspeak sentences, even when not containing newly-created words, would be barely intelligible to an English-speaker of our own day.

George Orwell (1949). The Principles of Newspeak. Appendix of *1984*.

Un sous-préfet doit inaugurer les comices agricoles de son chef-lieu. En manque d'inspiration, il se perd dans la rêverie et finit par faire des vers. Aujourd'hui, A. Daudet pourrait donner une autre fin à son conte : le sous-préfet se connecte sur le "pipotron"[1] et, en quelques clicks, il compose un discours aussi ronflant que creux, donc tout à fait adapté aux circonstances…

L'idée n'est pas neuve. En 1964, R. Escarpit avait imaginé le *Littératron* - ordinateur qui analysait et composait des textes littéraires et que des politiciens utilisaient pour leur campagne électorale dans la petite ville de Pédouillac. Après avoir analysé les conversations des Pédouillacais, le Littératron avait généré le discours, la profession de foi et l'affiche censés faire gagner le candidat…

L'ordinateur pourrait donc écrire à la place des politiciens, des chercheurs, des journalistes en panne d'idées ?

Si l'on s'en tient à des messages brefs et relativement simples, la réponse semble être positive. Qui n'a pas reçu de courriers contenant une offre de crédit, de placement ou d'autres publicités ciblées ? Naturellement, c'est l'ordinateur qui a sélectionné les destinataires dans une base de données. C'est également l'ordinateur qui a généré la lettre en remplissant un certain nombre de "trous" dans un courrier type : date, nom de la personne, adresse, personnalisation de l'offre, etc.

Les principaux générateurs de textes actuellement disponibles fonctionnent selon ce principe. Cette conférence commencera par en évoquer quelques-uns, spécialement le plus connu : SCIgen. Nous examinerons ensuite la question suivante : ces textes générés (TG) ressemblent-ils aux textes "naturels" (TN) qu'ils sont censés imiter ? Ce qui revient à se demander si les communications scientifiques générées automatiquement – des sortes de chimères -, peuvent passer pour de vrais textes ? Pour répondre à cette question, on utilisera une méthode de détection combinant le calcul de distance entre textes et la classification automatique. Cette méthode permet de constater qu'un nombre significatif de chimères se sont glissées dans des congrès et dans les grandes bases bibliographiques payantes.

Enfin nous envisagerons deux points : comment améliorer les capacités des générateurs ? Quelle utilité ?

[1] http://www.pipotron.free.fr/ ou http://bluepsi.free.fr/cybertechno/pipotron/pipotron.html.

1. Les générateurs de textes

Depuis plus d'un demi-siècle, la question de la génération du langage naturel par ordinateur suscite un intérêt croissant. On est ainsi parvenu à programmer des automates capables de parler et de lire un texte. Il existe également des programmes qui aident à résumer un texte à partir de son vocabulaire et de ses phrases caractéristiques, mais un opérateur doit toujours finir le travail. D'autres programmes évoqués en introduction, envoient des courriers personnalisés. D'autres encore peuvent rédiger des bulletins météos, ou le compte rendu d'une rencontre sportive ou d'une séance de bourse (généralement en anglais). Mais il s'agit toujours de textes brefs dans des champs très limités et assez stéréotypés.

En 2005, est apparu le premier véritable générateur automatique de papiers scientifiques (Ball 2005). Il a été réalisé et mis en ligne par des étudiants du Massachussetts Institute of Technology (http://pdos.csail.mit.edu/scigen/). Par la suite, ce programme a été adapté à la physique puis aux mathématiques (MATHgen : http://thatsmathematics.com/mathgen/). On trouvera dans Ghys 2012 un exemple de faux article écrit avec ce programme. Le tableau 1 ci-dessous donne le début de la liste des phrases possibles pour débuter un article SCIgen. Le tableau suivant reproduit le résumé, une figure et le début de la bibliographie d'un article SCIgen. On remarquera que les auteurs cités en bibliographie existent ou ont existé…

Tableau 1. Premiers mots des différentes phrases possibles pour le début d'un texte généré par SCIgen

Many SCI PEOPLE would agree that, had it not been for SCI GENERIC NOUN, ...
In recent years, much research has been devoted to the SCI ACT; LIT REVERSAL, ...
SCI THING MOD and SCI THING MOD, while SCI ADJ in theory, have not until ...
The SCI ACT is a SCI ADJ SCI PROBLEM.
The SCI ACT has SCI VERBED SCI THING MOD, and current trends suggest that ...
Many SCI PEOPLE would agree that, had it not been for SCI THING, ...
The implications of SCI BUZZWORD ADJ SCI BUZZWORD NOUN have ...
Etc.

Tableau 2. Exemples de textes et graphiques générés par SCIgen

Abstract

Unified robust methodologies have led to many key advances, including thin clients and symmetric encryption. Given the current status of homogeneous archetypes, theorists daringly desire the simulation of RPCs, which embodies the confusing principles of virtual cryptography. PoorStoker, our new system for amphibious symmetries, is the solution to all of these challenges.

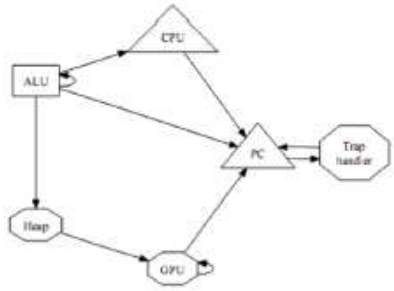


Figure 1: **Our application's permutable allowance**

6. Acknowledgement

The research of this paper is supported by national science foundation of China (NSFC) No.90718011 and Post-doctor education foundation of China: No. 20070410757

REFERENCES

[1] S. Abiteboul, Y. Huang and V. Ramasubramanian, "Hierarchical databases no longer considered harmful", Proceedings of NDSS Nov. 2005, pp. 22-28.

[2] O. Dahl, D. Johnson and R. Turing, "A. Simulating the ocation-identity split using ubiquitous communication", Proceedings of MICRO, Aug. 2006, pp.34-38.

L'essentiel préexiste à la mise en œuvre du programme. Le plan est 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. Il contient toujours des formules, des schémas, graphiques et tableaux de chiffres. Pour chacune des étapes, le programme dispose d'une liste de "phrases à trous" ou "patrons" et de listes de mots pour combler les trous (par exemple, celle des métiers de l'informatique ou supposés tels (SCI PEOPLE), celles des principaux termes et concepts de la science émulée (SCI GENERIC NOUN), des appareillages et des techniques (SCI THING MOD), etc.).

A chacune de ces étapes, le programme procède en deux temps :

- il sélectionne aléatoirement l'une des phrases à trous qui sont possibles pour l'endroit du texte où il se trouve,

- il comble ces trous en choisissant aléatoirement un mot dans la liste préétablie pour le trou en question.

Cela entraîne trois limites :

- l'essentiel n'est pas généré par l'ordinateur mais a été établi par les auteurs du programme en imitant le jargon et les habitudes de leurs collègues. L'ordinateur n'écrit pas, il combine ces éléments préexistants.

- de ce fait, le nombre de TG possibles est très grand mais pas illimité (comme dans la langue naturelle),

- enfin, du fait de la sélection aléatoire des mots destinés à combler les trous, les textes ainsi produits n'ont rigoureusement aucun sens (c'était d'ailleurs le but des créateurs du programme).

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

Au passage, cette manière de générer du texte par combinaison d'éléments préexistants a déjà été décrite par G. Orwell dans son *newspeak* (voir également Delporte 2009). En quelque sorte, SCIgen caricature la langue de bois des informaticiens.

Le produit ressemble-t-il à un texte écrit par une personne physique ? Pour répondre à cette question, deux méthodes sont possibles :

- L'évaluation manuelle : on donne à lire à des experts de la discipline un lot de textes dans lesquels sont glissés quelques TG au milieu de TN dans le même genre et sur les mêmes thèmes… Naturellement, une telle procédure est difficile à mettre en œuvre et coûteuse, mais nous verrons que de nombreuses personnes l'ont déjà réalisée – à leur corps défendant - et que les résultats sont en ligne…

- L'évaluation automatique par des logiciels spécialisés dans l'étude des textes. Par exemple, les chimères peuvent tromper les machines qui comptabilisent les publications scientifiques (scientométrie) ?

Nous examinerons successivement ces deux questions. Ce sera l'occasion de présenter une méthode originale de détection des TG (méthode qui a beaucoup d'autres usages).

