Dimulescu, A. & Dessalles, J-L. (2009). Prédire l'intérêt dans la communication événementielle. In N. Maudet, P.-Y. Schobbens & M. Guyomard (Eds.), *Modèles formels de l'interaction (MFI-09) - Actes des cinquièmes journées francophones*, 125-134. Lannion. www.dessalles.fr/papiers/pap.conv/Dessalles_09042202.pdf

Prédire l'intérêt dans la communication événementielle

A. Dimulescu[1] J.-L. Dessalles[1]
{adrian.dimulescu, jean-louis.dessalles}
@telecom-paristech.fr

[1]Telecom ParisTech – LTCI UMR-5141
75013 Paris – FRANCE

Résumé :
Cette étude vise à mesurer les variations de l'intérêt suscitées par un événement lorsque certaines dimensions définies sont manipulées. Les résultats sont comparés aux prédictions de la théorie du décalage de complexité, selon laquelle les événements sont d'autant plus intéressants qu'ils sont plus simples qu'attendu.
Mots-clés : Complexité, simplicité, intérêt, événement.

Abstract :
This study is an attempt to measure the variations of interest aroused by conversational narratives when definite dimensions of the reported events are manipulated. The results are compared with the predictions of the Complexity Drop Theory, which states that events are more interesting when they appear simpler than anticipated.
Keywords: Complexity, simplicity, interest, event.

1 Introduction

L'activité qui consiste à signaler ou rapporter des événements occupe une part significative du temps humain. Si l'on considère que le temps de conversation représente jusqu'à un tiers du temps d'éveil (Mehl & Pennebaker, 2003) et que les narrations peuvent constituer jusqu'à 40% de ce temps de parole (Eggins & Slade, 1997), nous pouvons passer plus du dixième de notre temps disponible à communiquer des événements dignes d'intérêt, ce qui est considérable. Cette activité présente un enjeu social important, comme l'ont révélé les études sociolinguistiques (p.ex. Labov & Waletzky, 1967; Labov, 1997; Tannen, 1984) : celui ou celle qui est capable d'intéresser son auditoire renforcera de ce fait ses liens sociaux et, inversement, ceux qui suscitent l'ennui en rapportant des événements sans intérêt courent le risque de se retrouver isolés. Malgré son importance, l'activité narrative spontanée n'a été décrite que récemment (voir notamment Norrick, 2000) et, à notre connaissance, n'a pas fait l'objet de modélisation.

L'analyse de la sélection des événements rapportés au cours des interactions spontanées et des paramètres qui contrôlent leur intérêt nous a permis de développer un modèle fondé sur la complexité (Dessalles, 2006, 2008b). Notre objectif ici est de tester cette théorie en analysant les préférences des sujets concernant les variantes de petites histoires.

Dans ce qui suit, nous rappelons les éléments principaux de la théorie du décalage de complexité, puis nous décrivons notre expérience. Nous terminons par une discussion sur la portée des résultats obtenus.

2 La théorie du décalage de complexité

Même si les facteurs de l'intérêt narratif des événements n'ont pas été répertoriés, il y a eu des tentatives dans des domaines connexes comme le journalisme ou l'étude de la mémoire. Selon les travaux classiques du journalisme (Galtung et Ruge, 1965), la récence, la proximité, le caractère exceptionnel ou négatif des événements et le nombre d'individus affectés ont une influence décisive sur la valeur de l'information (*newsworthiness*). Un moyen indirect d'évaluer les facteurs qui jouent sur l'intérêt narratif consiste à observer les paramètres qui influent sur la mémorisation des événements. Les individus mémorisent plus efficacement les comportements qui leur paraissent incohérents ou qui s'écartent des stéréotypes (Stangor et McMillan 1992), notamment les actions atypiques (Woll et Graesser 1982 ; Shapiro et Fox 2002). Nous suggérons qu'une seule propriété, l'*inattendu*, résume l'ensemble de ces caractéristiques. L'autre dimension de l'intérêt des événements, l'intensité émotionnelle (Rimé, 2005), ne sera pas considérée ici.

Selon la théorie du décalage de complexité (*Complexity Drop Theory*, CDT), les événements dignes d'intérêt sont ceux qui sont *plus complexes à produire qu'à décrire*. Formellement[1] :

$$U(s) = C_w(s) - C(s)$$

$U(s)$ mesure l'inattendu de la situation s. $C(s)$ est sa complexité, c'est-à-dire la taille de la description la plus concise de s, autrement dit la quantité d'information minimum permettant de désigner s sans ambigüité (Li & Vitányi, 1993). $C_w(s)$ est la complexité de *production*, mesurée par la quantité d'information minimum permettant de produire s à partir du monde w tel qu'il connu ou imaginé et en tenant compte de ses contraintes.

