Taxonomie de textes peu-naturels

Thomas Lavergne

ENST Paris – France

Abstract

In this paper, we define what is a natural text in a pragmatic way. Then, we present various types of unnatural texts and more particularly the simplest generators, which are also the most widespread in spamdexing. Finally, we describe some statistical tests which allow a first filtering of unnatural texts.

Résumé

Dans cet article nous définissons de manière pragmatique ce qu'est un texte naturel. Puis nous présentons différentes catégories de textes non-naturels et plus particulièrement les méthodes de génération les plus simples qui sont aussi les plus répandues dans le cadre du spamdexing. Enfin nous proposons quelques tests statistiques permettant un premier filtrage des textes non-naturels.

Mots-clés : TALN, naturalité, lexicométrie, génération de texte.

1. Introduction

Est-il possible, et, si oui, dans quelles limites, de distinguer de manière fiable un texte écrit par un humain d'un texte écrit par un ordinateur ?

Cette question intéresse beaucoup les acteurs du web confrontés aux diverses formes de faux contenus (spamdexing, spam, splogs... (Gyöngyi et Garcia-Molina, 2005)). Elle est aussi intéressante pour évaluer la qualité des outils de génération de texte.

Dans cet article, nous tenterons de définir aussi rigoureusement que possible ce que sont les textes naturels et ce qu'ils ne sont pas. Nous présenterons quelques catégories de textes artificiels et quelques tests statistiques simples pour les détecter.

2. Définitions

Au sens le plus strict, un texte est naturel s'il a été écrit par un humain. C'est ce que l'on appelle la « *naturalité de production* ».

Si on se base sur cette définition, la tâche d'identification du caractère naturel d'un texte consiste à déterminer si son producteur est humain. Cette identification s'apparente au classique Test de Turing (Turing, 1950), dans un contexte toutefois où il n'y a pas d'interaction avec le producteur du message. Cette absence d'interaction pose un problème. En effet, un être humain et une machine peuvent produire le même texte de manière indépendante.

Si l'on se base uniquement sur le texte pour statuer, on doit donc se contenter de considérer un texte comme naturel si sa naturalité fait l'objet d'un consensus parmi les humains. C'est la « *naturalité de perception* ».

Cette approximation de la naturalité de production repose sur la capacité des humains à détecter les textes écrits par d'autres humains. Il est toutefois possible que deux humains aient une perception différente d'un texte en fonction de leur degré de maîtrise de la langue employée, de leur connaissance du sujet du texte et de leur compréhension du texte lui-même.

Cette définition rend donc la qualification d'un texte indépendante de son producteur au prix d'une dépendance aux observateurs. Cette dépendance ainsi que la difficulté d'obtenir un consensus crée une zone de flou que l'on cherche à réduire.

Lorsque l'on demande à des humains pourquoi ils trouvent qu'un texte n'est pas naturel on obtient des réponses telles que :

> « Ce texte est incohérent. »
>
> « Les phrases n'ont pas de structure. »
>
> « Ces mots n'existent pas. »
>
> « Il passe du coq-à-l'âne. »

Ce qui nous amène à une définition plus structurelle des textes naturels construite à partir des critères que les humains utilisent pour les détecter :

> **Naturalité structurelle** : Un texte naturel est un texte respectant les contraintes syntaxiques, sémantiques, discursives et pragmatiques de la langue dans laquelle il est écrit.

Cette définition est une approximation de la naturalité de perception. Elle fixe les règles permettant de déterminer si un texte est naturel ou non :