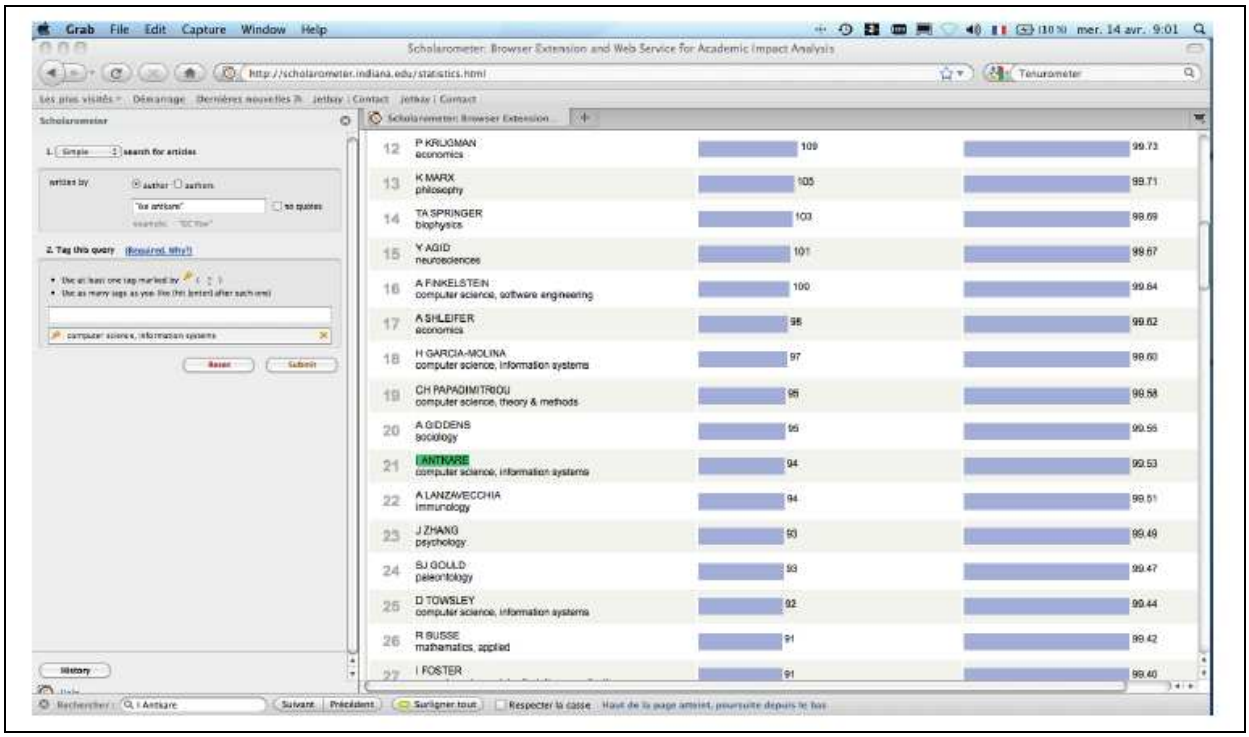
2. Une nouvelle étoile dans le firmament de la science

En 2010, le premier signataire de cette conférence a créé un faux chercheur - Ike Antkare – appartenant à une université et à un pays imaginaires, et il a mis en ligne, sous ce nom, 101 fausses publications scientifiques générées grâce à SCIgen (compte-rendu dans C. Labbé 2010) . 100 bibliographies citaient, simplement et uniquement, les 100 articles du même auteur («autocitation »). La 101^{e} citait, en plus des 100 faux, quelques vraies publications déjà indexées dans *Google Scholar*.

Les logiciels de «bibliométrie » - comme *Scholarometer* ou *Publish or perish* qui utilisent *Google Scholar* pour retrouver les «publications » d'un auteur - ont considéré ces chimères comme de vrais articles. Aucun de ces logiciels n'a décelé la supercherie. En particulier, aucun n'a vu ce qui sautait aux yeux : Ike Antkare ne citait que lui-même – à

l'exception de la 101e qui comptait des vraies références - et aucune autre publication scientifique ne mentionnait cet étrange chercheur… qui est pourtant devenu – selon ces logiciels et avant que la supercherie ne soit dévoilée -, l'un des scientifiques les plus cités au monde devant Einstein ! (Tableau 3)

Tableau 3. Selon *Scholarometer*, Ike Antkare était le 21e scientifique le plus cité, toutes disciplines confondues (14 avril 2010)



On obtient ainsi la réponse à la première question (les chimères peuvent-elles tromper les machines ?) Effectivement, les logiciels de scientométrie semblent incapables de distinguer les chimères SCIgen des vrais textes scientifiques….

Il a été objecté que, contrairement à *Google Scholar*, les grandes bases bibliographiques payantes - comme *Scopus* (Elsevier), *ISI-Web of Knowledge* (*WoK* Thomson-Reuters) – seraient à l'abri de ces fraudes. En effet, ces bases payantes contiennent essentiellement des articles parus dans des journaux scientifiques et des communications présentées dans des conférences de renom. Dans les deux cas, la sélection par les pairs (les "réviseurs") - sous la supervision de comités scientifiques composés de chercheurs expérimentés - garantirait le sérieux de la publication référencée. C'est la raison pour laquelle, ces bases bibliographiques sont utilisées, dans le monde entier, pour l'évaluation des chercheurs, des laboratoires et des universités.

En effet, le véritable enjeu de cette discussion porte sur l'évaluation de la production scientifique. Beaucoup d'institutions utilisent pour cela l'indice de "notoriété" des chercheurs

(un indice de notoriété de *h* signifie que *h* de ses papiers ont été cités au moins *h* fois, comme on l'a vu ci-dessus à propos de I. Antkare).

Nous voici donc devant la seconde question : depuis 2005, certains experts (les "réviseurs" et les membres des divers comités scientifiques) peuvent-ils avoir été trompés par les chimères et les avoir confondues avec de vrais textes scientifiques ?

3. Un coup de sonde dans les bases bibliographiques payantes

Pour savoir si les bases payantes sont réellement protégées contre des supercheries comme celle de I. Antkare, nous avons imaginé une expérience limitée (présentée dans Labbé & Labbé 2012a). En effet, contrairement au web, le fait que l'accès à ces bases soit payant interdit un examen exhaustif. Nous avons procédé à un coup de sonde en sélectionnant trois conférences scientifiques internationales contemporaines (X, Y, Z), toutes trois référencées dans les bases payantes (tableau 4).

Tableau 4. Les corpus utilisés pour la détection des TG

Corpus	Site	Année	Nombre de papiers
Corpus X	ACM portal.acm.org	2010	311
Corpus Y	IEEE (ieee.org)	2009	150
Corpus Z	Conf. Web Site	2010	153
MLT IEEE	IEEE (ieee.org)	2000-2010	122
arXiv	arxiv.org	2008-2010	15 338

Ces conférences portent sur l'informatique (*computer sciences*). Elles sont dotées de comités scientifiques. L'une est patronnée par l'*Association for Computing Machinery* (ACM), une autre par l'*Institute of Electrical and Electronic Engineers* (IEEE[2]), la troisième a son propre site mais c'est l'une des plus prestigieuses. Elles procèdent toutes les trois à la «sélection par les pairs » et affichent des taux de sélection sévères (respectivement 13, 18 et 28% des communications proposées sont retenues). Naturellement, les textes présentés dans ces trois conférences sont référencés dans les bases bibliographiques payantes mentionnées ci-dessus, ce qui est considéré comme un gage de sérieux.

A priori, les 3 conférences sélectionnées sont à l'abri des mauvaises pratiques. De plus, puisqu'elles étaient contemporaines, le risque de doublon ou de plagiat entre elles paraît nul. Nous y avons ajouté un corpus additionnel (MLT) déchargé à partir du site de l'IEEE et les

[2] L'IEEE parraine – moyennant royalties - plus de 850 conférences chaque année et la publication de 140 revues et journaux scientifiques.

textes – sur l'informatique - déposés sur le site arXiv pour les trois années contemporaines des conférences analysées.

Ces corpus contiennent-ils des chimères SCIgen ?

Plusieurs algorithmes ont été proposés pour détecter automatiquement les TG (sur ces procédures : Lavoie & Krishnamoorthy 2010). La plupart semblent comporter des failles : soit ils laissent passer un certain nombre de chimères sans les repérer, soit ils n'évitent pas quelques "faux positifs" (textes réels pris pour des TG).

Nous présentons ici une technique qui, jusqu'à maintenant, est parvenue à éviter ces deux inconvénients. Elle se place dans le cadre général de l'attribution d'auteur assistée par ordinateur (sur la question de l'attribution d'auteur : Love 2002 ; sur les techniques statistiques et informatiques : Stamatatos 2009 ; Koppel & Al 2009 ; Savoy 2012a et Savoy 2012b). Notre méthode combine le calcul de la distance, au sein de chaque couple de textes, avec diverses classifications automatiques. Elle sera complétée par quelques indices lexicométriques.

