Longues Phrases dans les codes de la Torah

(Traduit de l'anglais par E. Blum)

Art Levitt[1], *Nachum Bombach* [2],*Harold Gans* [3], *Robert Haralick* [4], *Leib Schwartzman* [5], *Chaim Stal*[6]

[1] Computer Research Analyst, Jerusalem, Israel
[2] Certified Public Accountant and Rabbi, Jerusalem, Israel
[3] Cryptologist, Baltimore, Maryland, U.S.A.
[4] Computer Science Department, City University of New York, U.S.A.
[5] Mathematics and Pattern Recognition, Jerusalem, Israel
[6] Torah Scholar, Jerusalem, Israel

artlevitt23@yahoo.com

Résumé

L'hypothèse de l'existence de codes dans la Torah stipule que la Torah (c'est-à-dire les cinq premiers livres de la Bible Hébraïque) contient des suites de lettres (des codes) qui ont été créées intentionnellement à des fins de communication avec le genre humain, les destinataires visés. Dans la présente étude, nous avons testé un aspect de cette hypothèse relative à des longues phrases, en proposant, dans un premier temps, une méthode pour estimer la probabilité que des suites de lettres extraites d'un texte quelconque puissent former des phrases intelligibles simplement par chance. Nous avons ensuite appliqué cette méthode, dans le cadre de la recherche sur les codes dans la Torah, à une suite de lettres extraite de la Torah relative à Ben Laden. La probabilité (p-level) que cette phrase ait été obtenue purement par chance a été évaluée à 0,000012.

1. Introduction

La recherche sur les codes dans la Torah porte sur un type particulier de suite de lettres que l'on obtient par le prélèvement de lettres régulièrement espacées dans un texte donné. C'est ce qu'on appelle une SLE (Suite de Lettres Equidistantes). Dans la présente analyse, nous nous sommes intéressés au cas d'une longue SLE, composée d'une ou de plusieurs phrases, ce que nous appellerons simplement une « expression ».

Le processus de prélèvement des lettres s'effectue en ignorant toute ponctuation ainsi que les espaces entre les mots. Par exemple, on peut trouver le mot « ecu » en partant du premier « e » de la phrase précédente et en effectuant un saut de +4 lettres (c'est-à-dire, en comptant 4 lettres à partir de la lettre de départ).

Selon l'hypothèse originale, stipulant l'existence de codes dans la Torah, des mots ayant un rapport entre eux sur le plan logique ou historique peuvent être trouvés sous la forme de SLEs de façon significativement plus fréquente et dans des configurations plus compactes dans le texte de la Torah que ce à quoi l'on pourrait s'attendre simplement par chance. De nombreuses études antérieures sur les codes dans la Torah ont porté sur l'analyse d'agrégats de multiples SLEs, par la mesure de la proximité de ces SLEs entre elles dans une matrice donnée [1]-[5].

La présente étude se propose d'ajouter aux méthodes existantes, une méthode d'évaluation plus simple. Le but est d'évaluer une SLE unique extraite d'un texte donné, comme à la figure 1, plutôt qu'un agrégat de SLEs. D'un point de vue géométrique, la configuration est par conséquent plus simple, ce qui permet à un examinateur d'observer le message sous la forme d'une ligne droite (le code représenté est en hébreu et est annoté de traductions françaises).