> *contraintes syntaxiques :* le texte doit respecter les règles de la langue qu'il utilise : son vocabulaire, sa grammaire et sa conjugaison. Le texte suivant par exemple, construit à partir d'une sélection aléatoire de mots de cet article, est non-naturel car il ne respecte pas ces contraintes :
>
> > *« texte le connaissance humain information doit et demander la probabilité tolérance elle »*
>
> *contraintes sémantiques :* en plus d'être syntaxiquement correcte, chacune des phrases composant le texte doivent avoir un sens. La phrase suivante, construite sur le principe S + 6 de l'Oulipo, même si elle est syntaxiquement correcte, n'est pas naturelle :
>
> > *« Si deux druides situés dans une plantation font avec une même seconde des animations intérieures de la même cotisation dont la sommité soit plus petite que deux dromadaires, ces deux druides se rencontrent dans cette cotisation. »*
>
> *contraintes discursives :* l'enchaînement des phrases doit lui aussi être cohérent. Un texte changeant de sujet à chaque phrase sans transition n'est pas naturel. Par exemple le texte suivant, qui est un patchwork de phrases de Wikipédia, n'est pas naturel :
>
> > *« Parnes est une commune française, située dans le département de l'Oise et la région Picardie. Elle se présente généralement sous la forme d'un cylindre. »*
>
> *contraintes pragmatiques* **:** le texte, placé dans son contexte, doit avoir un sens. En tenant compte d'informations extérieures au texte, mais que l'on suppose connues en fonction du sujet, le texte doit rester cohérent. Dans un contexte d'article de journal, le

texte suivant, s'il respecte les contraintes précédentes, n'a pas de sens car au moment où il a été écrit aucune piscine n'est disponible sur la lune :

« Un astronaute vient de battre le record du 400m nage libre sur la lune. »

Contrairement aux deux premières, cette définition des textes naturels repose uniquement sur leur structure. Mais ces contraintes sont elles-mêmes difficiles à exploiter. Il y a une certaine variabilité autour de la norme : quelques fautes d'orthographe ou de grammaire ne rendent pas un texte non-naturel, de même que les contraintes sémantiques seront différentes suivant le genre du texte.

3. Modélisation

Un système idéal détectant la naturalité de production est irréalisable. On se restreint donc à la naturalité structurelle qui permet de classer les textes suivant leur statut : naturel, non-naturel ou sujet à controverse.

Dans le cas d'une détection automatique on se trouve donc face à un problème de classification. La naturalité d'un texte étant la probabilité qu'il respecte les contraintes syntaxiques, sémantiques, discursives et pragmatiques. Cette probabilité reposant nécessairement sur une modélisation pertinente de la structure du texte qui doit tenir compte à la fois des connaissances limitées et donc incomplètes du système, mais aussi de la tolérance des humains aux erreurs.

L'évaluation du système peut se faire soit par rapport à la naturalité de production, à l'aide de corpus de textes dont le producteur est connu, soit par rapport à la naturalité de perception à l'aide de corpus étiquetés par des humains.

4. Producteurs de textes non-naturels

4.1. Producteurs humains

La première définition implique qu'un humain ne produira que des textes naturels, mais ce n'est pas le cas de la naturalité structurelle. Il est en effet possible pour un humain d'écrire un texte pour lequel il n'y aura pas consensus, que ce soit volontairement ou non.

Même si l'humain est tolérant aux fautes d'orthographe ou de grammaire, leur accumulation peut rendre un texte incompréhensible. La production de tels textes peut être la conséquence d'un trouble du langage tel que la *dysorthographie,* l'*aphasie,* un simple retard d'apprentissage ou due à des circonstances particulières comme lors de la prise de notes ou l'écriture de SMS, lorsque des contraintes de temps ou d'espace passent avant celles du langage.

Mais la production de textes non-naturels peut aussi être volontaire. Dans un but artistique, le producteur peut imposer des contraintes supplémentaires à la production de son texte : les *lipogrammes* écrits par les membres de l'*Oulipo*, par exemple. Enfin, elle peut être liée à des contraintes sociales, les argots tels que le *l33tspeak* : l'argot des hackers, ou le *loucherbem* : celui des bouchers, qui sont des déformations volontaires du langage.

4.2. Producteurs automatiques

Les outils tels que les abrégeurs et les traducteurs automatiques sont une première source de textes non-naturels générés par ordinateur. Ces outils produisent des textes à destination des humains, ils doivent donc être le plus naturel possible.

La plupart des outils de résumé automatique se contente de sélectionner les phrases les plus pertinentes du texte d'origine (Interjeet, 2001), le respect des contraintes syntaxiques et sémantiques ne leur pose donc pas de problèmes, mais l'enchaînement des phrases manque souvent de cohérence, rendant le texte non-naturel.

Les traducteurs automatiques (Hutchins et Somers, 1992), même s'ils se basent eux aussi sur des textes d'origine naturelle, doivent générer intégralement les textes qu'ils produisent. Cette génération est un processus complexe, même pour un humain, et actuellement les textes obtenus sont rarement naturels.