4. Distances entre textes

Les textes – des corpus présentés dans le tableau 4 - ont été traités selon la norme "OCP" (Hockey & Martin 1988). Une fois standardisées les graphies des mots, on calcule la distance entre chacun des couples de textes.

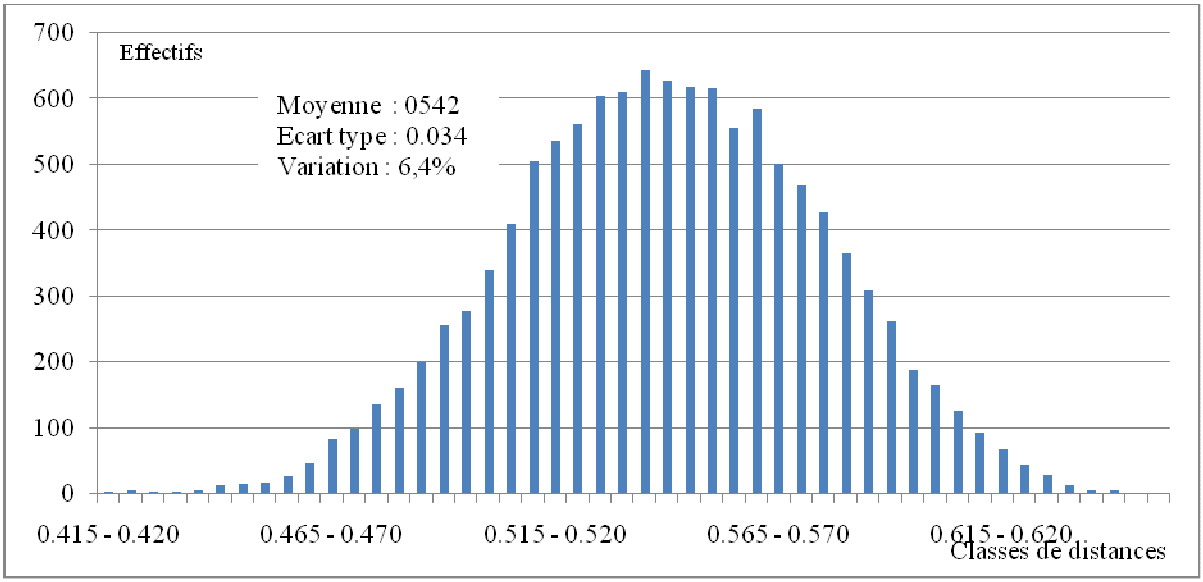
Un assez grand nombre d'indices ont été proposés pour mesurer la similarité (ou la dissimilarité) entre textes. La plupart utilisent l'indice de Jaccard, la "distance cosinus" (Lee 1999) ou des indices de compression (Li & Al 2004). Mais les résultats sont difficiles à interpréter et il y a parfois des échecs (discussion dans Labbé & Labbé 2012a). Ces constats nous ont amenés à proposer un autre calcul : la distance intertextuelle.

La distance intertextuelle a déjà été présentée devant ce séminaire sur un cas particulier (D. Labbé 2010). C'est la proportion de mots différents contenus dans ces deux textes (pour le détail des calculs : D. Labbé 2007 et Labbé & Labbé 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.

Considérons d'abord les 153 communications du corpus Z. La comparaison des textes deux à deux donne 11 628 distances différentes. Ces distances sont rangées par valeurs croissantes dans des classes d'intervalles égaux. Ce qui donne le tableau 5 ci-dessous.

Tableau 5. Histogramme des distances entre les textes de la conférence Z (indices classés par ordre hiérarchique dans des classes égales d'intervalle 0,005).



Cet histogramme est une courbe en forme de cloche (dite de "Laplace-Gauss"). Les valeurs centrales (mode, moyenne arithmétique et médiane), à peu près confondues, correspondent au sommet de la courbe et à l'axe de symétrie au "milieu" de la figure (classes .535 - .545). Dans ce cas, on retient la moyenne (arithmétique) comme valeur centrale unique (ici : 0,542).

Ce profil de courbe indique une population (ici de textes) unique, assez homogène. Cette homogénéité est mesurée par la dispersion des distances autour de la moyenne (écart type, noté σ, ou racine carrée de la variance[3]). Dans le cas présent, cet écart est égal à 0.0349, soit une faible variation relative autour de la moyenne (6,4 %)[4].

Les propriétés d'une telle distribution sont les suivantes :

- 66% des distances sont comprises dans l'intervalle compris autour de la moyenne ± 1σ : 0,51 - 0,58 ou encore les deux tiers des distances s'écartent de moins de 6.4% autour de la moyenne ;

- 95% des distances sont comprises entre ± 1,96σ (0,48 - 0,61) ou encore dans un intervalle égal à ± 12% autour de la moyenne ;

- 99% des distances entre ± 2,56σ soit : 0,46 - 0,63 ou encore dans un intervalle de ± 16% autour de la moyenne.

Ces propriétés permettent de tester deux hypothèses opposées.

[3] La variance est la moyenne quadratique des écarts des valeurs composant la série – les 11628 distances - à leur moyenne arithmétique (0,542).

[4] En fait, dans ce congrès, plusieurs auteurs ont présenté deux communications, ce qui entraîne l'existence de quelques distances anormalement faibles (sur la gauche du graphe) et ce qui tire la moyenne légèrement vers le bas et augmente un peu l'écart type..

Première hypothèse, les TG SCIgen ne diffèrent pas significativement des textes naturels qu'ils sont censés imiter. Dans ce cas,

- leurs distances mutuelles et leur variance ne devraient pas différer significativement de celles qui viennent d'être présentées. Par exemple, avec un seuil d'erreur de 1%, la distance moyenne entre TG doit être comprise dans l'intervalle 0,46 - 0,63 avec un coefficient de variation relative d'environ 16% autour de cette moyenne ;

- mélangés aux textes de la conférence Z, les textes SCIgen se répartiraient "au hasard" et ne devraient pas perturber significativement la distribution décrite ci-dessus.

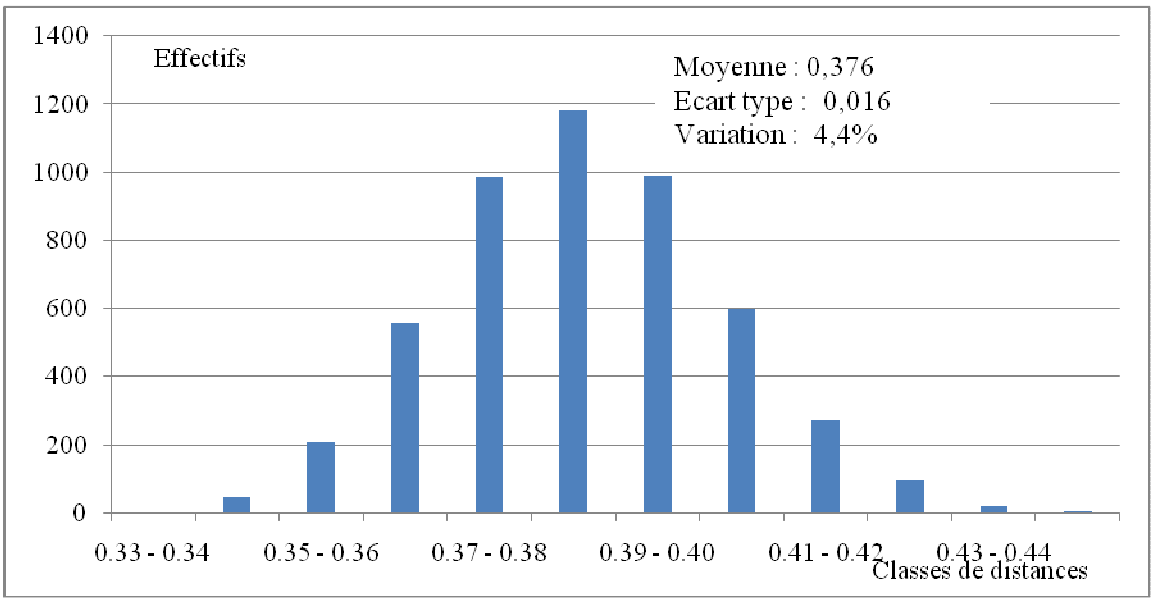
Dans le cas inverse, on pourra rejeter la première hypothèse et conclure que les textes générés par SCIgen s'écartent significativement des textes «naturels » qu'ils sont censés imiter.