La notion de complexité apparaît de plus en plus comme fondamentale en sciences cognitives (Chater, 1999 ; Chater & Vitányi, 2003). Par exemple, la reconstitution des parties cachées obéit à un principe de complexité minimale. Il en est de même pour l'induction (Solomonoff, 1978) et pour la formation d'hypothèse (Chaitin, 2004). Le reproche qui est parfois adressé aux théories utilisant la complexité est que cette notion reste abstraite, puisqu'on ne peut pas la calculer en général. Ce reproche est infondé dans de nombreux cas d'intérêt pratique (Delahaye & Zenil, 2007), et particulièrement dans le cas de la complexité cognitive (Dessalles 2008b). La suite du présent article va permettre d'illustrer la fécondité du concept de complexité en testant ses prédictions concernant l'intérêt des événements.

3 Description de l'expérience

Pour tester les prédictions de la CDT, nous avons constitué un corpus de 18 histoires en français, pour partie inspirées d'événements rapportés sur un site Web communautaire (viedemerde.fr) sur lequel les utilisateurs décrivent des petits événements de la vie quotidienne dans un style informel. Les histoires ont été proposées à 95 participants par l'intermédiaire d'un questionnaire sur le Web. Notre objectif était de démontrer que, derrière la richesse et la variété phénoménologique apparente de ces histoires, un déterminisme simple est à l'œuvre. Voici un exemple d'histoire, suivi des trois options possibles.

> Les policiers du commissariat d'Antibes ont découvert dans l'après midi dans la baie des Anges les corps flottants de deux femmes aux élégantes tenues vestimentaires. Apparemment les deux noyades se sont passées à peu près en même temps. De plus, les deux femmes [...].
>
> a: avaient chacune un tatouage rouge représentant le dragon Tsuba-Kasai sur le bras droit
> b: avaient chacune un tatouage rouge sur le bras droit
> c: avaient chacune un tatouage sur le bras droit

On ne saurait être trompé par le fait que les réponses attendues pour ce test semblent souvent « évidentes ». L'explication ne l'est pas. Cette situation rappelle celle des études sur la syntaxe, où l'explication scientifique d'une intuition largement partagée se révèle très difficile à dégager.

Toutes les histoires ont été présentées dans un ordre aléatoire à tous les participants. La plupart des participants étaient des ingénieurs et des étudiants travaillant dans les domaines autres que les sciences cognitives. Pour chaque histoire, nous avons isolé un paramètre considéré comme important pour la pertinence de l'histoire et avons défini trois options qui affecteraient graduellement l'intérêt. Chaque participant ne voyait que deux options pour chaque histoire, de sorte que le but global du test ne soit pas transparent. Le test était proposé sous la forme d'une application Web. Les participants devaient compléter un passage manquant pour chaque histoire en cliquant sur une option parmi deux choisies aléatoirement parmi les trois possibles.

À la fin de l'expérience, nous avons donc obtenu pour chaque histoire un ensemble de préférences deux-à-deux entre les trois versions possibles. Pour comparer ces résultats aux prédictions de CDT, il nous fallait agréger les préférences en une hiérarchie. En suivant (Saaty, 1994), nous avons calculé le rang r_i de la version i à partir de sa dominance sur les autres versions, représentée par les poids w_{ij} =

[1] Pour un tutoriel sur CDT, voir le site www.unexpectedness.eu.

win(i)/win(j), où win(i) est le nombre de réponse où la version i a été choisie. Pour ne pas diviser par zéro, un vote a été accordé par défaut à chaque option. Par exemple, si 30 participants ont eu à choisir entre les options i et j d'une histoire donnée et que 23 ont choisi la première version tandis que les 7 participants restants choisissent la seconde, nous avons w_{ij} = 24/8 = 3. Notez que w_{ij} = 1/ w_{ji}. Nous voulons favoriser le rang d'une version d'histoire qui a été préférée aux autres versions, en tenant compte du rang des versions délaissées et de l'amplitude de la dominance. Un calcul possible est $r_i = k \Sigma_j w_{ij} r_j$, que l'on peut réécrire comme $R = kWR$, ce qui revient à trouver un vecteur propre de la matrice W (Farkas 2007). Les rangs sont ensuite normalisés pour que leur somme soit 100. Par exemple, les options de l'histoire précédente sur le double suicide obtiennent les rangs suivants :

Préférences		Rangs
a-b	17/14	a 40
b-c*	34/5	b 49
a-c*	23/9	c 11

Pour chaque histoire, les versions sont ordonnées de la plus intéressante à la moins intéressante selon CDT. Les préférences identifiées par * ont été trouvées statistiquement significatives ($p < .05$) sur un test binomial standard visant à détecter une préférence marquée pour une version particulière.