4.3. Grammaires formelles

Les outils de traduction et les abrégeurs ne peuvent produire des textes, naturels ou non, qu'à partir d'un texte déjà existant. Mais il est aussi possible de produire des textes de manière entièrement automatique. Les grammaires formelles sont un outil fondamental pour effectuer une telle génération. À partir d'un ensemble fini de règles elles permettent de générer une quantité éventuellement infinie de textes.

L'exemple suivant montre une grammaire simple mais qui permet d'engendrer des textes corrects, à la fois d'un point de vue syntaxique et sémantique ; mais cette grammaire ne peut produire que 6 phrases différentes, les contraintes discursives ne seront donc pas respectées.

<texte> → <phrase> <texte> | ε

<phrase> → <sujet> <verbe> « . »

<sujet> → « Le chat » | « Le chien »

<verbe> → « mange » | « boit » | « dort »

La production de textes de taille importante nécessite des grammaires plus riches, mais les grammaires les plus simples, appelées générateurs à trous, ont un intérêt. Elles sont utilisées notamment dans les programmes de questions/réponses ou pour les annonces automatiques dans les aéroports. Par exemple, lorsque l'utilisateur demande la capitale d'un pays, il suffit au générateur d'utiliser un modèle tel que : *« La capitale de <pays> est <capitale> »* en complétant les deux trous. Ces systèmes verbalisent une information structurelle sur la manière d'exprimer la relation entre les différents trous.

Les grammaires plus riches apportent une plus grande qualité dans les textes générés, au prix d'une difficulté de création supérieure. Elle permettent l'écriture de textes complets et variés ayant plus ou moins de sens, tels que ceux produits par le générateur d'articles scientifiques (Stribling et al.) ou le générateur d'essais post-modernes (Pomo) dont est issu l'exemple suivant :

> *If one examines Sontagist camp, one is faced with a choice: either accept neocapitalist discourse or conclude that class, ironically, has significance. In a sense, the premise of the deconstructive paradigm of discourse holds that reality is unattainable. Debord uses the term "cultural materialism" to denote a subdialectic reality.*

4.4. Générateurs purement probabilistes

Lors de la génération de texte à l'aide d'une grammaire, le choix des règles à utiliser peut se faire à l'aide d'une base de connaissance, comme dans l'exemple de générateur à trous ; mais il peut aussi se faire en fonction de probabilités associées à chaque règle. Dans ce cas l'objectif est de produire à moindre coût de grandes quantités de textes et non plus de produire un texte de qualité transmettant une information.

4.4.1. Salades

Les salades sont la forme la plus élémentaire de grammaires. Elles sont constituées d'éléments de textes naturels provenant d'un ensemble de documents, tirés de manière aléatoire et mis bout-à-bout. Différentes granularités sont possibles, de la salade de lettres à la salade de paragraphes. Le tirage des éléments pouvant être uniforme ou pondéré de manière à obtenir la même distribution que dans l'ensemble de documents d'origine. Elles sont à la fois simples à engendrer et permettent la production d'une grande quantité de texte, mais de faible qualité comme le montrent les exemples suivants générés à partir d'articles du journal *Le Monde* :

> ***salade de caractères pondérée :*** *lademOninrrte It u ra rr tsuc uet soucnu la léd' n eeéég nostennrn .ftuu stneecs rbdsetns ie deratnél'uatdqa.rtdil ms imlsu*
>
> ***salade de mots uniforme :*** *veulent oublier. productrice 1985 strictement Rausch, paralyserait affirmait relevaient Censure cadran,personnel atteint mois,appels*
>
> ***salade de phrases :*** *Il croit encore aujourd'hui dur comme fer que les bassistes de jazz ont des bretelles dont ils tirent des doum-doum-doum sonores. Nous userons de tout notre poids pour que se créent (avec la CEE) des liens nouveaux, concrets et mutuellement enrichissants avec la Roumanie.*

4.4.2. Générateurs Markoviens

Les générateurs markoviens sont une forme plus complexe de grammaires qui prennent en compte les dépendances entre les éléments consécutifs du texte. Pour un générateur d'ordre *n*, l'étape d'apprentissage consiste à associer à chaque séquence de *n - 1* éléments du corpus d'apprentissage la distribution de probabilité d'apparition de chaque élément du vocabulaire. Lors de la production de texte, la génération d'un nouvel élément se fait selon la distribution associée aux *n - 1* derniers mots générés. Le générateur utilisant juste une fenêtre de *n* mots, appelée *n-grams* (Manning et Schütze, 1999), dont les *n - 1* premiers constituent le contexte de génération. De même que pour les salades, il est possible de considérer différentes granularités, mais l'utilisation d'éléments de la taille d'une phrase est irréaliste, car elle demanderait des corpus d'apprentissage dans lesquels chacune des phrases apparaîtrait plusieurs fois.

