Le Codage des données textuelles

Paul A. Fortier

Department of French, Spanish and Italian — 231 St. Paul's College —
University of Manitoba — Winnipeg, Man. R3T 2N2 — CANADA

ABSTRACT

Virtually all textual data analysis, whether of literature or of other documents like answers to open-ended questions, begins with the identification of semantic categories in the text. This paper reports results of a large-scale experiment looking into the reliability of the coding process, by asking ten different people to categorize 1830 short passages from literary texts as to whether or not each passage contains an allusion to human solitude. The results indicate that a large number of coders is required if one wants to have a reliable level of consistency in the coding.

RÉSUMÉ

La plupart des analyses statistiques des données textuelles, qu' il s' agisse d' oeuvres littéraires ou de réponses à des questions ouvertes, débute par l' identification des catégories sémantiques qui se trouvent dans le texte. Cette communication présente les résultats de deux expériences visant à définir à quel point il est raisonnable de se fier à ce que produit ce processus de codage. Une dixaine de personnes ont codé 1830 courts extraits de textes littéraires, indiquant si chaque extrait contient une évocation de la solitude humaine, ou bien toute autre chose. Les résultats obtenus démontrent qu' il faut prévoir un très grand nombre de codeurs pour obtenir un codage avec une consistance statistiquement significative.

MOTS CLÉS: Codage, Sémantique, Données, Statistique

1. Le Problème

Qu' il s' agisse de textes littéraires ou de réponses aux questions ouvertes, l' analyse des données textuelles comporte une étape de codage. Il est nécessaire d' identifier la présence et la fréquence de certains sujets, thèmes ou idées dans le texte avant de procéder à des analyses plus avancées. Autrement dit une désambiguation sémantique précède et prépare tout traitement statistique des données textuelles.

Il est, bien entendu, possible de repérer par ordinateur le vocabulaire susceptible d' évoquer une certaine idée, mais le résultat n' est pas toujours immédiatement utilisable, car, étant donné le phénomène linguistique de la polysémie, il y a une différence inconnue entre la fréquence du vocabulaire potentiel de tel concept et les évocations réelles du même concept. Par exemple, "il m' a *abandonnée*" suggère la solitude, alors que "je me suis *abandonnée*" désigne tout autre chose. Un codage humain semble être la solution à ce problème. Il est raisonnable, pourtant, de vouloir définir à quel point on peut se fier aux résultats d' un codage humain.

Les spécialistes d'études littéraires qui se sont penchés sur ce problème (Singh, 1986; Fokkema, 1988) l' examinent d' une perspective théorique et spéculative, sans mettre en jeu des

méthodes empiriques. Les psychologues, surtout les spécialistes de pathologies langagières, fondent leurs analyses sur des observations et évaluations humaines de performance linguistique; Scarsellone (1998) note l' absence d' attention au problème de la confiance qu' on peut attacher à ces jugements, ce qui fait contraste avec les évaluations faites par les spécialistes de pédagogie. Un examen des numéros récents de *Computers and the Humanities* aussi bien que de *Literary and Linguistic Computing* ne révèle aucune étude de la confiance qu' on peut attacher aux résultats d' un codage humain.

D' une perspective ces manques n' étonnent pas outre mesure. La signification de tel élément dans la chaîne linguistique relève du domaine culturel, où les concepts de vérité et de fausseté s' appliquent difficilement, où il s' agit de degré, de jugement, et assez souvent de flair. Dépourvu de la capacité d' identifier la vérité ou la fausseté de tel codage, le chercheur doit se limiter au concept beaucoup plus humble de la consistance. Dans le contexte, il est raisonnable d' évaluer un codage de la perspective de son degré d' accord ou de désaccord avec un codage des mêmes données fait par au autre codeur. C' est la perspective choisie pour cette étude.

2. Données

Le vocabulaire de la solitude constitue le point de focalisation de cette étude. Ce thème est important pour des études psychologiques et sociologiques, puisqu' il indique une intégration imparfaite d' un locuteur à son milieu. C' est aussi un thème qui paraît souvent dans les oeuvres littéraires. Les dictionnaires des synonymes indiquent environ soixante lemmes susceptibles d' évoquer la solitude, comme "seul, solitude, abandonner, isolement". Ces soixante lemmes se réduisent à une trentaine de chaînes de caractères, du type "seul.*", utilisables pour un recensement des textes choisis par l' intermédiaire d' un logiciel.

Le corpus choisi pour cette expérience se constitue de neuf romans du 20e siècle (Bernanos, 1935; Camus, 1942, 1956; Céline, 1932; Gide, 1902, 1909; Mauriac, 1932, Proust, 1925; Sartre, 1938). Le logiciel utilisé a identifié la présence 1830 occurrences du vocabulaire potentielle de la solitude dans l' ensemble des neuf textes. Les fréquences pour chaque texte se trouvent dans la colonne "*brut*" de la Table 1.

