

Université d'été de Carcassonne sur l'« Analyse de Données Textuelles – Méthode ALCESTE »

Cette année aura lieu la vingtième édition de l'Université d'été de Carcassonne sur l'« Analyse de Données Textuelles – Méthode ALCESTE » du 19 au 21 Août 2009. A cette occasion, nous réserverons la journée du 21 pour une journée d'étude et de discussion sur le contenu de cette méthodologie pour chacun, l'objectif étant un éclaircissement des principes de base sur laquelle elle repose. Afin de faciliter la discussion, nous vous joignons un argumentaire.

Pour le comité d'organisation,

Max Reinert (max.reinert@wanadoo.fr)

L'argument de la journée :

La simplicité des principes à la base de la méthode ALCESTE s'accompagne d'une complexité de leur mise en œuvre dans des applications réelles. L'objet de cette journée serait déjà de s'entendre sur le contenu de ces principes afin de démêler ce qui se joue dans chaque pratique entre analyse de contenu et analyse de discours, de ce qui se joue aussi au plan expérimental... A cet égard on ne peut oublier que la conception de la méthode a été interdépendante de la conception du logiciel, l'algorithme informatisé ayant évolué avec les différents environnements de travail de son auteur. En fait, il s'agirait plutôt d'essayer d'extraire de manière explicite l'essentiel des principes de cette méthodologie afin de pouvoir en transmettre les principes (hors d'une pratique informatisée) pour les discuter, les compléter, voire les réinterpréter dans des pratiques nouvelles...

En voici une première version sous forme de questions qui servira d'argumentaire :

1. Dans quelle mesure l'analyse statistique exploratoire est-elle une aide à la lecture de l'analyste ? Quel est l'objet d'une analyse ?

Le corpus textuel soumis à l'analyse tient sa légitimité non pas en tant qu'il serait construit en fonction d'un plan expérimental, mais par le seul fait qu'il constitue, pour cet

analyste, un discours porteur de " contenu ". Bien entendu le contenu ne définit pas l'objet d'étude, mais il introduit une dynamique de recherche pour lui donner forme. A ce titre, cette approche ne saurait être autre qu'exploratoire en offrant un cadre indépendant de tout a priori conceptuel sur les conditions de production du corpus pour élaborer une lecture..

2. Quelle signification donner au découpage du texte en unités de contexte ?

On sait que ce découpage du texte en unités de contexte de "longueur" comparable et relativement courte par rapport à la " longueur " du corpus est un des points les plus controversés de la méthode (J. Jenny, BMS, n°57). Reinert le situe conceptuellement comme essentiel en ce sens que, sans aucun a priori sur le sens du texte étudié, ce découpage permet d'évaluer un jeu contextuel pour définir la cooccurrence.. avec l'hypothèse forte qu'une stabilisation des résultats est généralement soutenable au moins pour un ordre de grandeur donné de l'unité de contexte. Cette hypothèse forte est confortée depuis 1990 par la conception d'une double analyse obtenue par variation systématique de la longueur des unités de contexte (Reinert, BMS 1990) et, aujourd'hui, par la mise en évidence de « mondes lexicaux stabilisés » (Reinert, JADT 2008).

3. Quelle signification donnée au choix des mots pleins pour le calcul statistique des classes ?

Ce choix, également controversé, n'a jamais varié, alors même que la frontière entre mots pleins et mots outils a été bien fluctuante (cas de certains adverbes, des locutions prépositionnelles, de certains figements, etc.). De plus ce critère impose logiquement une lemmatisation du texte, mais il s'est avéré que cette lemmatisation peut être assez souvent omise en pratique..

4. Pourquoi une analyse statistique de type descendant ?

Rappelons que la classification descendante hiérarchique (C.D.H.) mise en œuvre dans l'algorithme d'ALCESTE est une méthode d'analyse statistique proche de l'analyse factorielle des correspondances de J.P. Benzécri dont elle utilise l'algorithme à chaque pas (Reinert, 1979, 1983). Dans les deux cas, l'aspect descendant de la méthode implique que l'analyse porte d'abord sur la forme globale de l'ensemble des distributions plutôt que sur une analyse précise des

liens deux à deux. Cette forme globale des répétitions est perceptible statistiquement à travers la notion de valeur propre de la matrice de covariance et de facteurs propres (ici pour la métrique du χ^2). Dès les premières analyses factorielles des correspondances que J. P. Benzécri a introduites au cours des années 1960-1970, les propriétés synthétisantes des représentations des premiers plans factoriels ont été mis en évidence (voir par exemple, l'analyse du graphe de la carte des départements français proposée par L. Lebart in Benzécri & col, « l'analyse des données », tome 2, 1973, p244-252). Reinert considère la classification descendante hiérarchique comme une adaptation de l'analyse factorielle des correspondances (simple) pour le traitement des grands tableaux binaires clairsemés du type « unités de contextes par mots pleins »...

La réponse à ces quatre questions nous semblent se référer à ce qui constitue le cœur de « la méthodologie ALCESTE ». Voilà ce que je souhaite soumettre à la discussion et à votre expérience propre des analyses statistiques textuelles, en fonction de vos objectifs et de vos pratiques, pour une discussion franche et amicale, afin d'éclaircir au mieux le rôle des principes de base dans une interrogation sur le sens des résultats.