Les textes suivants ont été produits par des générateurs markoviens de mots à l'aide du logiciel SRILM (srilm) et d'un corpus d'articles du journal *Le Monde* :

> ***markovien d'ordre 2 :*** *on s' était tellement apprécié sur ses valeurs symboliques dans la paix sur des populations ont déposé des balkans et ses croûtes. la seule juridiction internationale de littérature. depuis, les elections regionales et russes*
>
> ***markovien d'ordre 4 :*** *le défi du futur que nous aurons contigus les uns aux autres et pas seulement par ce à quoi elle s est mise à incarner l intérêt général ce gouvernement formé soutiendra lors des élections législatives qui suivront des*

Dans ces exemples on peut voir que la qualité du texte produit augmente avec l'ordre du générateur. Le générateur d'ordre 2 capture certains éléments des textes naturels tels que des accords ou des conjugaisons mais le texte reste très peu naturel pour un humain. En effet, le texte n'a pas de réelle cohérence, contrairement au générateur d'ordre 4 qui produit des portions de texte relativement grandes semblant naturelles, tout en donnant l'impression de passer du coq-à-l'âne.

5. Détection automatique

La majorité des systèmes de génération automatique de textes sont facilement identifiables par des humains, mais leur détection de manière automatique est beaucoup plus difficile. Nous présentons ici quelques tests statistiques simples qui permettent un premier filtrage des textes.

5.1. Patchworks

Les textes naturels peuvent être constitués d'un seul texte de grande taille, dans le cas par exemple d'un livre ou d'une page web de grande taille traitant d'un sujet. Mais ils peuvent aussi être composés de plusieurs textes mis bout à bout, dans le cas par exemple d'une page contenant plusieurs articles de presse ou pour un blog où l'on retrouvera sur la même page le message d'origine et les commentaires des visiteurs. Dans ce second cas, la séparation des différents textes n'est pas forcement évidente, ni même souhaitable.

Ces patchworks de textes naturels sont à considérer avec précaution. En effet, ils peuvent être considérés comme des salades de paragraphes. La principale différence étant que dans le premier cas, chacun des textes est complet et peut être considéré de manière isolée comme naturel. Dans le cas d'une salade de paragraphes, les différentes parties de texte mises bout à bout n'ont de sens que prises de manière isolée.

5.2. Les corpus

Différents corpus de textes ont été utilisés pour les expérimentations suivantes. Les deux corpus de textes naturels sont constitués de textes provenant de différentes sources : des articles de journaux (*Le Monde*), des textes littéraires, des débats au parlement européen, des articles de blogs ainsi qu'un échantillon d'articles de Wikipedia, représentant environ 40 millions de mots. Le corpus nommé *naturel 1* est composé de textes de grande taille alors que le corpus nommé *naturel 2* est composé de patchworks de petits textes.

En plus de ces corpus de textes naturels, des corpus de textes artificiels ont étés constitués pour évaluer les tests proposés :

- des corpus de salades de différentes granularités produits par un générateur développé par nos soins,
- des corpus de textes produits par des générateurs markoviens d'ordre 2 et 3 à l'aide du logiciel SRILM (srilm),
- ainsi qu'un corpus de textes générés par des grammaires en utilisant le logiciel Dada Engine (dada).

5.3. Tests statistiques

Une première approche pour détecter les textes artificiels est de vérifier s'ils respectent bien certaines propriétés statistiques qui sont caractéristiques des textes naturels. Ces tests sont

simples et rapides et permettent déjà un premier filtrage avant d'envisager des méthodes plus complexes (voir (Ntoulas et al., 2006) pour d'autres tests similaires mis en œuvre dans une optique de détection du spamdexing).