Dans un premier temps, le codage de ces données brutes s' est fait à partir de contextes courts: environ dix mots, ou 60 caractères. Deux professeurs de littérature française et quatre étudiants de troisième cycle en ce domaine jouaient le rôle de codeurs. Les instructions fournies étaient simples: marquer les mots qui d' après leur contexte ne semblent pas évoquer la solitude humaine. Chaque codeur avait donc à faire 1830 décisions d' après sa lecture du contexte et sa connaissance générale de la langue. La Table 1, côté gauche, résume les résultats de ce codage.

Le lecteur notera la grande disparité entre les résultats. En beaucoup de cas la valeur maximum est plus que le double de la valeur minimum, et pour deux des textes la fréquence maximum est plus de trois fois plus grande que la minimum. Transformés en statistique *t*, les écarts-type sont tous signficatifs au niveau $p = 0,01$. Il semblerait que l' absence de contexte suffisant indroduise trop d' incertitude dans l' esprit des codeurs, incertitude qui se traduit par des résultats peu consistants.

Afin de résoudre ce problème, une deuxième expérience, qui s' est déroulée dix-huit mois plus tard, présentait le même vocabulaire centré dans trois lignes (environ 300 caractères) de contexte. Les codeurs étaient trois professeurs dont un avait participé à la première expérience

et trois étudiants de troisième cycle dont un avait participé à la première expérience. Les instructions données aux codeurs étaient identiques à celles de la première expérience. Un sommaire des résultats de ce deuxième codage se trouve dans la Table 2, côté droit.

Dans tous les cas, que la moyenne augmente ou diminue, la différence entre la fréquence minimum et la fréquence maximum accroît. Les écarts-type sont maintenant signfiicatifs au niveau p = 0,001, lorsqu' ils sont convertis en statistique *t*. Ce phénomène semble digne d' analyse plus avancée, ce qui sera faite en fonction des 21.960 décisions individuelles qui sont résumées dans la Table 1.

Texte			Contexte: 60 Caractères 6 Codeurs					Contexte: 300 Caractères 6 Codeurs			
	brut		Min	Moy	Max	E.T.		Min	Moy	Max	E.T.
BJC	260		44	74	100	23,04		44	78	154	42,83
CEt	74		11	19	26	5,90		6	19	48	15,46
CCh	115		26	35	41	5,96		23	35	61	13,64
CVN	488		41	104	153	41,57		53	102	199	59,85
GIm	89		18	37	48	11,73		13	30	49	13,30
GPE	108		26	37	54	10,46		10	34	69	19,47
MNV	159		28	58	92	23,43		28	56	113	33,40
PFu	304		44	65	82	16,71		44	68	126	30,90
SNa	233		77	101	122	17,44		67	96	151	32,11

Sigles:

brut: Le nombre d'évocations potentielles de la solitude trouvées par le logiciel.
Min: Le plus petit nombre d'évocations de la solitude humaine enregistré par l'un des codeurs.
Moy: La moyenne des fréquences enregistrées par les six codeurs.
Max: Le plus grand nombre d'évocations de la solitude humaine enregistré par l'un des codeurs.
E.T.: L'écart-type des fréquences enregistrées par les six codeurs.

Textes:

BJC: Bernanos, *Journal d'un curé de campagne*
CEt: Camus, *L'Étranger*
CCh: Camus, *La Chute*
CVN: Céline, *Voyage au bout de la nuit*
GIm: Gide, *L'Immoraliste*
GPE: Gide, *La Porte Étroite*
MNV: Mauriac, *Le Noeud de Vipères*
PFu: Proust, *La Fugitive*
SNa: Sartre, *La Nausée.*

Table 1: La Solitude humaine

3. Méthode

Les chiffres de la Table 1 totalisent les décisions faites par un codeur devant un texte donné. Ce processus occulte les décisions individuelles. Par exemple, le fait qu' un codeur a trouvé 36 évocations de la solitude humaine dans *La Porte Étroite* d' André Gide en examinant 60 caractères de contexte, et qu' un autre en a trouvé 37 pour le même texte avec le même contexte, n' indique nullement que les deux codeurs soient d' accord sur 36 des cas, avec un seul point de différence. Cela est possible, mais il est tout aussi possible qu' il n' ait aucun cas où les deux codeurs marquent le même mot comme évocateur de la solitude humaine. C' est pour cela qu' il est nécessaire d' analyser les 21.960 choix individuels.