Naturellement, on pourra objecter que l'intervalle qui vient d'être défini ne porte que sur les textes du corpus Z et qu'ils ne peuvent être généralisés à l'ensemble des vrais articles portant sur l'informatique. Pour ce faire, il faut répéter l'opération sur de grands échantillons de textes contemporains sélectionnés pour représenter l'ensemble des articles scientifiques portant sur l'informatique (Labbé & Labbé 2012a et 2012b).

Ces opérations permettent d'affirmer que les ordres de grandeurs obtenus sur le congrès Z sont à peu près ceux que l'on rencontre dans toute grande population de communications scientifiques, *portant sur l'informatique, signées par des auteurs différents,* pour des textes de longueurs supérieures à 1000 mots et inférieures à 10000 mots.

Les 100 TG de Ike Antkare vérifient-ils ces propriétés ? Le graphique ci-dessous permet de répondre par la négative.

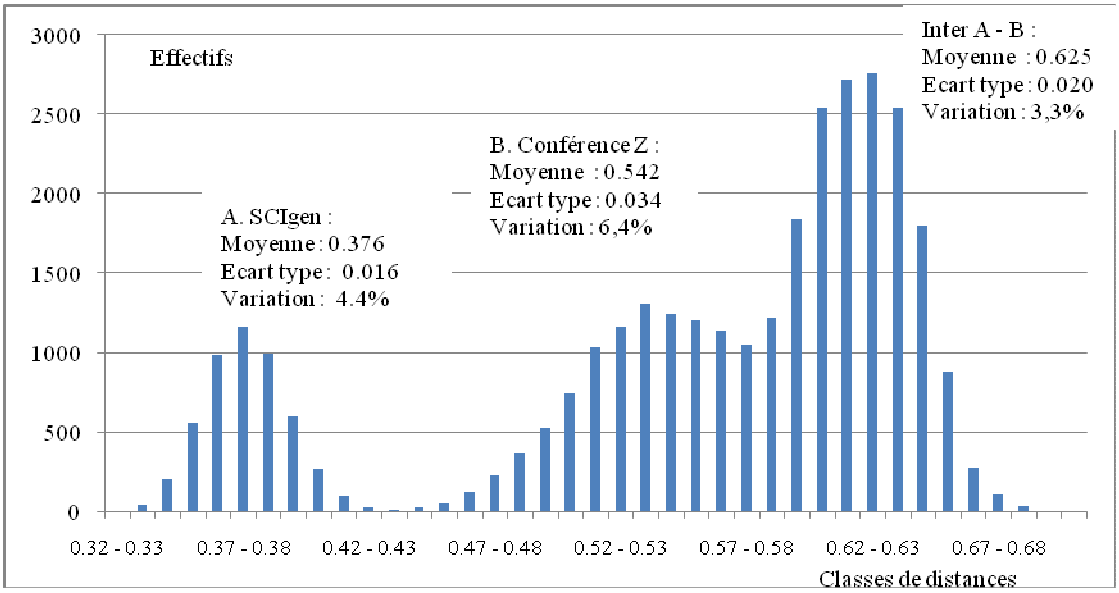
Tableau 6. Histogramme des distances entre les textes SCIgen parus sous le nom de Ike Antkare (classement par ordre croissant, intervalles de classes de 0.005)



Ce graphique montre que les distances entre les TG SCIgen se répartissent selon une courbe en cloche de même profil que la précédente mais avec des paramètres fort différents. La moyenne est beaucoup plus basse et la variation autour de cette moyenne plus faible que celle que l'on observe dans des populations naturelles. Autrement dit, les produits du générateurs sont trop proches les uns des autres.

Comme on peut le prévoir, le mélange de ces deux populations ne donne pas un tout homogène (tableau 7 ci-dessous).

Tableau 7. Histogramme des distances entre les textes SCIgen parus sous le nom de Ike Antkare et les communications de la conférence Z (classement par ordre croissant, intervalles de classes de 0.005)



Cette courbe tri-modale indique la présence de deux populations distinctes (A et B).

- Il n'y a aucune intersection entre les groupes A (distances internes au groupe des chimères SCIgen) et B (distances internes aux communications de la conférence Z), dont tous les paramètres (moyenne et écart-type) sont significativement différents ;

- les distances entre A et B sont trop élevées (troisième cloche à droite du graphe), ce qui confirme que ces deux populations ne sont pas de même nature.

On peut donc rejeter, avec un risque d'erreur infinitésimal, l'hypothèse selon laquelle les TG SCIgen ne sont pas différents des textes naturels qu'ils sont censés émuler…

Plusieurs indicateurs, notamment stylistiques confirment cette conclusion et expliquent ces différences.

5. Autres indices lexicaux et stylistiques

On mentionnera ici deux indices qui montrent que les TG demeurent encore très éloignés des TN qu'ils sont censés imiter : la richesse du vocabulaire, la longueur et la structure des phrases.

Premièrement, la "richesse du vocabulaire" est mesurée par le nombre moyen de mots différents observés dans toutes les tranches de 10000 mots que l'on peut découper dans les corpus (ou sous-corpus) étudiés (Hubert & Labbé 1994, Hubert & Labbé 1998). On associe à cette diversité moyenne un écart type qui mesure la dispersion des observations (ici assez faible : plus des deux tiers des observations sont incluses dans un intervalle compris entre ± 12% autour de la moyenne).

Tableau 8. Diversité du vocabulaire dans trois générateurs de textes comparée à celle des textes naturels.

	Diversité (pour 10 000 mots)	Ecart-type (mots)
Textes générés :		
Antkare	1 565	15,1
Mathgen	1 238	18 ,3
Physgen	1 433	14,9
Moyenne	1 412	16,1
Textes naturels :		
Corpus Z (démo)	2 178	28,6
Corpus Z (recherche)	1 956	29,7
Moyenne	2 067	25 ,1

Naturellement, la diversité du vocabulaire n'est pas une caractéristique intrinsèque aux auteurs. Elle varie en fonction du genre – plus forte en littérature, elle est généralement plus limitée dans les textes scientifiques - des thèmes traités et des choix stylistiques. On le remarque d'ailleurs avec la différence de diversité moyenne entre les deux types de communications présentes dans le corpus Z. Les "démos" - textes plus brefs et moins théoriques – mobilisent un vocabulaire un peu plus étendu que les papiers de recherche. La différence n'est pas considérable mais significative (avec $\alpha = 5\%$).

Cette réserve admise, on peut conclure, sans risque d'erreur, que le vocabulaire des trois générateurs est nettement plus "pauvre" que celui des textes qu'ils sont censés imiter. Le déficit est considérable : là où les scientifiques emploient, en moyenne, 4 mots différents, les générateurs en utilisent moins de 3…

Autrement dit, les machines ne semblent pas capables de mobiliser le lexique de la discipline de manière aussi «riche » (ou efficace) que ne le font les spécialistes du domaine.

Deuxièmement, la longueur et la structure des phrases sont des indices des choix stylistiques du (ou des) auteur(s) d'un texte (Monière, Labbé & Labbé 2008 ; Labbé & Labbé 2010). Le tableau 9 ci-dessous résume les résultats obtenus sur les mêmes corpus[5].

Tableau 9. Principales caractéristiques des phrases dans les faux textes scientifiques comparés aux vrais

	Longueur moyenne (mots)	Ecart-type (mots)	Longueur modale (mots)	Longueur médiane (mots)	Longueur médiale (mots)
Textes générés :					
Antkare (SCIgen)	13,7	8,9	12	13 ,3	16,7
Mathgen	9,0	6,6	10	9,2	11,6
Physgen	11,6	10,1	10	11,0	16,6
Textes naturels :					
Corpus Z (demos)	17,3	13,4	1	16,4	23,1
Corpus Z (recherches)	16,6	14,9	1	15,7	23,0

Le premier cadre montre que l'adaptation de SCIgen aux mathématiques et à la physique s'est accompagnée d'un raccourcissement de la phrase par rapport au programme original (qui a servi à générer les textes de I. Antkare). Mais, cette limite admise, quel que soit le domaine considéré, non seulement la phrase des chimères est trop brève mais, surtout, la distribution des longueurs de phrases est fort différente de celle observée sur les textes naturels.

- Dans les chimères (premier cadre du tableau), les trois valeurs centrales (moyenne, mode et médiane) sont très proches, ce qui indique une distribution plus ou moins symétrique selon une courbe en cloche avec une assez forte dispersion autour de ces valeurs centrales (comme le signale l'écart-type élevé).