4 Analyse

4.1 Coïncidences

La théorie CDT permet de prédire la fascination qu'exercent les coïncidences, ainsi que la manière dont cette fascination dépend des paramètres de l'histoire (Dessalles 2008a). Dans une coïncidence comme celle qui est rapportée dans l'histoire du double suicide, l'intérêt résulte de l'observation de deux situations analogues s_1 et s_2. Nous pouvons écrire $C(s_1\&s_2) \leq C(s_1) + C(s_2|s_1)$, où la complexité conditionnelle $C(s_2|s_1)$ mesure la taille de la description la plus concise de s_2 lorsque s_1 est disponible. Si les deux situations sont indépendantes, nous pouvons écrire par ailleurs : $C_w(s_1\&s_2) = C_w(s_1) + C_w(s_2)$. Si l'on considère que les deux événements ont la même complexité : $C_w(s_1) \approx C_w(s_2)$ et que, séparément, ils ne sont pas inattendus : $C_w(s_1) = C(s_1)$, alors :

$$U(s_1\&s_2) \geq C(s_1) - C(s_2|s_1)$$

Le modèle CDT prédit ainsi qu'une coïncidence est d'autant plus intéressante que les événements coïncidents sont complexes et qu'ils sont fortement analogues, puisque l'analogie a pour effet de diminuer $C(s_2|s_1)$ (Cornuéjols, 1996). Cette prédiction est globalement vérifiée dans l'histoire suivante.

> J'ai acheté une petite Peugeot 106 ColorLine, que j'ai payée 2000 euros. Je l'avais essayée la veille, elle roulait très bien. Je tourne la clé, je démarre, je sors de la propriété de l'ancien propriétaire de la voiture quand, venant de gauche sans regarder, une autre [...] me rentre dedans.
>
> a: Peugeot 106 ColorLine
> b: Peugeot 106
> c: Peugeot

La contribution de la marque de la voiture à l'inattendu peut faire l'objet d'un calcul précis. Pour estimer $C_w(s_1) = C(s_1)$, on peut considérer que le « choix » de la voiture impliquée dans l'accident demande une complexité $\log_2 N$, où N est le nombre de toutes les voitures de la région. Le calcul de $C(s_2|s_1)$ peut utiliser la marque f de la voiture, qui est disponible gratuitement dans s_1. Il suffit alors de $\log_2 n_f$ pour discriminer la voiture impliquée, si n_f est le nombre de voitures de la marque f. Ainsi :

$$U(s_1\&s_2) \geq \log_2 N - \log_2 n_f$$

La prédiction est donc que la caractéristique f commune aux deux véhicules augmentera d'autant plus l'intérêt qu'elle sera précise. C'est bien ce qui a été observé.

Préférences		Rangs
a-b*	24/8	a 61
b-c*	32/4	b 32
a-c*	28/5	c 7

Dans l'histoire du double suicide mentionnée plus haut, chaque option offre des caractéristiques qui augmentent $C(s_1)$ mais,

puisqu'elles sont communes aux deux situations coïncidentes, ces caractéristiques laissent $C(s_2|s_1)$ inchangée. La prédiction du modèle se trouve globalement vérifiée dans cette histoire, puisque l'option (c) y est largement dominée. Noter toutefois que l'option (a) n'obtient pas le meilleur rang, contrairement à la prédiction de CDT (voir partie discussion).

4.2 Déviation quantitative

La CDT prévoit que les situations atypiques sont intéressantes parce qu'il est facile de les singulariser. Nous supposons que la situation atypique est vue comme appartenant à un cadre de référence *r* et qu'elle s'écarte du prototype de *r* de *k* écarts-types pour la caractéristique *f*. Si le « monde » est comparé dans ce cas à une loterie, alors $C_w(s|r) = \log_2 N$, où N est le nombre de situations correspondant à *r*. La complexité pour décrire *s* à partir de *r* respecte l'inégalité : $C(s|r) \leq C(f|r) + C(s|r\&f)$. Nous supposerons pour simplifier que *f* n'est pas conceptuellement liée à *r* : $C(f|r) = C(f)$. Les situations couvertes par *r* peuvent être rangées selon les valeurs de *f* avec une complexité additionnelle négligeable (noter que la complexité se rapporte à la taille des algorithmes, pas à leur temps d'exécution). Si *s* est extrême pour *f*, elle apparaîtra parmi les premiers éléments de ce tri. Nous pouvons écrire $C(s|r\&f) \approx \log_2 N - A(k)$, où la fonction A dépend seulement de la distribution statistique de *r* le long de *f* (pour une distribution gaussienne, $A(k) \sim k^2$). Nous obtenons finalement :

$$U(s|r) \geq A(k) - C(f)$$

L'histoire suivante teste la prédiction.