Par contraste avec les études de pédagogie, il n' est pas possible de désigner tel choix comme vrai ou comme faux. Semblable en cela à beaucoup de cas qui se rencontrent dans les sciences sociales, il est possible de mesurer à quel degré des series de décisions s' accordent, sans rien pouvoir conclure absolument sur la vérité des décisions prises (Shrout, 1995). Dans un contexte culturel ou linguistique, comme celui de cette étude, il est pourtant raisonnable conclure que la consistance indique fort probablement une bonne décision, alors que moins de consistance crée le doute. Il est donc raisonnable de mesurer en quelle proportion les codeurs font des décisions identiques, et à quel degré leurs décisions divergent. Une telle analyse doit nécessairement prendre en compte la possibilité que pour telle décision deux codeurs sont d' accord par hasard.

Il s' agit pour cette analyse de comparer douze séries de 1830 décisions. Le coefficient de corrélation de Pearson, méthode classique et disponible partout, semblerait toute désignée pour cette sorte de comparaison. Cette méthode, pourtant, ne prend pas en compte la possibilité qu' une certaine proportion des choix des codeurs soient identiques par l' action du pur hasard (voir Kuder et Richardson 1937, Hoyt 1941 et Cronbach 1951). Le coefficient de corrélation entre classes (Cohen 1960, Shrout et Fleiss 1979) montre le degré de corrélation entre une série de décisions, après avoir corrigé pour les cas d' accord résultant du pur hasard. Le CCEC, s' interprète toutefois comme le coefficient de Pearson: 1,0 indiquant un accord parfait, et 0,0 correspondant à une absence totale de relation entre les variables. Également comme le coefficient de Pearson, il est souvent utile de présenter les résultats non pas par une valeur unique, mais sous forme d' intervalle de confiance soit à ±5% ou même ±1%.

À partir du CCEC, il est possible de dériver le nombre de codeurs nécessaire pour produire une consistance à tel degré, puisqu' un nombre infini de codeurs produirait, paradoxalement, un coefficient de 1,0 (Feldt, 1965). Il semble raisonnable de déterminer pour chaque texte le nombre de codeurs nécessaires pour produire un CCEC de 0,95 dix-neuf fois sur vingt.

4. Résultats

La Table 2 montre les résultats de l' application de la formule CCEC aux données et les présente en fonction d' intervalles de confiance au niveau de±5%. Les résultats pour un contexte de 60 caractères se trouvent à gauche, et ceux pour un contexte de 300 caractère à droite. Les valeurs pour 300 caractères sont systématiquement plus petites que les coefficients pour 60 caractères. Dans le cas de trois romans–Bernanos, *Journal d'un curé de campagne,* Céline, *Voyage au bout de la nuit,* et Proust, *La Fugitive*–les intervalles de confiance ne se recouvrent en aucun point. Manifestement, comme le contexte s' étend, le codage devient

moins consistant. Il est fort probable que cette divergence résulte du fait que le contexte accru fournit plus d' occasions pour le jeu des opinions et valeurs personnelles.

Texte	60 Caractères		300 Caractères	
	-5%	+5%	-5%	+5%
BJC	0,56	0,66	0,43	0,54
CEt	0,48	0,68	0,29	0,51
CCh	0,68	0,79	0,54	0,69
CVN	0,51	0,59	0,42	0,50
GIm	0,40	0,59	0,28	0,48
GPE	0,43	0,60	0,33	0,50
MNV	0,47	0,60	0,33	0,50
PFu	0,63	0,71	0,49	0,59
SNa	0,59	0,69	0,48	0,59

Sigles: Voir la Table 1

Table 2: Le CCEC: Intervalles de confiance à 95%

Lorsque les intervalles de confiance sont convertis en nombre de codeurs nécessaires pour produire un CCEC de 0,95 dix-neuf fois sur vingt, les valeurs présentées dans la Table 3 se produisent. Comme l' on pourrait prévoir, il faudrait plus de codeurs pour obtentir une consistance raisonnable lorsque le contexte est plus étoffé. Notable est le fait que les textes réputés les plus difficiles dans les milieux d' études littéraires, comme *La Chute* de Camus et *La Fugitive* de Proust exigent les plus petits nombres de codeurs, alors que ceux dont les spécialistes de littérature louent la clarté stylistique, comme *L'Étranger* de Camus et les romans de Gide ont besoin de beaucoup plus de codeurs pour arriver au même niveau de certitude.

Texte	60 Caractères	300 Caractères
BJC	15	25
CEt	19	43
CCh	9	16
CVN	18	26
GIm	28	45
GPE	25	37
MNV	21	25
PFu	12	19
SNa	14	21

Sigles: Voir la Table 1.

Table 3: Le Nombre de codeurs nécessaires pour produire un CCEC de 0,95 19 fois sur 20.

Des analyses subséquentes, qui dépassent le cadre de cette communication démontrent que ces résultats de découlent de l' influence ni d' un seul codeur, ni de la langue maternelle ou niveau d' instruction des codeurs, ni d' un mot donné, ni d' une quelconque classe de fréquences des mots codés. Il est nécessaire d' accepter le fait que le nombre de codeurs joue un rôle prépondérant dans la consistance du codage, et donc dans la confiance qu' il serait légitime d' y mettre.