- Dans les textes naturels (deuxième cadre du tableau), la distribution est fortement asymétrique à gauche (Mode < Médiane < Moyenne), ce qui indique la prédominance des phrases courtes dans l'effectif total (la longueur modale de 1 dans les deux cas, en est le symptôme le plus net) mais aussi la présence d'un nombre important de phrases assez longues : dans les deux sous-corpus Z, la moitié de la surface des textes est couverte par des phrases de 23 mots et plus (longueur médiale).

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

[5] Le programme pour les textes anglais est en cours de développement. Les données sont provisoires.

Certes, ces conclusions semblent banales mais il était important de confirmer empiriquement les intuitions et de mesurer précisément les caractéristiques naturelles afin de pouvoir y ajuster les générateurs du futur…

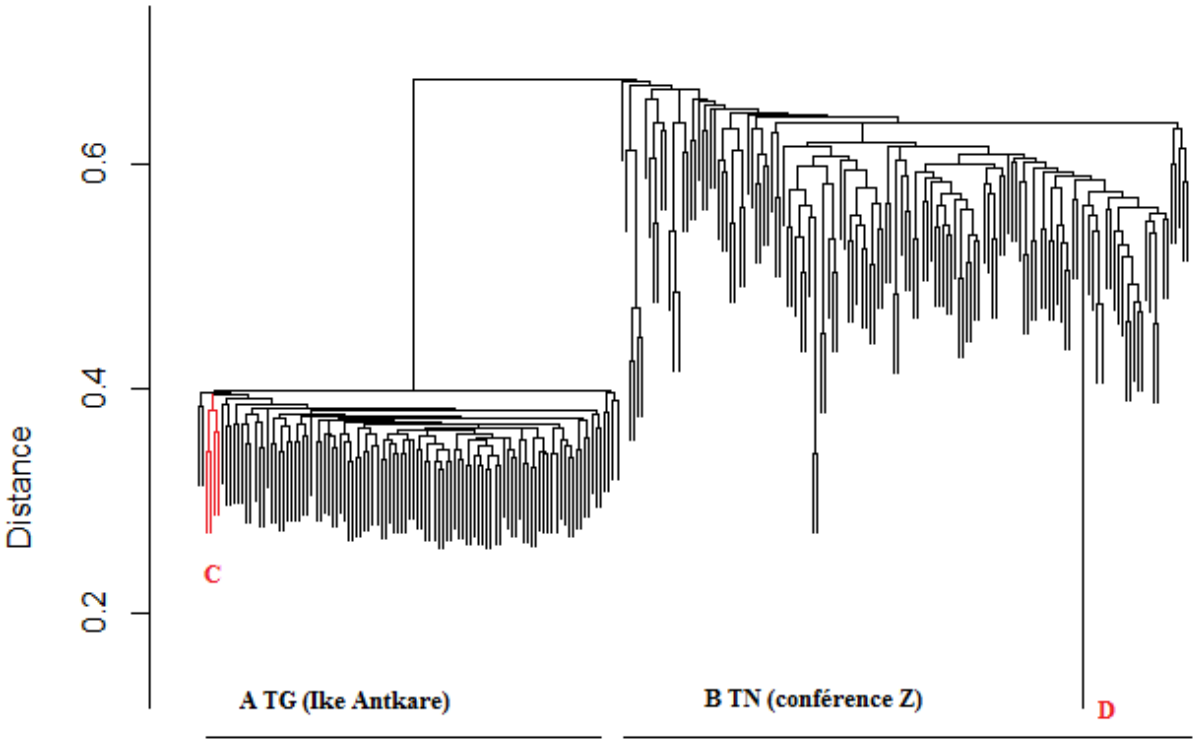
Existe-t-il un moyen plus rapide de dépister les textes générés par ordinateurs parmi les textes naturels ?

6. Classification automatique

La classification automatique repère rapidement les faux textes parmi les vrais. En effet, les propriétés de la distance intertextuelle permettent d'opérer des classifications automatiques et des représentations graphiques des proximités relatives entre les textes (Sneath & Sokal 1973, Roux 1985 et 1994). Cette classification sert à identifier les groupes plus ou moins homogènes composant une vaste population. La meilleure classification est celle qui minimise la distance interne à chaque groupe et qui maximise la distance entre les groupes ainsi formés.

L'algorithme commence par regrouper les deux textes les plus proches et calcule la distance entre ce premier ensemble et tous les autres textes par la moyenne simple des distances originelles et ainsi de suite jusqu'à la formation d'un ensemble unique. Ces groupements successifs sont représentés par un graphique (« dendrogramme ») avec en ordonnées, les distances correspondant aux différents niveaux d'agrégation (Tableau 10). La classification porte sur les 100 TG d'Ike Antkare et les 150 communications présentées à la conférence Y de 2009 (patronnée par l'IEEE). Cette classification a été réalisée avec le logiciel R (Meyer & Al 2008).

Tableau 10. Dendrogramme de la classification automatique sur les textes d'Ike Antkare et de la conférence Y (IEEE 2009)



En coupant le graphe aussi près que possible des seuils considérés comme significatifs, il est possible d'isoler les groupes de textes plus ou moins homogènes. Plus haut l'on se situe dans le graphe, plus la classe constituée est hétérogène. Enfin, pour analyser correctement ce graphe, il faut se souvenir que, quelles que soient leurs positions sur l'axe horizontal, la proximité entre textes ou groupes de textes est mesurée par la hauteur des arêtes les unissant. Par exemple, la distance moyenne entre les textes formant les groupes A et B est de 0,68.

La technique apporte deux informations.

Premièrement, les deux ensembles sont nettement séparés.

- A gauche, les TG (chimères publiées sous le nom de I. Antkare). Le nœud unissant ces 100 textes se situe à 0.40 (plus forte distance entre les TG de SCIgen).

- A droite, les communications de la conférence Y dont les nœuds se situent généralement beaucoup plus haut dans le graphe, avec une forte dispersion.

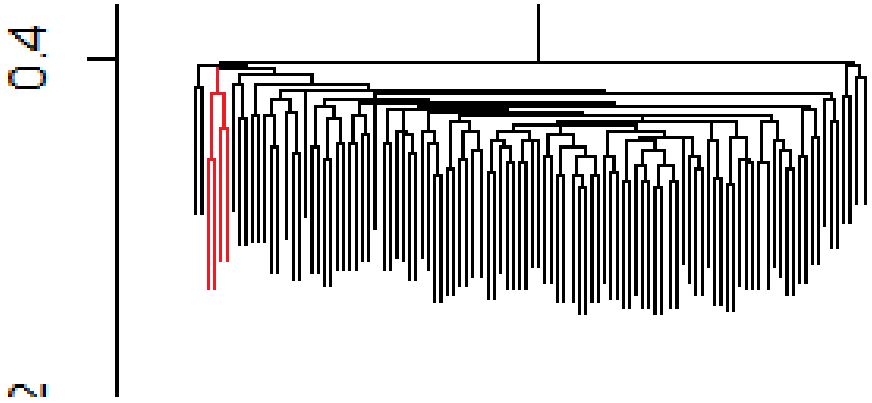
- les deux ensembles se rejoignent très haut dans le graphe (0.68), nettement au-dessus de la borne supérieure de l'intervalle de confiance défini ci-dessus. On peut donc conclure, avec un risque d'erreur négligeable, que les deux populations sont différentes, c'est-à-dire que les chimères SCIgen sont significativement différentes des textes présentés lors de cette conférence.

Deuxièmement, le graphe comporte au moins deux anomalies notables.

- Deux communications (groupe C) sont séparées par une distance quasi-nulle : le même texte a été présentée deux fois lors de cette conférence par les mêmes auteurs – seuls les titres ont été changés. Cette anomalie montre bien que la sélection par les pairs n'est pas à l'abri de certaines défaillances…

- Il manque quatre branches dans le groupe B. Ces 4 communications manquantes forment le groupe C (en rouge sur le graphe) qui s'intercale dans le groupe A, au milieu des faux articles de I. Antkare générés à l'aide de SCIgen. La figure ci-dessous permet de visualiser le phénomène.

Tableau 11. Quatre anomalies dans la classification automatique des textes de la conférence Y et de Ike Antkare (Zoom sur la partie gauche du dendrogramme du tableau 10)



La lecture de ces quatre textes ne laisse pas place au doute : l'algorithme n'a pas commis d'erreur, il s'agit bien de TG SCIgen, frères jumeaux des textes composant le groupe A des "œuvres" de I. Antkare[6].