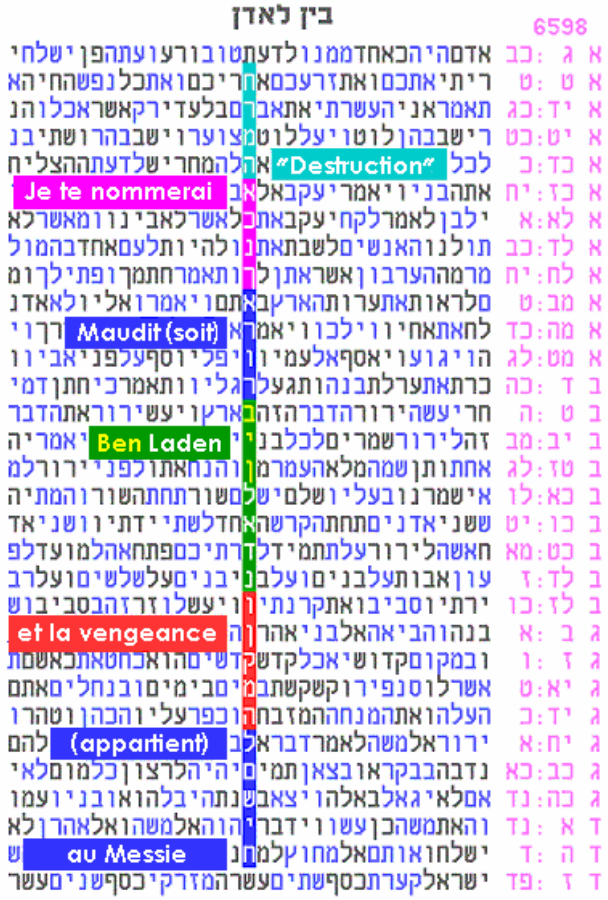

Figure 1: Une SLE dans la Torah d'un intérêt particulier

Une des qualités les plus élémentaires d'un message a trait à son intelligibilité pour celui qui le reçoit. Si

nous formulons l'hypothèse initiale qu'il n'existe pas de codes dans la Torah, nous nous attendons à ce que les expressions trouvées dans le texte de la Torah ne soient pas plus intelligibles que celles que nous serions susceptibles de trouver dans d'autres textes (comparatifs).

2. Présentation de la méthode générale

2.1 Aperçu

Notre méthode, destinée à estimer le caractère signifiant d'une SLE d'une expression trouvée dans un texte particulier, consiste à comparer son intelligibilité à celle d'un vaste ensemble d'expressions compétitrices. Pour cela, nous avons demandé à un nombre important d'examinateurs de classer chaque expression en fonction de son intelligibilité. Sans savoir quelle expression provient du texte original, les personnes en question ont procédé à la classification de cette expression et des expressions compétitrices extraites d'un ensemble de textes comparatifs. La popularité d'une expression est définie comme étant le nombre d'examinateurs qui la qualifient d'intelligible. La popularité relative de l'expression originale parmi les expressions compétitrices, détermine son caractère signifiant. Une expression – provenant d'un quelconque de ces textes – est acceptée et soumise à l'analyse d'un examinateur seulement si les mots qui la composent proviennent d'un lexique de la langue analysée.

2.2 Préparation des données

2.2.1. Le lexique et l'expression originale

Le lexique doit couvrir autant que possible la langue analysée. La SLE de l'expression originale soumise à l'étude est typiquement trouvée en partant d'un mot-clé particulier appelé « (mot) ancrage ». Nous recherchons toutes les SLEs intelligibles inhabituellement longues d'expressions contenant cet ancrage et composées de mots faisant partie du lexique.

2.2.2. Les textes comparatifs

Un texte original peut être modifié pour créer des textes comparatifs. Par exemple, à partir de permutions aléatoires des mots ou des lettres d'un texte original, il est possible de créer par ordinateur un texte comparatif. Un tel texte est qualifié de texte randomisé du fait de son caractère aléatoire. On peut utiliser en plus, le texte original non modifié. Une position de départ aléatoire ainsi qu'une distance de saut dans un tel texte peuvent être choisies par ordinateur. Les positions dans le texte ainsi définies peuvent servir d'ancrage pour une expérience, le processus pouvant être répété des milliers de fois.

2.2.3. A la recherche d'expressions compétitrices

L'obtention de phrases, exprimées sous la forme de SLEs dans un texte comparatif, s'effectue en recherchant de façon exhaustive toutes les possibilités d'espacements permettant d'obtenir des mots composants ces phrases avec comme seule exigence que le tout forme une SLE continue incluant (le mot) l'ancrage choisi. Cette recherche peut conduire à un arbre de possibilités. Par exemple, une expression contenant les lettres « parlesensami », possède au moins 2 branches principales, une branche se situant après le mot « par » (« par le », etc) et une autre après le mot « parles » (« parles en » , etc) ; d'autres sous branches existent aussi. Une expression compétitrice doit avoir une longueur totale ainsi qu'une longueur de mot moyenne égales ou supérieures à celles de l'expression originale.

2.3 Deux sessions d'analyse

2.3.1. Collecte des popularités observées

Les expressions compétitrices provenant des textes comparatifs sont soumises, au cours de deux sessions différentes, à un ensemble important d'examinateurs appartenants à deux groupes distincts. Ces personnes ont procédé – dans le cadre d'un protocole en double aveugle – au classement des expressions listées en les qualifiant « d'intelligibles » ou bien de « non intelligibles ».

Durant la session n°1, chaque expression est soumise à un seul examinateur. Par exception : (1) l'expression originale est mélangée aux expressions compétitrices et occupe une position aléatoire non repérable dans la liste fournie à chaque examinateur, (2) un même petit nombre d'expressions de contrôle (créées manuellement pour paraître intelligibles) mélangées à la liste de chaque examinateur. Le critère pour que les résultats fournis par un examinateur soient retenus est que ce même examinateur choisisse au moins une expression de contrôle mais pas toutes. Cela permet d'écarter ceux qui sont trop stricts ou au contraire trop laxistes dans l'acceptation d'une expression.