> La Police d'Amiens a effectué mercredi une saisie [...] d'héroïne au numéro 13 de la rue Fafet.
>
> a: de 10 kilos
> b: de 5 kilos
> c: de 2 kilos

Pour cette histoire, nous devons supposer que les participants ont une idée des saisies typiques de drogue telles qu'elles sont rapportées dans les nouvelles, même si cette idée ne porte que sur l'ordre de grandeur. Il leur suffit de comprendre que les options proposées impliquent des valeurs différentes de *k*. Dans ce cas, la prédiction est qu'ils trouveront l'histoire plus intéressante lorsque la saisie est plus importante.

Préférences		Rangs
a-b*	28/13	a 54
b-c	20/11	b 28
a-c*	21/8	c 18

L'expérience montre bien que la hiérarchie (a) > (b) > (c) est respectée.

4.3 Déviation qualificative

Certains événements se produisent même si l'on pouvait penser qu'ils étaient presque impossibles. Cette propriété signifie simplement que la complexité $C_w(s)$ nécessaire pour produire un tel événement *s* est très élevée. Elle s'élève à $C(H)$, où *H* est le scénario causal le plus simple qui explique *s* en tenant compte du fonctionnement du monde connu (voir la partie discussion). Selon CDT, l'inattendu vaut :

$$U(s) = C(H) - C(s)$$

Considérons l'histoire suivante.

> Je marchais tranquillement dans la rue quand un parfait inconnu s'arrête devant moi, me regarde et [...] avant de continuer sa route.
>
> a: me donne une gifle phénoménale
> b: me donne une gifle
> c: me demande l'heure

Dans cette histoire, expliquer pourquoi quelqu'un demande l'heure (option (c)) peut être aussi simple que « parce qu'il a oublié sa montre », tandis qu'une explication d'une gifle immotivée exigerait un scénario autrement plus compliqué, et par conséquent une complexité $C(H)$ beaucoup plus élevée. D'autre part, la complexité $C(s)$ peut être calculée en utilisant un cadre de référence *r* et une caractéristique *f* de *s* : $C(s) \leq C(r) + C(f|r) + C(s|r\&f)$. Dans l'histoire précédente, *r* correspond à une scène typique dans la rue. Pour les options (a) et (b), *f* correspond au fait d'être giflé par un étranger. Si une telle description peut être considérée comme capturant une situation unique, alors

$C(s|r\&f) = 0$. En revanche, $C(s|r\&f)$ a une valeur significative pour l'option (c) (demander l'heure), puisque les scènes où une personne demande l'heure dans la rue sont nombreuses et ainsi difficiles à distinguer. L'inattendu sera ainsi élevé pour les options (a) et (b) et proche de zéro pour (c).

Préférences		Rangs
a-b	22/18	a 53
b-c*	25/4	b 41
a-c*	27/3	c 6

L'expérience montre que, conformément à CDT, l'option (c) est rejetée.

4.4 Proximité

Des événements sont plus intéressants s'ils se sont produits à proximité de l'observateur ou à proximité d'un élément déjà connu. La proximité peut être vue comme une relaxation de l'analogie. Dans l'analogie, $C(s_2|s_1)$ est diminuée par la présence d'éléments communs entre s_1 et s_2. La proximité diminue également la quantité d'information requise pour définir la situation s_2 à partir de s_1, en utilisant cette dernière comme point de référence à partir duquel il est plus facile de localiser s_2. La CDT prévoit ainsi que des événements indépendants se produisant à proximité l'un de l'autre seront ensemble plus intéressants que si leurs occurrences étaient éloignées.

Dans le domaine temporel, si une classe d'événements se produit avec la densité D_t, la CDT prévoit que la contribution à l'inattendu d'une occurrence de l'événement observée à distance t_1 est[2] :

$$U = -\log_2 (D d)$$

L'intérêt de l'événement décroît comme le logarithme de la distance temporelle. L'histoire qui suit offre un exemple de proximité temporelle inattendue.