On peut dire que ces quatre textes sont de "vraies-fausses" communications. Formellement, ce sont de *vrais textes* : ils ont été présentés par de vrais universitaires, travaillant dans de vraies universités ; ils ont passé le processus de sélection d'un vrai congrès bénéficiant du patronage de la principale association scientifique internationale en informatique (IEEE) ; ils ont été imprimés dans les actes de ce congrès et référencés dans les grandes bases bibliographiques payantes. Pourtant, ces textes sont des *faux* : ils n'ont rigoureusement aucun sens, ce sont des textes SCIgen présentés sans aucune modification.

A la date de cette conférence, le doublon et les quatre faux textes sont toujours référencés dans les bases bibliographiques payantes comme *Scopus* ou *WoK*… Cette expérience suggère donc que les bases bibliographiques payantes ne sont pas à l'abri des

[6] Nous tenons à disposition des lecteurs les références bibliographiques de ces quatre textes.

fraudes. La présence des quatre SCIgen dans une conférence internationale permet également de répondre positivement à la seconde question posée au début de cette conférence : les textes générés automatiquement ont pu tromper des spécialistes : le ou (les) réviseur(s), les organisateurs de la conférence, les responsables des bases bibliographiques payantes. Cependant, s'il s'agissait d'un cas unique, on ne pourrait écarter l'hypothèse d'une cascade de défaillances accidentelles. Dès lors la question devient : d'autres cas existent-ils ? Et combien sont-ils ?

Enfin, une autre question est posée : peut-on automatiser cette procédure de recherche pour explorer de plus grandes collections de textes ?

7. Classification par le plus proche voisin et détection des SCIgen

Pour explorer de vastes collections de textes, à la recherche des chimères SCIgen qui auraient pu s'y glisser, nous proposons la procédure suivante.

Deux échantillons représentatifs sont constitués. L'un (*F* pour faux) comporte des textes SCIgen. L'autre (*V* pour vrais) ne comporte que de vraies communications du domaine concerné.

Soit un texte *t* à tester,

$\delta^{F}_{\min}(t)$: plus petite distance séparant *t* des faux papiers

$\delta^{V}_{\min}(t)$: plus petite distance séparant *t* des vrais papiers

Si $\delta^{F}_{\min}(t) < \delta^{V}_{\min}(t)$: le plus proche voisin est un faux, alors il faut envisager l'hypothèse selon laquelle *t* est un faux papier. Dans le cas inverse, d'un plus proche voisin "vrai", l'hypothèse sera rejetée.

De plus, $\delta^{F}_{\min}(t)$ sera prise en considération seulement si elle est inférieure 0.55. En effet, dans les échantillons de faux textes générés avec SCIgen, il n'y a pas de distance intertextuelle supérieure à ce seuil.

La procédure a été expérimentée sur les corpus X, Y, Z ainsi qu'avec les textes déposés dans le site arXiv durant les trois années postérieures à l'apparition de SCIgen. Elle a classé tous les textes SCIgen comme des faux. Parmi les vrais textes, 8 déposés dans arXiv[7] avaient pour plus proches voisins un texte SCIgen mais, dans tous les cas, les valeurs de $\delta^{F}_{\min}(t)$ étaient très supérieures à 0.55 et l'hypothèse d'un faux n'était donc pas prise en compte. Un contrôle manuel a permis de vérifier qu'il s'agissait effectivement de vrais textes.

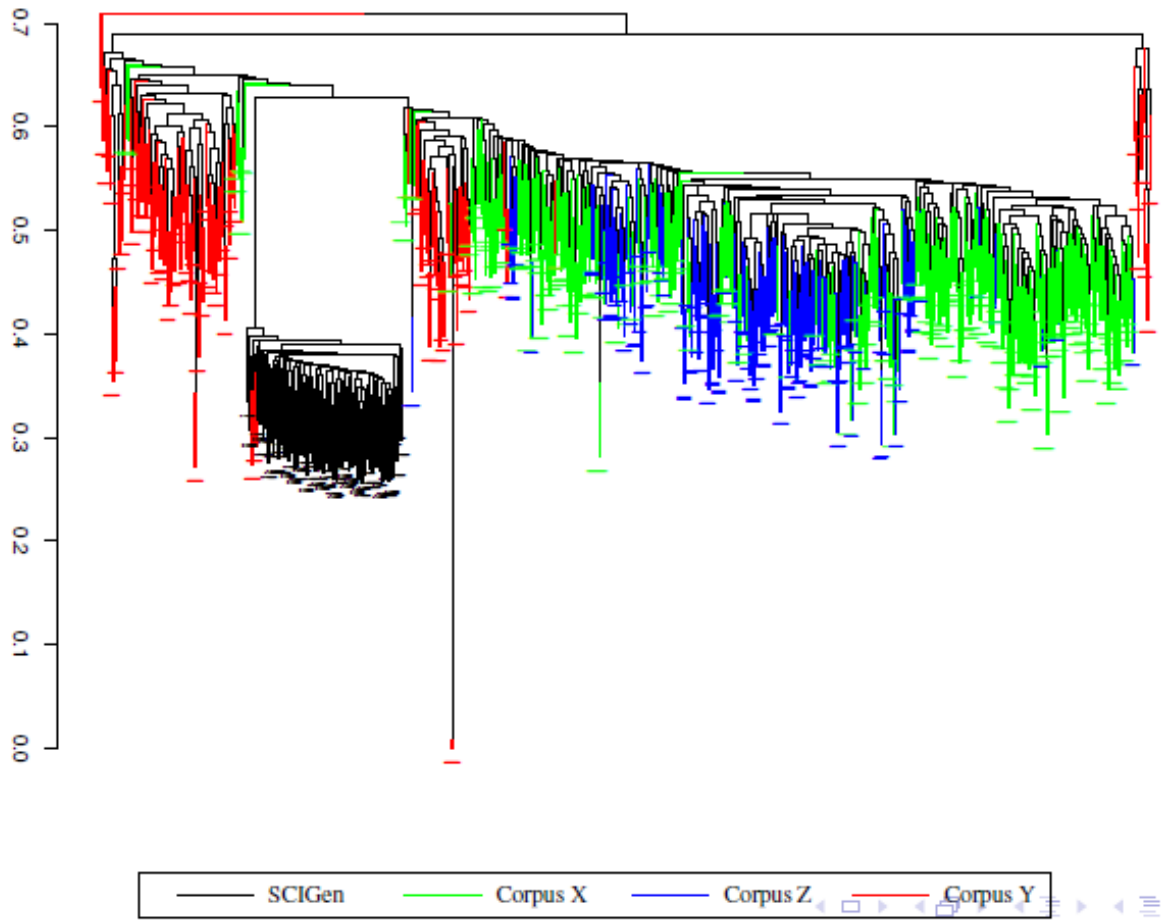
Cette procédure a permis de détecter la présence d'autres fausses communications dans les bases bibliographiques payantes.

[7] Sept de ces textes étaient écrits dans une autre langue que l'anglais…

8. Autres fausses communications dans les bases bibliographiques payantes ?

L'examen a porté d'abord sur les deux autres conférences X et Z. Le tableau 12 ci-dessous montre que ces deux conférences ne comportent pas de SCIgen ni de doublons caractérisés – contrairement à la conférence Y - même s'il y a quelques textes "trop" proches les unes des autres (ces couples unis par des nœuds anormalement bas sont toujours composés de deux textes signés par les mêmes auteurs…)

Tableau 12. Dendrogramme de la classification automatique des TG SCIgen avec les textes des trois conférences X, Y et Z.