Toutes les expressions (y compris l'expression originale) choisies par les examinateurs valides de la session n°1 sont réunies. La liste obtenue est ensuite dupliquée et mélangée aléatoirement à l'attention des examinateurs de la session n°2. Par conséquent, chaque expression acceptée par un examinateur de la session n°1 est jugée par l'ensemble des examinateurs de la session n°2.

2.3.2. Dérivation d'une popularité inhérente

Une attention doit être portée au fait que nous devons avoir l'assurance que notre mesure de la popularité constitue une estimation raisonnable de l'intelligibilité perçue. Nous devons tenir compte du fait que chaque examinateur de la session n°1 agit comme une sorte de garde-barrière qui peut, par conséquent, empêcher que d'authentiques expressions compétitrices atteignent la session n°2. Par conséquent, les tendances des popularités observées – c'est-à-dire, les résultats de la session n°2 – doivent être affinées. Nous sommes vraiment intéressés par les tendances des popularités

inhérentes, c'est à dire les mesures que nous aurions obtenues si toutes les expressions avaient eu la possibilité de franchir le cap.

Les résultats réels de la session n°2 sont utilisés pour effectuer une simulation sur ordinateur afin d'en obtenir les popularités inhérentes. Notre algorithme assigne à chaque expression une popularité inhérente de départ et la soumet ensuite à un examinateur de session n°1 simulé, lui donnant la possibilité de franchir le cap à un niveau ou à une probabilité dictée par la popularité inhérente assignée. Par exemple, si le niveau de la popularité inhérente est fixé à 2 (parmi, par exemple, 22 examinateurs de session n°2 simulés) l'expression concernée a une chance de 1/11 de franchir le cap lors de la simulation. Si elle franchit le cap, elle possède une chance équivalente d'être retenue par chacun des examinateurs de la session n°2 simulés. Des milliers d'itérations sont effectuées, chacune permettant d'ajuster légèrement les popularités inhérentes, jusqu'à obtenir la meilleure adéquation avec les popularités réelles observées durant la session n°2.

2.4. Obtention du niveau du caractère signifiant

Le niveau du caractère signifiant, **P** (p-level), pour l'expérience est donné par la formule:

$$P = s / t \tag{1}$$

où **s** représente le nombre d'expressions – non comprises les expressions de contrôle – ayant une popularité inhérente supérieure à l'expression originale, plus la moitié du nombre d'expression – non comprises les expressions de contrôle – ayant une popularité inhérente égale à l'expression originale ; et où **t** représente le nombre de textes acceptés, nombre estimé, à partir du ratio d'expressions classées, comme suit : **m** représente le nombre total d'expressions – non comprises les expressions de contrôle – fournies à l'ensemble des examinateurs de la session n°1 ; **v** représente le nombre des expressions ayant effectivement fait l'objet d'un classement par les examinateurs valides de la session n°1 (sont exclues les expressions non examinées du fait d'une indécision ou par manque de temps). **N** représente le nombre total de textes utilisés lors de l'expérience. La valeur estimée de **t** est donnée par la formule suivante:

$$t = (v / m)N \tag{2}$$

3. Cas étudié

3.1. Description

Le cas étudié (figure 1) est relatif à une SLE particulière, trouvée dans la Torah, d'une expression qui consiste en une phrase dont la traduction est **: Je te nommerai « destruction ». Maudit (soit) Ben Laden et la vengeance (appartient) au Messie.** Notre propos n'a pas pour objet d'essayer d'interpréter cette expression mais seulement de juger son intelligibilité.

3.2. Les résultats de la préparation des données

3.2.1. Le léxique

Notre léxique est construit à partir de deux sources, l'une ancienne, l'autre moderne : (1) Nous avons utilisé environ 40000 mots issus de la Bible hébraïque (le livre de Daniel a été exclu dans la mesure où il contient de nombreux mots non hébreux (araméens). (2) Nous avons utilisé tous les mots du site Internet d'actualités en hébreux, Arutz-7, de l'année 2002. Cette seconde source a porté la taille totale du lexique à plus de 107000 mots.