> Cela fait un an que je dois changer de portable chez SFR. Je me décide enfin à y aller même s'il faut que je paye une partie car je n'ai pas assez de points carrés rouges. J'ai acheté le portable à 13h00. [...] je reçois un message de SFR : "Changez de mobile, SFR vous offre 15 000 points carrés rouges."
>
> a: A 13h10
> b: A 14h00
> c: Deux semaines plus tard

Dans cette histoire, on s'attend à des offres promotionnelles de la part de l'opérateur environ tous les deux mois. La production qu'une telle offre arrive à la minute précise t_2 où elle est arrivée demande une complexité égale à $C_w(t_2) = \log_2 86400 = 16.4$ bits. La description de l'événement peut utiliser le moment t_1 de l'achat du mobile comme point de référence. Ainsi, $C(t_2|t_1) = \log_2 10 = 3.3$ bits. Ce contraste contribue à hauteur de 13 bits à l'inattendu (noter que la valeur serait la même avec une résolution temporelle différente). Les trois options de l'histoire précédente produisent un inattendu temporel de 13 bits, 10.5 bits et 2 bits respectivement. Les préférences des participants confirment nettement cet effet.

Préférences		Rangs
a-b*	29/4	a 79
b-c*	28/6	b 16
a-c*	31/3	c 5

L'histoire suivante illustre l'influence de la proximité sociale.

> Deux semaines après le vol de ma voiture, la police m'apprend qu'une voiture qui a des chances d'être la mienne est en vente sur Internet. Ils me présentent l'annonce, le numéro de portable est identifié comme étant [...].
>
> a: celui de mon collègue de bureau
> b: celui d'un collègue de mon frère
> c: celui de quelqu'un qui habite le même quartier

L'inattendu, ici, vient du fait que le voleur se trouve être plus simple qu'attendu. Le fait qu'une personne inconnue P soit le voleur a une complexité $C_w(P)$. On peut l'estimer, en comparant la désignation du voleur à une loterie, par le logarithme $\log_2 N$ de la population de la région que le sujet prend comme référence, probablement la ville. Si P se révèle être quelqu'un qui vit dans un voisinage de taille n, alors la contribution à l'inattendu est $U = \log_2 N - \log_2 n - C(d)$, où d

[2] Voir les détails sur http://www.unexpectedness.eu/NextDoor.html

est le concept de voisinage considéré. Ce calcul extensionnel représente un dernier recours, car un calcul sur le graphe social fournit souvent une plus petite valeur pour $C(P)$. Dans ce cas-là, la complexité de P est l'information minimum requise pour atteindre le nœud de P dans le graphe social. Dans l'histoire du voleur, l'expression « mon collègue » suggère que la complexité de P est zéro une fois que le concept du collègue est installé, car P vient en premier dans la liste. Les résultats confirment clairement la prévision : l'option (a) (mon collègue) est sensiblement préférée à (b) (un collègue de mon frère), et (a) et (b) sont préférées à (c) (quelqu'un vivant dans mon voisinage).

Préférences	Rangs
a-b* 25/6	a 65
b-c* 24/10	b 22
a-c* 26/7	c 13

4.5 Rencontres fortuites

Les rencontres fortuites fournissent la matière pour de nombreuses conversations. L'inattendu qu'elles produisent s'écrit :

$$U = C(l) - C(P)$$

où $C(l)$ est la complexité du lieu de la rencontre et $C(P)$ celle de la personne rencontrée (Dessalles, 2008a). Il est donc important, pour l'intérêt de l'histoire, que la personne rencontrée soit simple. L'histoire suivante a été proposée aux participants pour tester ce phénomène.

> Je passais dans une rue en plein milieu de Paris quand j'ai entendu quelqu'un m'appeler: c'était un gars [...] quand il était gamin. Nous avons échangé nos adresses. Ca m'a fait super plaisir.

a: que j'ai babysitté pendant 2 ans
b: que j'ai babysitté pendant 2 mois
c: que j'ai babysitté quelques soirs

Une période de fréquentation plus longue place l'individu P plus haut dans la liste des personnes que l'on connaît personnellement. La complexité $C(P)$ peut être estimée par le logarithme du rang dans cette liste. Les résultats confirment cette prédiction.

Préférences	Rangs
a-b* 25/11	a 58
b-c 20/12	b 26
a-c* 25/7	c 16

Noter que la préférence de (b) à (c) est faible, ce qui s'explique par le fait que les deux options ne modifient pas sensiblement le degré d'accointance passée avec P.

4.6 Structure

L'une des prévisions les plus immédiates de la CDT est que l'occurrence de structures remarquables augmente l'intérêt. Le « monde » exige les mêmes efforts pour produire une structure typique non remarquable s_r et pour produire une structure remarquable s : $C_w(s_r) = C_w(s)$. Mais les structures remarquables sont simples et ainsi inattendues :

$$U(s) = C(s_r) - C(s)$$

Dans l'histoire suivante, l'option (a) (numéro 4444) économise trois instanciations, car le chiffre 4 est simplement copié. L'inattendu vaut donc $U_a = 3 \times \log_2 10 = 10$ bits. Dans l'option (b), pour la même raison, l'inattendu s'élève à au moins $U_b = 6.6$ bits, tandis qu'il vaut zéro pour l'option (c).