5. Conclusion

Ces expériences n' examinent pas la vérité des décisions prises par les codeurs. Elles se limitent au concept beaucoup plus humble de la consistance. Les résultats obtenus démontrent un fait inéluctable: de différentes personnes codent le contenu des texte de façon tellement peu semblable, qu' un seul ou même plusieurs codages produisent des résultats fortement sujets à caution. D' ailleurs, dans la plupart des cas, une dizaine de codeurs ne suffisent pas pour garantir un accord statistiquement significatif entre les résultats. Cela explique la pléthore de jugements contradictoires sur la valeur, et même sur le contenu, des textes littéraires qui caractérise le domaine des études littéraires. En fin de compte, chaque critique, consciemment ou non, appuie ses évaluations sur des jugements multiples concernant la signification des éléments sémantiques du texte étudié.

Pour nous autres que font des analyses statistiques des données textuelles, la leçon est un peu différente. Du moment qu' on dépasse la trivialité le plus banal, comme le nombre de mots entre deux signes de ponctuation, nos données sont codées. Les procédures qu' on utilise pour les analyser comportent chacune sa marge d' erreur. Il est important de rappeler qu' il s' agit là de l' imprécision inhérente à la méthode même. La marge d' erreur se produit quand les données sont parfaitment exactes. Lorsqu' on travaille sur des données codées, une source d' erreur substantielle provient du processus de codage lui-même, et risque de noyer toute possibilité de certitude dans une marée d' erreur présente dans les données de base et renforcée par l' action des procédés statistiques. Tout au moins devrait-on prendre en compte la fragilité des données en faisant le choix d' une méthode d' analyse statistique des données codées.

Nécessaire au processus d' analyse des données textuelles, le codage humain comporte des difficultés qu' il serait prudent de ne pas négliger.

Références:

Bernanos, G. (1935). *Journal d'un curé de campagne.* In Béguin, A. editor, *Oeuvres Romanesques.* Bibliothèque de la Pléiade. Paris: Gallimard, 1961.

Camus, A. (1942). *L'Étranger.* In Quilliot, R. editor, *Théâtre, Récits, Nouvelles.* Bibliothèque de la Pléiade. Paris: Gallimard, 1962.

Camus, A. (1956). *La Chute.* In Quilliot, R. editor, *Théâtre, Récits, Nouvelles.* Bibliothèque de la Pléiade. Paris: Gallimard, 1962.

Céline, L.-F. (1932). *Voyage au bout de la nuit.* In Mondor, H. editor, *Romans.* Bibliothèque de la Pléiade. Paris: Gallimard, 1962.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, vol.20: 37-46.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, vol.16: 297-334.

Feldt, L. S. (1965). The approximate simpling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika,* vol.30: 357-370.

Fokkema, D. (1988). On the reliability of literary studies. *Poetics Today,* vol.9: 529-543.

Gide, A. (1902). *L'Immoraliste.* In Davet Y, and Thierry, J.-J., editors, *Romans, Récits, Soties, Oeuvres lyriques.* Bibliothèque de la Pléiade. Paris: Gallimard, 1958.

Gide, A.(1909). *La Porte Étroite.* In Davet Y, and Thierry, J.-J., editors, *Romans, Récits, Soties, Oeuvres lyriques.* Bibliothèque de la Pléiade. Paris: Gallimard, 1958.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, vol.*6: 153-160.

Kuder, G. F., and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika,* vol.2: 151-160.

Mauriac, F. (1932). *Le Noeud de Vipères.* Paris: Grasset.

Proust, M. (1925). *La Fugitive.* In Clarac, P. and Ferré, A. editors, *À la recherche du temps perdu.* Bibliothèque de la Pléiade. Paris: Gallimard, 1954.

Sartre, J.-P. (1938). *La Nausée.* In Contat, M. and Rybalka, M. editors, *Oeuvres Romanesques.* Bibliothèque de la Pléiade. Paris: Gallimard, 1981.

Scarsellone, J. M. (1998). Analysis of observation data in speech and language research using generalizability theory. *Journal of Speech, Language, and Hearing,* vol.41: 1341-1347.

Shrout, P. E., and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin,* vol.86: 420-428.

Shrout, P. E. (1995) Measuring the degree of consensus in personality judgments. In Shrout, P. E. and S. T. Fiske editors *Personality research, methods and theory: A festchrift honoring Donald W. Fiske.* Hillsdale, N.J.: Lawrence Erlbaum Associates, pages 79-92.

Singh, G. (1986). The identity of the literary text: Problems and reliability of reader response: An inquiry into theoretical positions. *The Literary Criterion,* vol.21: 49-57.