3.2.2. Génération des textes comparatifs

Les textes comparatifs ont été créés à partir de deux sources différentes : (1) un ensemble de 307200 textes de la Torah permutés – résultat de la permutation de 25600 textes selon 12 méthodes différentes de permutation : permutation des lettres dans un mot, dans un verset, dans un chapitre, dans un livre, dans un texte ; permutation des mots dans un verset, dans un chapitre, dans un livre, dans un texte ; et permutation des versets dans un chapitre, dans un livre, dans un texte ; (2) un segment de texte de la Bible de même longueur que la Torah permettant de générer un nombre important de cas supplémentaires. Ce segment commence tout de suite après le dernier mot de la Torah (le début du livre de Josué) pour se terminer avec le $7^{ième}$ mot du livre des Rois II, 18:24. A partir de ce segment et pour chaque essai, nous avons choisi aléatoirement la position (du mot) de l'ancrage (« Ben Laden »). Nous avons examiné approximativement le même nombre de SLEs pour chacune des deux sources et par conséquent nous considérons que le nombre de textes examinés (**N**) est de 2*307200 = 614400.

3.2.3.Identification des expressions compétitrices

Pour chaque occurrence (du mot) de l'ancrage repérée dans un texte comparatif, l'ordinateur cherche les expressions ayant une longueur au moins égale à l'originale (c'est à dire 29 lettres) et possédant une longueur de mot moyenne au moins égale à l'originale (soit 29/6, car nous considérons que (le mot) l'ancrage ainsi que sa lettre préfixe optionnelle forment un seul mot). Cette opération a fourni **m** = 13430 expressions.

3.3. Résultats des sessions d'analyse

Les 13430 expressions compétitrices ont été distribuées à 64 examinateurs de la session n°1, de telle sorte que chacun en a reçu approximativement 210 avec en plus les 8 expressions de contrôle ainsi que l'expression originale, le tout ayant été mélangé de façon aléatoire. 62 des 64 étudiants se sont révélés valides, chacun ayant qualifié d'intelligible de 1 à 7 expressions de contrôle (il est à noter que 41 d'entre

eux ont accepté l'expression originale). Un total de **v** = 12880 expressions – non comprises les expressions de contrôle – ont fait l'objet d'un classement, 204 d'entre elles ont été qualifiées « d'intelligibles ».

Table 1: Résultats pour les expressions – non comprises les expressions de contrôle

Niveau de Popularité	Nombre d'expressions (Observées)*	Nombre d'expressions (Inhérentes)
1	36	1331
2	16	150
3	9	39
4	3	9
5	2	5
6	2	4
7	3**	5**
8	0	0
9	0	0
10	1	1

* 133 expressions – non comprises les expressions de contrôles – n'ont reçu aucun vote lors de la session n°2.
** Comprend l'expression originale

Chacun des 27 examinateurs de la session n°2 a reçu la liste complète des 204 expressions, liste triée d'une seule façon, avec en plus les 8 expressions de contrôle ainsi que l'expression originale. 22 examinateurs de la session n°2 se sont révélés valides, chacun ayant retenu de 1 à 7 expressions de contrôle. 3 parmi les expressions de contrôle ont reçu plus de vote que l'expression originale. Les résultats pour les expressions – non comprises les expressions de contrôle – figurent dans la table 1.

La table présente les valeurs observées ainsi que l'estimation des valeurs inhérentes obtenues à partir d'une simulation (effectuée comme proposée à la section 2.3.2). La simulation est en accord avec une évaluation logique, à savoir que pour un niveau de popularité inhérente bas, il y a un plus grand excès d'expressions. Cela est dû au fait qu'à ce niveau de popularité, le « garde-barrière » éprouve une plus grande difficulté à accepter l'expression.

3.4. Résultats des calculs du caractère signifiant

3.4.1. Résultat initial

A partir des équations (1) et (2), nous obtenons d'abord **s** = 3.5 (qui correspond à la moitié des 5 expressions ayant une popularité inhérente égale à 7 plus 1 expression dont la popularité inhérente est égale à 10). Avec les différents paramètres : **m** = 13430; **v** = 12880 et **N** = 614400, nous obtenons **t** = 589238. Donc:

P = **s** / **t** = 5.9 e-06 (valeur préliminaire)

3.4.2. Ajustement du fait d'une variation dans l'orthographe

Il existe une autre orthographe en hébreu, pour Ben Laden, qui ne contient pas la lettre « youd » et qui est utilisée environ 50% du temps d'après une recherche effectuée avec le moteur de recherche Google en hébreu. Par conséquent, nous divisons par 2 notre niveau du caractère signifiant. Nous obtenons alors la valeur finale de p:

P = 1.2 e-05, environ 1 sur 83,000.