> C'est marrant, j'ai trouvé ça sur Internet: la ville de St-Chéron compte [...] habitants.

a: 4444
b: 4000
c: 3856

Préférences	Rangs
a-b* 35/1	a 92
b-c* 21/11	b 4
a-c* 30/2	c 3

Les résultats confirment ces prévisions. La structure la plus simple (a) a été fortement préférée à (b) et à (c). L'option (b) a été préférée à (c), bien que moins de manière significative.

5 Discussion

La plupart des résultats confirment les prédictions de la CDT. Il y a quelques légères

différences, que nous allons commenter. Auparavant, nous devons répondre à la question récurrente concernant la non-calculabilité théorique de la complexité.

5.1 Mesurer la complexité

La non-calculabilité théorique de la complexité de Kolmogorov est un résultat facile à démontrer. Nous évitons cette difficulté en considérant deux restrictions. D'abord, nous définissons la complexité, non comme un minimum objectif, mais comme la longueur de la description la plus concise *disponible*. Ceci peut mener à une sous-estimation de l'inattendu et expliquer pourquoi quelques sujets peuvent ne pas saisir l'intérêt que d'autres sujets trouvent dans une histoire. Notre deuxième restriction est que la complexité cognitive est calculée sur une machine spécifique, « la machine cognitive », c.-à-d. l'ensemble des outils cognitifs dont les humains sont censés disposer. Ceci présuppose que le modèle cognitif que nous employons soit rendu explicite dans chaque cas. Dans les exemples commentés précédemment, la plupart des calculs ont été faits en utilisant des hypothèses minimales en ce qui concerne ces capacités cognitives. Comme nous l'avons montré, le calcul de la complexité cognitive est parfaitement possible dans la plupart des cas. Trois aspects de la complexité sont, cependant, externes au modèle : la complexité conceptuelle, la complexité structurelle et la complexité causale.

La *complexité conceptuelle* intervient quand des prototypes et des propriétés caractéristiques, notés *r* et *f* dans nos exemples, sont impliqués. Une méthode pour évaluer $C(r)$ ou $C(f)$ serait d'utiliser des graphes ontologiques ou des distances dans des corpus comme Wikipedia. Nous sommes en train d'explorer cette possibilité.

La *complexité structurelle* peut être approchée en utilisant des outils de compression, tels que gzip ou bzip2 (Cilibrasi et Vitányi, 2007). Ces outils permettent de reproduire la hiérarchie des options dans l'histoire de St Chéron (en moyennant sur 1000 mesures pour compenser les limitations des compresseurs pour les chaînes courtes). Bien que les valeurs obtenues sous-estiment l'inattendu, de tels outils peuvent être utiles quand la complexité doit être évaluée par un système automatisé.

La *complexité causale* $C(H)$ est la complexité du scénario le plus simple qui permet de produire la situation dans le monde considéré. Elle est équivalente à la parcimonie d'explication (Feldman, 2004 ; Chaitin, 2004). La complexité d'un scénario causal dépend de la capacité du sujet d'imaginer des causes plausibles par abduction. La complexité de génération $C_w(s)$ est alors calculée de manière récursive à partir de la complexité de génération de ses causes (Dessalles, 2008b). Comme modéliser l'abduction se révèle problématique dès que le sens commun est impliqué (Magnani, 2001), nous sommes à la recherche d'une méthode indirecte. Une technique prometteuse consiste à employer une distance informationnelle telle que la distance normalisée Google (NGD) (Cilibrasi et Vitányi, 2007). Le Web contient la trace textuelle d'innombrables événements non inattendus. La co-occurrence statistique des termes décrivant un événement *s* doit donc être négativement corrélée avec la complexité $C_w(s)$. Dans l'histoire de la gifle, il est censé être plus difficile de produire l'occurrence d'un homme qui en gifle un autre dans la rue qu'un événement dans lequel l'homme demande simplement l'heure. Une application directe de la formule NGD donne une distance (normalisée entre 0 et 1) entre « rue » et « gifler » de 0.74 qui est sensiblement plus élevée que la distance entre « rue » et « demander » (0.15). Par comparaison, « rue » est proche de « bâtiment » (0.08) et éloignée d'un concept abstrait comme « configurer » (0.84). Ces distances pourraient offrir des estimations fiables de la complexité (Cilibrasi et Vitányi, 2007), bien qu'il faille encore travailler pour rendre ce type de calcul plus robuste en fonction du choix du moteur de recherche et des corpus de textes utilisés (Lindsey *et al.*, 2007).