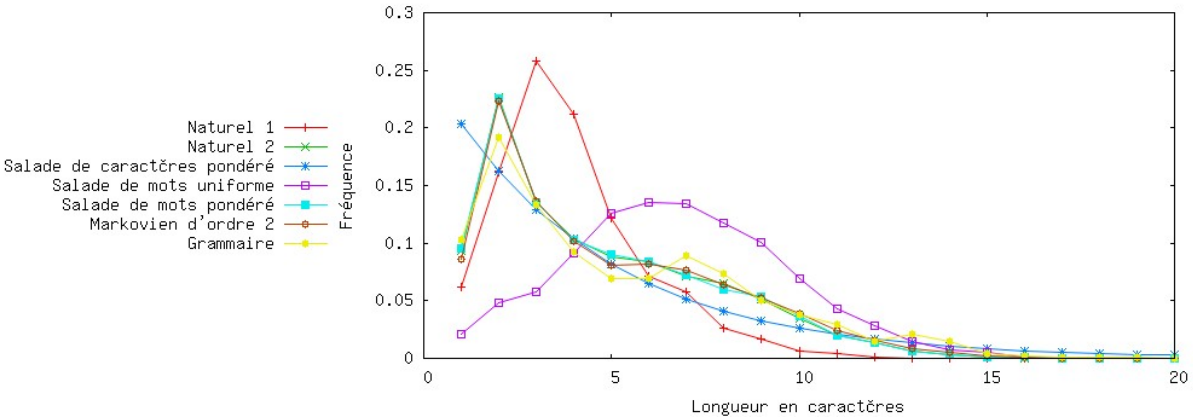

Figure 1 : Distribution de probabilité des mots de différents types de textes en fonction de leur longueur.

5.3.1. Longueur des mots et phrases

La distribution de la longueur des mots d'un texte naturel a tendance à suivre approximativement une loi de Poisson centrée différemment en fonction du type de texte : complet ou patchwork, comme on peut le voir sur la figure 1. La majorité des générateurs respecte la même distribution, à l'exception des salades de caractères et des salades uniformes de mots. En effet, les salades de caractères ne construisent pas de mots réels, mais combinent aléatoirement des lettres. Il en résulte une distribution décroissante de la longueur des mots produits. Les salades uniformes de mots, même si elles génèrent des mots issus du corpus d'origine ne peuvent elles non plus reproduire la distribution originale.

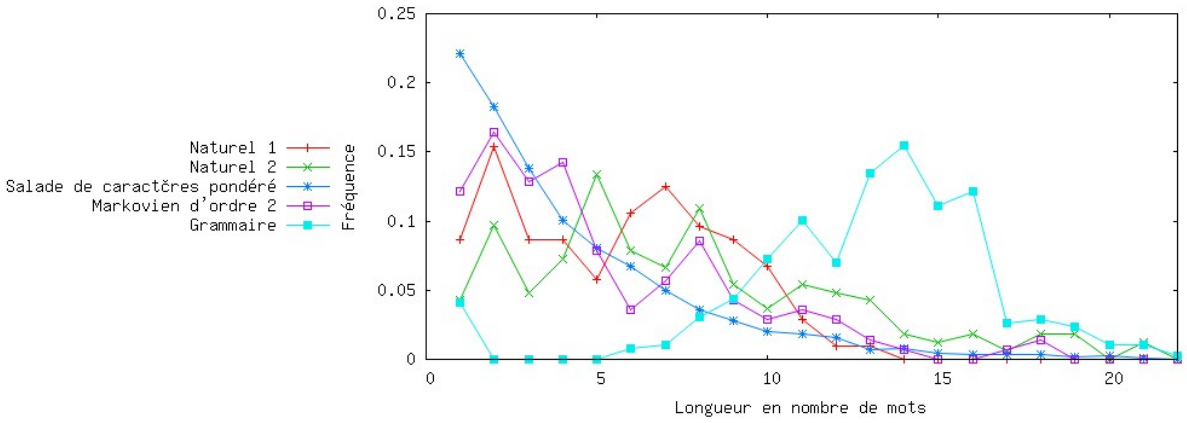

Figure 2 : Distribution de probabilité des phrases de différents types de textes en fonction de leur longueur.