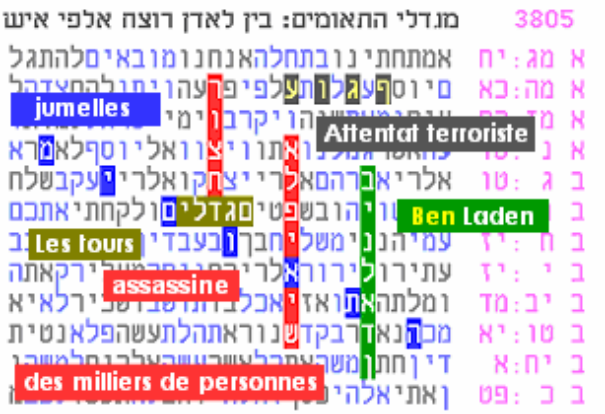

Figure 2: Un exemple d'une matrice en rapport avec Ben Laden

4. Discussion

Le résultat présent ne doit pas être considéré comme un cas isolé. En effet, il existe 5 autres matrices sur le même sujet [6], mêlant de façon hautement significative le pattern colinéaire (la longue phrase) et deux autres types de pattern mis en exergue ces dernières années dans le cadre de différentes études : les SLEs parallèles et horizontales. La figure 2 ne représente qu'une de ces 5 matrices. Du fait d'une part, que la grande majorité des mots clés trouvés dans ces matrices sont des mots choisis à priori (figurant parmi les mots les plus cités dans les nouvelles en hébreu relatives au 11 septembre) et d'autre part, d'une répétition hautement significative de mots ou de thèmes dans les différentes matrices apparentées, il se trouve que certaines d'entre ces matrices ont une probabilité (p-level) au moins aussi significative que celle de notre étude.

La présente méthode pourrait être adaptée afin de tester une implication ultérieure de l'hypothèse de l'existence de codes dans la Torah : si les codes constituent véritablement un message intentionnel plutôt que des occurrences dues au hasard, il en résulte que les capacités de l'auteur de ces codes dépassent de très loin celles de l'être humain. Dans le cadre de futures études portant sur des longues phrases, il serait intéressant d'examiner les qualités perçues de chaque auteur de phrase. Par exemple, les examinateurs pourraient évaluer le « niveau de sagesse » de chaque phrase comme étant du niveau du « singe », de « l'enfant », de « l'adulte », du « prophète » ou bien « surnaturel ». Cette quantification pourrait au final fournir une sorte

d'évaluation mixte relative à l'intelligibilité et la sagesse de la phrase analysée.

5. Conclusion

Nous avons proposé une méthode permettant d'estimer le caractère signifiant d'une longue SLE d'une expression trouvée dans un texte, cette estimation étant le résultat de l'appréciation de son intelligibilité par un nombre important d'examinateurs. En utilisant cette méthode, nous avons démontré que dans la Torah, le niveau du caractère signifiant d'une SLE particulière d'une expression concernant Ben Laden était de 0,000012. Ce résultat ainsi que les matrices apparentées, mentionnées à la section 4, suggèrent fortement que ce sujet est intentionnellement codé dans la Torah. De plus, dans la mesure où ces résultats concernent une des figures les plus mentionnées dans les nouvelles d'aujourd'hui, cela démontre implicitement qu'un tel niveau du caractère signifiant a été obtenu assez facilement sans qu'il ait été nécessaire d'effectuer des recherches exhaustives ; ce qui renforce de façon remarquable l'hypothèse de l'existence de codes dans la Torah.

6. Remerciements

Les auteurs adressent leurs remerciements à Eliyahu Rips ainsi qu'à Yechezkel Zilber pour leurs précieux conseils tout au long de l'étude. La session n°1 s'est déroulée le 18 novembre 2003 à la Yeshiva Shaalbim ; la session n°2 s'est déroulée le 15 janvier 2004 à la Yeshiva Nehora

7. Références

[1] D. Witztum, E. Rips, Y. Rosenberg; Equidistant Letter Sequences in the Book of Genesis; *Statistical Science,* 9(3):429-438, 1994.

[2] H. Gans (2001); Torah Codes Primer; http://aish.com/seminars/discovery/Codes/codes.htm

[3] R. Haralick (2003); Testing the Torah Code Hypothesis; http://www.torahcodes.net/hypoth.html

[4] R. Haralick (2003); Torah Codes: Redundant Encoding; http://www.torahcodes.net/redun.html

[5] D. Witztum; Torah Codes; http://www.torahcodes.co.il

[6] A. Levitt (2004); Twin Tower Codes; http://www.torahcodes.net/alltwin.html