5.2 Résultats inattendus

La plupart des écarts entre les résultats et les prévisions peuvent être attribués à deux

causes. L'une vient du caractère artificiel de la situation expérimentale, qui diffère de la situation idéale dans laquelle les sujets ne sauraient pas qu'ils sont testés. Par exemple, il est étonnant que trois participants jugent plus intéressant que l'offre promotionnelle de l'opérateur arrive deux semaines après l'achat du mobile, plutôt que juste dix minutes après. La consigne demandait de choisir « l'alternative qui rend l'histoire aussi intéressante que possible », en proposant une récompense (une clé USB) pour le participant qui ferait les choix considérés comme intéressants par les autres participants. Il est possible que certains participants aient pu mal interpréter la consigne.

L'autre source d'écart par rapport aux prédictions vient du caractère peu vraisemblable de certaines options, parfois mentionné de manière explicite dans les commentaires que nous avons recueillis. Par exemple :

`Pour la deuxième fois je coche un truc moins intéressant mais plus crédible.`

ou encore :

`la réponse la plus "choc" n'est pas forcément celle qui rend l'auditoire le plus attentif - ça en fait trop.`

Ce phénomène a certainement joué dans l'histoire du double suicide, dans laquelle le tatouage le plus détaillé n'a pas été retenu. Pour l'histoire de la saisie de drogue, nous avons eu le commentaire suivant :

`5 kilos, c'est une affaire internationale. 2 kilos, peut-être les gens n'en auront pas entendu parler quand je leur raconterai. Donc ca m'intéresse plus.`

Certaines réponses peuvent résulter d'erreurs ou de réflexions complexes, comme le révèle le commentaire suivant à propos de l'histoire de la gifle, pour laquelle le participant a préféré « demander l'heure » :

`C'est intéressant parce qu'on se demande ce qui pousse la personne à raconter ca. Ce type doit être vraiment spécial pour que cet événement soit marquant. Mais qu'est-ce qu'il a de spécial? On veut en savoir plus.`

La seule vraie surprise de l'expérience est venue de l'histoire suivante.

> J'ai dû payer 400 euros une amende vieille d'un an à cause du trésor public qui n'a pas envoyé les rappels à la bonne adresse. Au moment du prélèvement, j'avais exactement [...] euros sur mon compte.
>
> a: 400
> b: 401
> c: 419

Préférences		Rangs
a-b	13/20	a 33
b-c*	29/5	b 56
a-c*	24/9	c 11

Les participants ont préféré 401 à 400 euros, alors que CDT prévoit *a priori* que 400 est plus simple. Une explication possible est que 400 évoque un nombre arrondi, et donc une classe de situations, tandis que 401 est à la fois simple et unique (comme pour 1001 nuits *vs.* 1000 nuits). Nous projetons de tester cette hypothèse plus avant.

6 Perspectives

On peut penser à une variété d'explications séparées pour les différents types d'histoire : une explication pour des analogies, une pour la proximité, une pour la structure, et ainsi de suite. L'expérience décrite en cet article conforte l'idée inverse qu'un principe cognitif unique et simple est à l'œuvre : la baisse de complexité permet de reproduire de manière correcte toutes les prédictions sans recourir à des hypothèses *ad hoc*. En particulier, le sentiment d'improbabilité qui est souvent mentionné en lisant les histoires de notre test n'est pas une dimension séparée d'intérêt. La probabilité subjective p peut être déduite de l'inattendu U (Dessalles, 2006 ; 2008b) par la formule :

$$p = 2^{-U}$$

La difficulté principale avec notre approche expérimentale de l'intérêt narratif est que les histoires dans ce type de test se situent à mi-chemin entre la fiction et la non-fiction. Nous travaillons actuellement à une nouvelle conception de l'expérience dans laquelle les

jugements d'intérêt seront plus conformes à ceux d'une mise en situation réelle. Une autre perspective est de rendre des calculs de complexité automatiques, en utilisant des distances issues de mesures de co-occurrence sur le Web. Les applications pratiques de ces recherches sont diverses, et vont du calcul de l'intérêt médiatique (*newsworthiness*) aux moteurs de recherche événementiels.

Références

Chaitin, G. J. (2004). On the intelligibility of the universe and the notions of simplicity, complexity and irreducibility. In Hogrebe & Bromand (Eds.), *Grenzen und Grenzüberschreitungen, XIX*, 517-534. Berlin: Akademie Verlag.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle?. *The Quaterly Journal of Experimental Psychology, 52* (A), 273-302.