Tout comme la longueur des mots, il est aussi possible de s'intéresser à la longueur des phrases comme illustré dans la figure 2. Les résultats sont comparables à l'exception des textes générés par des grammaires riches, dont la distribution est très particulière et reflète le

faible nombre de phrases de longueur différente que peut produire la grammaire utilisée. Cet artefact est particulier à cette grammaire, mais la difficulté de production de ce type de grammaire rend cette particularité difficile à éliminer.

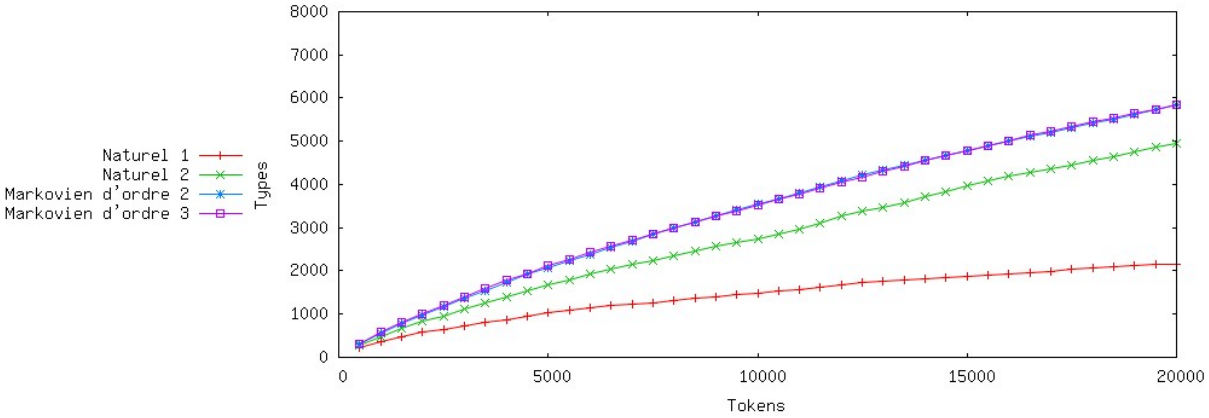

Figure 3 : Nombre de mots du vocabulaire en fonction du nombre d'occurrences de ces mots.

5.3.2. Croissance du vocabulaire

Le test suivant consiste à observer la vitesse de croissance du vocabulaire qui, pour un texte naturel, a tendance à commencer par une croissance forte, très peu de mots ayant déjà été utilisés. Le rythme de croissance se réduit ensuite petit à petit mais ne devient jamais nul. De nouveaux mots sont introduits au fil du texte, le vocabulaire d'une langue étant trop important pour qu'il soit possible de l'utiliser entièrement (Baayen, 2001). On notera *type* les mots du vocabulaire utilisé et *token* les occurrences des types.