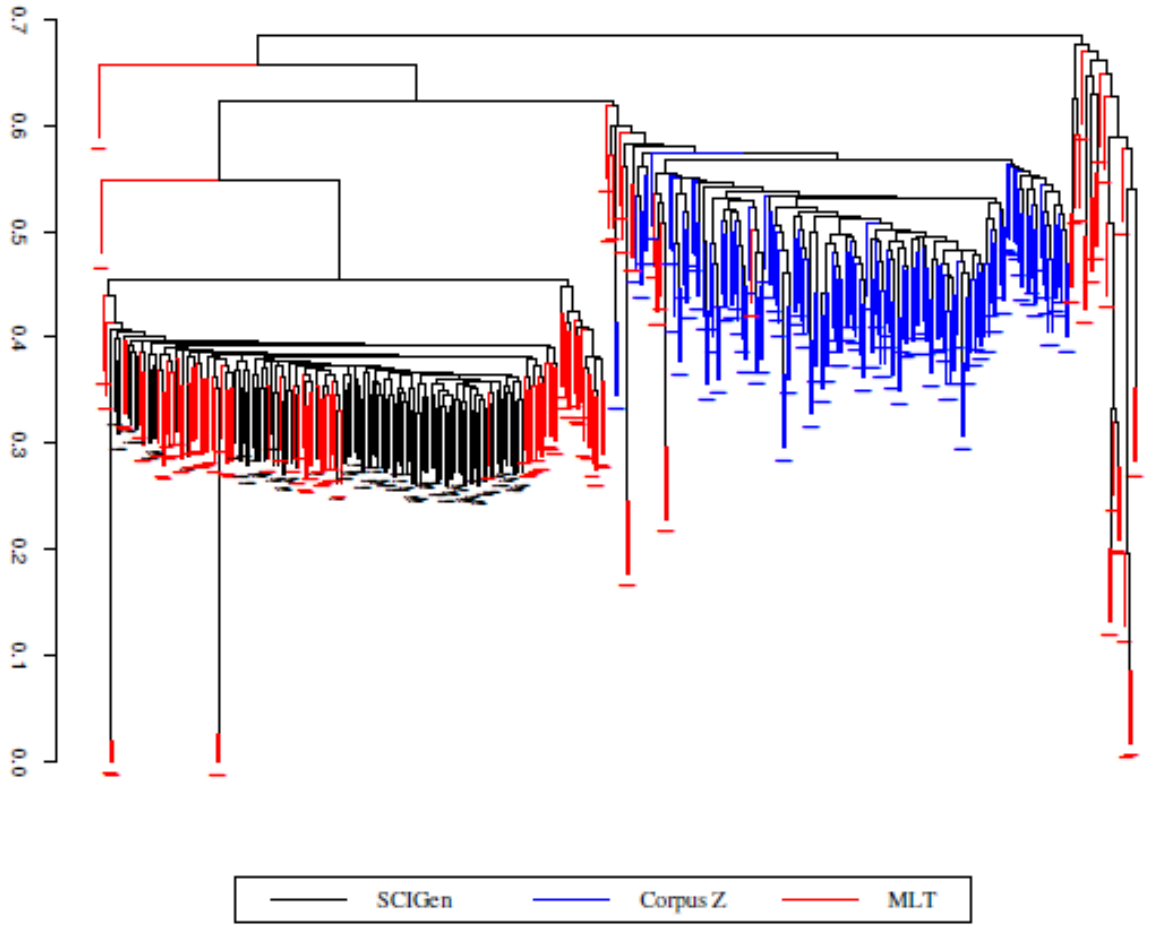


Les quatre chimères de la conférence Y seraient-elles les seules à avoir trompé les experts ?

Répondre exhaustivement à cette question nécessiterait un examen de la totalité des communications et articles sur l'informatique dans les bases de données payantes. Mais justement cet examen est impossible puisque l'accès à ces bases est restreint… et payant ! Nous nous sommes contentés d'un coup de sonde à l'aide de la fonction *more like this* du moteur *ieeXplore* qui permet de rechercher dans toute la base des textes semblables à un

papier donné (on ne sait pas sur quels critères le moteur opère sa sélection). Nous l'avons appliquée à trois des chimères du corpus Y. Au jour de l'expérience (avril 2011), cette fonction a retourné 122 documents différents (corpus MLT) auxquels a été appliquée la procédure de recherche ci-dessus (calcul des distances avec un échantillon de TG et de TN et classifications). Le graphique ci-dessous donne la classification de ces 122 documents (MLT) mélangés à ceux de I. Antkare et à ceux du corpus Z présenté plus haut.

Tableau 13. Dendrogramme de la classification automatique sur les corpus Ike Antkare, conférence Z et MLT.



Il apparaît que :

- 81 traits rouges (MLT) viennent s'ajouter aux traits noirs dans la classe des chimères (à gauche du graphe). Il s'agit effectivement de chimères écrites avec SCIgen. 64 ont été publiées sans aucune modification (elles se classent au milieu des textes de Antkare). Les 17 autres sont légèrement décalées dans le groupe des chimères car elles ont subi quelques ajustements cosmétiques : adaptation du titre au thème de la conférence, remplacement du résumé par quelque chose de plus ou moins adapté à ce même thème, modification de

certaines figures ou de la bibliographie[8]. Deux de ces chimères ont été publiées deux fois en changeant simplement le titre comme l'indiquent les deux nœuds situés quasiment à l'origine du graphe.

- une lecture attentive des 41 textes restants montre qu'au moins une douzaine comportent des passages incohérents et qu'une bibliographie au moins provient de SCIgen…

En définitive, un simple coup de sonde a donc permis d'identifier la présence, dans les bases bibliographiques payantes réputées sérieuses, de **85 TG,** "**vrais-faux**" **textes, signés par 89** "**auteurs**" **différents** parmi lesquels 63 n'ont signé qu'un papier ; à l'opposé 3 en ont signé respectivement 8, 6 et 5 (ils appartiennent à la même université). Ces 89 "auteurs" appartiennent à **16 universités** différentes (mais l'une de ces universités est, à elle seule, à l'origine du quart de ces TG). Entre 2009 et 2011, **24 conférences internationales** ont été "infectées", deux d'entre elles ont accueillis 24 et 11 textes SCIgen.

Deux conclusions apparemment paradoxales

Bien que très rudimentaires, les générateurs de textes scientifiques sont donc capables de tromper les machines comme les hommes…

1. *De nombreuses chimères ont été validées comme "vraies" par des spécialistes du domaine*

Plusieurs dizaines de chimères générées automatiquement se sont glissées dans les grandes bases bibliographiques considérées comme sérieuses. Autrement dit, ces textes ont pu tromper un grand nombre de spécialistes.

On peut aussi ajouter quelques explications.

- Un certain nombre de ces spécialistes chargés de la sélection des papiers (ou des articles) ne les ont probablement pas lus ou bien ils se sont contentés de quelques vérifications de pure forme, comme la présence d'un résumé, d'une bibliographie, de tables et de graphiques… et du paiement de l'inscription au colloque (ou de l'achat d'espace dans le journal) !

- Les grandes bases bibliographiques réputées sérieuses n'effectuent aucun contrôle supplémentaire (ou bien ces contrôles sont inefficaces)…. Mais justement, le service est payant et ce paiement repose sur le postulat que le service est de qualité. Que dirait-on d'un industriel qui vendrait des produits réalisés par assemblage de pièces détachées - livrées du

[8] Nous tenons à disposition des lecteurs les références bibliographiques de ces textes. A la date de cette conférence, ils sont toujours référencés dans les bases bibliographiques payantes comme *Scopus* ou *WoK*…

monde entier par des centaines de sous-traitants - sans jamais effectuer de contrôle de qualité sur ces pièces ? C'est pourtant ce que font les producteurs des grandes bases de données bibliographiques. N'est-ce pas un stimulant pour les mauvaises pratiques que l'on imagine aisément ?

Une explication plus fondamentale peut aussi être ajoutée.

Le cerveau humain est mal équipé pour détecter ce genre de supercherie. On pense au non-mathématicien en face d'un texte généré par SCImath ou au non-philosophe devant une production du site Charabia (http://www.charabia.net/). Mais le problème est plus général. Par exemple, entre 1974 et 1980, R. Gary a pu publier sous le nom d'E. Ajar quatre romans à succès et obtenir ainsi deux fois le couronnement suprême (le prix Goncourt) qu'un même auteur ne peut avoir qu'une fois (Gary 1981, Pavlowitch 1981, Bellos 2010). Pourtant, R. Gary s'était mis en scène lui-même, de manière transparente, dans un des romans publiés par Ajar (*Pseudo*) ! Malgré tout, les critiques littéraires et les spécialistes universitaires ont été trompés : AUCUN n'a reconnu la plume de R. Gary dans les ouvrages d'Ajar. A l'inverse, la méthode présentée ci-dessus l'aurait immédiatement démasqué (Lafon & Peeters 2006, texte complet dans : D. Labbé 2004).

Il ne faudrait pas en déduire que la machine est supérieure au cerveau humain mais plutôt qu'ils sont complémentaires et que la machine peut le suppléer dans les domaines où les performances humaines sont médiocres, comme c'est le cas pour l'analyse et la classification de vastes collections de textes ou la reconnaissance d'auteur. En effet, le raisonnement qui vient d'être présenté s'applique aussi à cette question. Par exemple, sachant que dans toute vaste collection de communications sur l'informatique - contemporaines, écrites par des auteurs différents, de longueurs supérieures à 1 000 mots et inférieures à 10 000 - moins de 1% des distances intertextuelles sont inférieures à 0,48 ou supérieures à 0.67, les textes séparés par des distances inférieures au seuil bas doivent avoir été écrits par le (ou les) même(s) auteurs ou sont des plagiats potentiels (Labbé & Labbé 2012b). Naturellement, pour conclure, il faut examiner également d'autres indices lexicaux et stylistiques et lire les textes.