Chater, N. & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science?. *Trends in cognitive sciences, 7* (1), 19-22.

Cilibrasi, R. & Vitányi, P. (2007). The Google similarity distance. *ACM Transactions on Knowledge and Data Engineering*. http://www.cwi.nl/~paulv/papers/amdug.pdf

Cornuéjols, A. (1996). Analogie, principe d'économie et complexité algorithmique. In *Actes des 11èmes Journées Françaises de l'Apprentissage*. Sète. http://www.lri.fr/~antoine/Papers/JFA96-final-osX.pdf

Delahaye, J-P. & Zenil, H. (2007). On the Kolmogorov-Chaitin Complexity for short sequences. In C.S. Calude (Ed.), *Randomness and Complexity: From Leibniz to Chaitin*. Singapore: World Scientific. http://arxiv.org/abs/0704.1043

Dessalles, J-L. (2006). A structural model of intuitive probability. In D. Fum, F. Del Missier & A. Stocco (Eds.), *Proceedings of the seventh International Conference on Cognitive Modeling*, 86-91. Trieste, IT: Edizioni Goliardiche. http://www.dessalles.fr/papiers/pap.cogni/Dessalles_06020601.pdf

Dessalles, J-L. (2008a). Coincidences and the encounter problem: A formal account. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2134-2139. Austin, TX: Cognitive Science Society. http://www.dessalles.fr/papiers/pap.conv/Dessalles_08020201.pdf

Dessalles, J-L. (2008b). *La pertinence et ses origines cognitives - Nouvelles théories.* Paris: Hermes-Science Publications. http://pertinence.dessalles.fr

Eggins, S. & Slade, D. (1997). *Analysing casual conversation.* London: Equinox.

Farkas, A. (2007). The analysis of the principal eigenvector of pairwise comparison matrices. *Acta Polytechnica Hungarica, 4* (2). http://bmf.hu/journal/Farkas_10.pdf

Feldman, J. (2004). How surprising is a simple pattern? Quantifying 'Eureka!'. *Cognition, 93*, 199-224.

Galtung, J. & Ruge, M. H. (1965). The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four foreign newspapers. *Journal of International Peace Research, 2* (1), 64-91.

Labov, W. (1997). Some further steps in narrative analysis. *Journal of Narrative and Life History, 7* (1-4), 395-415. http://www.ling.upenn.edu/~wlabov/sfs.html

Labov, W. & Waletzky, J. (1967). Narrative analysis: Oral versions of personal experience. In J. Helm (Ed.), *Essays on the verbal and visual arts*, 12-44. Seattle, WA: University of Washington Press. http://www.clarku.edu/~mbamberg/LabovWaletzky.htm

Lindsey, R., Veksler, V. D., Grintsvayg, A. & Gray, W. D. (2007). Be wary of what your computer reads: The effects of corpus selection on measuring semantic relatedness. In *Proceedings of the 8th International Conference on Cognitive Modeling*. Ann Arbor, MI. http://www.cogsci.rpi.edu/cogworks/publications/271_ICCM_final.pdf

Li, M. & Vitányi, P. (1993). *An introduction to Kolmogorov complexity and its applications.* New York: Springer Verlag, ed. 1997.

Magnani, L. (2001). *Abduction, reason and science - Processes of discovery and explanation.* New York: Kluwer Academic.

Mehl, M. R. & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, *84* (4), 857-870.

Norrick, N. R. (2000). *Conversational narrative: storytelling in everyday talk.* Amsterdam: John Benjamins Publishing Company.

Rimé, B. (2005). *Le partage social des émotions.* Paris: PUF.

Saaty, T. L. (1994). How to make a decision: The analytic hierarchy process. *Interfaces*, *24* (6), 19-43.

Shapiro, M. A. & Fox, J. R. (2002). The role of typical and atypical events in story memory. *Human Communication Research*, *28* (1), 109-135.

Solomonoff, R. J. (1978). Complexity-based induction systems: Comparisons and convergence theorems. *IEEE transactions on Information Theory*, *24* (4), 422-432. http://world.std.com/~rjs/solo1.pdf

Stangor, C. & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: a review of the social and social developmental literatures. *Psychological Bulletin*, *111* (1), 42-61.

Tannen, D. (1984). *Conversational style - Analyzing talk among friends.* Norwood: Ablex Publishing Corporation.

Woll, S. B. & Graesser, A. C. (1982). Memory discrimination for information typical and atypical of person schemata. *Social Cognition*, *1*, 287-310.