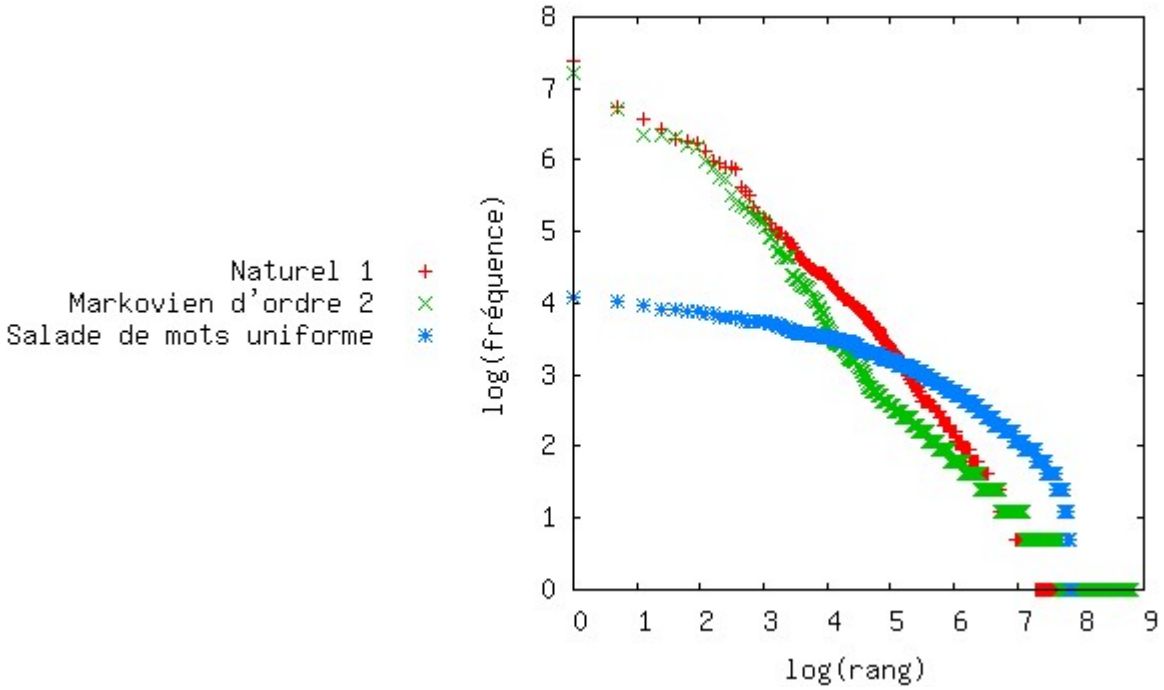

Figure 4 : Courbe de Zipf – Relation entre la fréquence et le rang des mots dans le plan logarithmique

Les patchworks de textes naturels ont une croissance plus rapide due à l'hétérogénéité des thèmes des documents qui les composent. Chaque fois qu'un nouveau texte commence, tout le vocabulaire spécifique à son domaine est introduit comme on peut le voir sur la figure 3.

Les générateurs markoviens se distinguent du corpus *naturel 1* par une plus forte croissance du vocabulaire. En effet, les relations à courte portée qu'ils prennent en compte ne sont pas suffisantes et la thématique du texte change rapidement. Le vocabulaire se diversifie donc plus que pour des textes naturels.

5.3.3. Loi de Zipf

Les tests précédents mettent en évidence le fait que certaines salades, mais aussi des générateurs plus avancés tels que certaines grammaires riches, ne respectent pas la distribution des mots. En effet dans les langues naturelles, la distribution des mots tend à suivre la loi de Zipf, c'est-à-dire que dans un texte naturel la relation entre la fréquence d'un mot et son rang est linéaire dans le plan logarithmique (Zipf, 1949 ; Baayen, 2001).

On peut en effet voir sur la figure 4 que la distribution du texte naturel est globalement linéaire, et qu'il en est de même pour le générateur markovien. Ce n'est en revanche pas le cas de la salade uniforme de mots.

6. Conclusion

On a pu voir dans cet article les formes très diverses de textes non naturels et quelques méthodes statistiques simples qui permettent d'en filtrer une large portion.

Ces méthodes de filtrage trouvent des applications dans de nombreux domaines, tels que l'évaluation des systèmes de génération ou la détection du spam.

Elles constituent une étape préalable à la constitution de filtres plus complexes, mais aussi plus coûteux, permettant la détection de textes artificiels plus sophistiqués.

Références

Baayen R. H. (2001). *Word Frequency Distributions*. Kluwer Academic Publishers.

Dada *The Dada Engine*. http://dev.null.org/dadaengine/, consulté en Septembre 2007.

Gyöngyi Z. and Garcia-Molina H. (2005). Web spam taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.

Hutchins W. J. and Somers H. L. (1992). *An Introduction to Machine Translation*. London: Academic Press.

Interjeet M. (2001). *Automatic Summarization.* John Benjamins Publishing.

Ntoulas A. and Najork M. and Manasse M. and Fetterly D. (2006). Detecting Spam Web Pages through Content Analysis. *International World Wide Web Conference.*

Pomo *Postmodern Essay Generator.* http://www.elsewhere.org/pomo/, consulté en septembre 2007.

Srilm *The SRI Language Modeling Toolkit.* http://www.speech.sri.com/projects/srilm/, consulté en septembre 2007.

Stribling J. and Krohn M. and Aguayo D. *SCIgen – An Automatic CS Paper Generator.* http://pdos.csail.mit.edu/scigen/, consulté en septembre 2007.

Turing A. (1950). Computing machinery and intelligence. *MIND*, Vol.(LIX): 433-460.

Zipf G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, Cambridge, MA.