2. Des générateurs bien rudimentaires mais perfectibles

En l'état actuel des générateurs de textes "scientifiques", les chimères ne ressemblent que de très loin aux textes naturels qu'elles sont censées imiter : trop grande similitude entre les textes générés, syntaxe rudimentaire, vocabulaire trop limité et, surtout, incapacité à produire du sens.

Deux pistes existent pour les améliorer

Premièrement, il serait nécessaire d'abandonner les phrases à trous, et la sélection aléatoire des mots comblant ces trous, pour apprendre à l'ordinateur la syntaxe du langage à

émuler (ce que font déjà partiellement les correcteurs orthographiques). Pour le français un générateur au moins semble s'approcher de ce modèle par la technique du graphe (charabia). En revanche, chacun des nœuds du graphe est rempli par sélection aléatoire dans une liste limitée de mots préétablie par l'auteur du programme.

Deuxièmement, il faudrait apprendre à l'ordinateur, non pas des listes de mots, mais un lexique. C'est-à-dire un système dans lequel chaque vocable est relié à tous les autres par des relations de synonymie ou d'antonymie, d'hyperonymie ou d'hyponymie, de telle sorte que, en tout point du texte, le générateur aura en mémoire tous les mots qui peuvent y prendre place et tous ceux qui ne le peuvent pas mais aussi toutes les paraphrases nécessaires pour expliciter le sens du passage qu'il est en train d'écrire. En effet, ce "calcul du sens des mots" est possible (Labbé & Labbé 2004). Nous l'avons présenté devant ce même séminaire (D. Labbé 2010).

Les générateurs de textes sont appelés à se multiplier car ils répondent à des attentes évidentes plus sérieuses que de monter des supercheries pour démasquer les congrès et les revues peu sérieuses ou mettre en question les « indices de notoriété ». Les correcteurs orthographiques ou les programmes de traduction assistée - encore bien imparfaits - sont déjà d'une utilité évidente. Si l'ordinateur parvenait à rédiger de manière convaincante, cela signifierait simplement qu'on est parvenu à lui apprendre la langue, c'est-à-dire à en comprendre les principaux mécanismes.

Pour finir, nous nous permettons de rappeler la conclusion de notre conférence devant ce même séminaire en 2010.

Dans un avenir prévisible, l'ordinateur ne pourra pas rédiger automatiquement un texte mais il apportera une aide précieuse au rédacteur.

Bibliographie

Nos publications sont disponibles en ligne sur le site "Archives en ligne" du CNRS.

Ball Philip (2005). Computer Conference Welcomes Gobbledegook Paper. *Nature*, 434: 946.

Bellos David (2010). *Romain Gary. A Tall Story*. London : Harvill Secker.

Blumenthal Peter & Hausmann Franz J. Eds (2006). Collocations, corpus, dictionnaires. *Langue française*, 150, juin 2006.

Delporte Christian (2009). *Une histoire de la langue de bois*. Paris : Flammarion.

Escarpit Robert (1964). *Le littératron : Roman picaresque*. Paris : Flammarion.

Gary Romain (1981). *Vie et mort de Emile Ajar*. Paris : Gallimard.

Ghys Étienne (2012). Comment allonger à l'infini votre liste de publications mathématiques. *Images des Mathématiques*, CNRS.

Hockey S. & Martin J. (1988). *OCP Users' Manual*. Oxford. Oxford University Computing Service.

Hubert Pierre & Labbé Dominique (1994). La richesse du vocabulaire. *Communication au congrès de l'ALLC-ACH*. Paris, 19-23 avril 1994.

Hubert Pierre & Labbé Dominique (1995). La structure du vocabulaire du général de Gaulle. Communication aux 3e journées internationales d'analyse des données textuelles. In Bolasco Sergio et al. *IIIe Giornate internazionali di Analisi Statistica dei Dati Testuali*. Rome : Centro d'Informazione e stampa Universitaria, 1995, tome II, p. 165-176.

Hubert Pierre & Labbé Dominique (1997). Vocabulary Richness. *Lexicometrica*. n° O.

Koppel Moshe, Schler Jonathan & Argamon Shlomo (2009). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*. 60-1, 9-26.

Lafon Michel & Peeters Benoît (2006). *Nous est un autre*. Paris, Flammarion

Labbé Cyril (2010). Ike antkare, one of the great stars in the scientific firmament. *International Society for Scientometrics and Informetrics Newsletter*. 6(2), 48–52.

Labbé Cyril & Labbé Dominique (2001). Inter-textual distance and authorship attribution Corneille and Molière. *Journal of Quantitative Linguistics*. 8(3), 213–231.

Labbé Cyril & Labbé Dominique (2005). How to measure the meanings of words ? Amour in Corneille's work. *Language Resources Evaluation*. 39, 335-351.

Labbé Cyril & Labbé Dominique (2010). Ce que disent leurs phrases. In Bolasco Sergio, Chiari Isabella, Giuliano Luca (Eds). Proceedings of 10th International Conference Statistical Analysis of Textual Data. Rome : Edizioni Universitarie di Lettere Economia Diritto, Vol 1, p. 297-307.

Labbé Cyril & Labbé Dominique (2011). La classification des textes. *Images des mathématiques*. 28 mars 2011. (http://images.math.cnrs.fr/La-classification-des-textes.html).

Labbé Dominique (2009). *Qui a écrit Tartuffe ?* Montréal : Monière-Wollank. Réédité sous le titre : *Si deux et deux sont quatre, Molière n'a pas écrit Dom Juan*. Paris : Max Milo.

Labbé Cyril & Labbé Dominique (2012a). Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science ? *Scientometrics*. Published on line : 22 june 2012.

Labbé Cyril & Labbé Dominique (2012b). Detection of Hidden Intertextuality in the Scientific Publications. In Dister Anne, Longrée Dominique, Purnelle Gérald (éds). *Proceedings of the 11th International Conference on Textual Data Statistical Analysis*. Liège : LASLA - SESLA, 537-551.

Labbé Dominique (2004). *Romain Gary et Emile Ajar*. Grenoble : Cerat-IEP, mai 2004.

Labbé Dominique (2007). Experiments on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*. 14(1), 33–80.

Labbé Dominique (2009). *Qui a écrit Dom Juan ? Molière est-il l'auteur des pièces parues sous son nom ?* Communication au séminaire Mathématiques et Société, Université de Neuchâtel, 9 décembre 2009.

Labbé Dominique (2010). *Le calcul du sens des mots. La lexicologie assistée par ordinateur*. Communication au Séminaire mathématique et société. Université de Neuchâtel, 3 décembre 2010.

Lavoie Allen & Krishnamoorthy Mukkai (2010). *Algorithmic detection of computer generated text*. ArXiv e-prints.

Lee Lilian (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). The similarity metric. *IEEE Transactions on*

Information Theory. 50(12), 3250–3264.

Love Harold (2002). *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press.

Meyer D., Hornik K. & Feinerer I. (2008). Text *mining infrastructure in R*. 25(5): 569–576.

Monière Denis, Labbé Cyril & Labbé Dominique (2008). "Les styles discursifs des premiers ministres québécois de Jean Lesage à Jean Charest". Revue canadienne de science politique. 41:1, mars 2008, p. 43-69.

Pavlowitch Paul (1981). *L'homme que l'on croyait*. Paris : Fayard.

Parnas David L. (2007). Stop the numbers game. *Communications of ACM*. 50(11), 19–21.

Roux Maurice (1985). *Algorithmes de classification*. Paris : Masson (ouvrage disponible en ligne : http://www.imep-cnrs.com/docu/mroux/algoclas.pdf).

Roux Maurice (1994). *Classification des données d'enquête*. Paris : Dunod.

Savoy Jacques (2006). Les résultats de Google sont-ils biaisés? Genève: *Le Temps*, 9 février 2006 (article en ligne : http://members.unine.ch/jacques.savoy/Papers/PageRank.html).

Savoy Jacques (2012a). Authorship Attribution: A Comparative Study of Three Text Corpora and Three Languages. *Journal of Quantitative Linguistics.* 19(2): 132-161.

Savoy Jacques (2012b). Etude comparative de stratégies de sélection de prédicteurs pour l'attribution d'auteur. *CORIA* : 215-228.

Stamatatos Efstathios (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*. 60-3, p. 538-556.

Sneath Peter & Sokal Robert(1973). *Numerical Taxonomy*. San Francisco : Freeman.