Alceste, un logiciel d'analyse textuelle

Valérie Delavigne

Alceste, un logiciel d'analyse textuelle

Il semble difficile aujourd'hui de négliger l'apport de logiciels d'analyse textuelle extrêmement performants. Sans être une panacée, ces outils constituent néanmoins une assistance très efficace et garantissent une systématicité d'analyse. Ils permettent d'objectiver les intuitions que l'on peut avoir vis à vis d'un corpus tout en mettant en évidence certains aspects qu'une analyse manuelle ne relèverait sans doute pas. C'est une aide à une lecture non séquentielle des textes, utile pour différents types d'analyses.

Les travaux de Harris sur l'analyse distributionnelle ont contribué à développer en France dans les années 60 diverses réflexions d'analyse textuelle. C'est au sein d'une école originale d'analyse de données textuelles, née des travaux du mathématicien Jean-Paul Benzécri, que Max Reinert a conçu le logiciel d'analyse Alceste (Analyse lexicale par Contexte d'un Ensemble de Segments de Texte). Alceste est orienté vers l'analyse de contenu. Il éclaire sur des faits statistiques du corpus, qui peut être de diverse nature (textes littéraires, articles de journaux, entretiens…), et offre des pistes interprétatives.

LES PRINCIPES D'ALCESTE

Généralement, les logiciels d'analyse statistique partent des mots et, en recherchant leurs cooccurrents, forment des classes. Ils procèdent ainsi à une « classification hiérarchique ascendante ». Inversant la démarche, Alceste utilise une méthode de classification originale, une méthode de « classification hiérarchique descendante » : le logiciel fractionne de façon successive le texte et en extrait des classes de mots représentatives.

Le logiciel est fondé sur une analyse statistique distributionnelle. Les mécanismes qu'il met en œuvre sont indépendants du sens : Alceste classe les « phrases »[1] du corpus (dénommées « unités de contexte » ou « u.c. »), en fonction de la distribution du vocabulaire présent dans ces unités de contexte. Le logiciel repère le vocabulaire dans les différentes unités de contexte et les met en relation. Autrement dit, il *relie les contextes qui ont des mots communs*.

Alceste permet d'effectuer une analyse sur un volume important de documents numérisés. Inversement, afin que les résultats statistiques gardent leur pertinence, le corpus soumis à Alceste doit être suffisamment volumineux.

Le logiciel réclame, afin d'être optimisé, que le corpus qu'on lui soumet présente une certaine cohérence thématique : dans la mesure où le principe du logiciel est de comparer les distributions des vocables au sein d'un corpus, encore faut-il que les différentes parties du corpus aient des mots en commun…

[1] Qui ne correspondent pas forcément à des phrases, tout particulièrement pour des transcriptions d'oral dont la syntaxe na peu de choses à voir avec celle de l'écrit.

Comment Alceste procède-t-il ? A partir d'un corpus mis en forme, le logiciel découpe le texte en unités de contexte, puis il reconnaît les formes[2] dans ces unités de contexte. Deux types d'unités de contextes sont à distinguer.

- Les plus grandes unités de contexte sont des parties de textes du corpus, dites « unités de contexte initiales » ou « u.c.i. », auxquelles l'analyste affecte des variables. Ce peut être n'importe quel découpage qui convienne à l'analyste : différents articles de journaux par exemple, des entretiens, différentes parties du corpus… Les variables attribuées à ces segments de texte permettent ensuite de les croiser. Ces variables sont notées à l'aide du signe « étoile » qui signale au logiciel que ces mots sont à considérer hors corpus.

Ce découpage en unités de contexte initiales est facultatif : on peut très bien traiter un corpus sans procéder à cette signalisation.

- Le deuxième découpage du corpus est effectué par Alceste qui définit les « phrases » à partir desquelles le logiciel va procéder à l'analyse (les « unités de contexte élémentaire » ou « u.c.e. »). Cette segmentation se fonde éventuellement sur la ponctuation, si elle existe, puis sur le nombre de mots.

Un autre type d'unité nécessaire au logiciel concerne l'unité lexicale. Le logiciel identifie les occurrences de chaque forme grâce à un dictionnaire. Les « mots outils » et les locutions sont ainsi distingués des « mots pleins ». Alceste regroupe l'ensemble des formes susceptibles d'appartenir à une même famille morphologique quelle que soit leur catégorie syntaxique. Il classe ainsi sous la même forme *présent+* : *présenter, présent, présentation, présentable, présentée, présentait*, etc. Le logiciel établit ainsi un dictionnaire des « formes réduites » du corpus à partir de la racine des mots.

A partir de cette partition des unités de contexte et des formes, le corpus est modélisé par un tableau de données, qui comprend les fragments en lignes et les mots pleins en colonnes. Alceste effectue ensuite diverses classifications. Il croise les unités de contexte et la présence/absence des formes, puis il forme des classes à partir des unités de contexte qui contiennent les même mots. De façon itérative, le logiciel fait varier le nombre de mots par unités de contexte, compare les classes obtenues et conserve les classes qui sont associées au plus grand nombre d'unités de contexte.

En fin de course, on obtient un certain nombre de classes de mots, représentatives du texte analysé. Le logiciel met ainsi en évidence les principaux « mondes lexicaux » du corpus traité, c'est-à-dire des ensembles de mots plus particulièrement associés à une classe.

LES PHASES DE L'ANALYSE

L'analyse se déroule en quatre phases, subdivisées en plusieurs opérations (le « plan d'analyse »).

[2] Chaque suite de caractère située à l'intérieur de caractères délimiteurs est une occurrence ; les occurrences identiques constituent des formes (cf. Lebart et Salem, 1994 : 36).

L'étape A

Lors de la première étape, Alceste reconnaît les u.c.i. (les unités de contexte choisies par l'analyste) et les mots « étoilés » (les variables). Puis il effectue trois traitements successifs : il découpe le corpus en formes, il procède à une catégorisation syntaxique et à une lemmatisation.

L'étape B

La deuxième étape découpe le corpus en unité de contexte élémentaire (grossièrement, en phrases) et les classe en fonction de leur distribution. Alceste constitue ainsi des classes sur la base du contenu lexical de chaque u.c.e. : il rapproche les u.c.e. contenant les mêmes formes lexicales.

L'étape C

Lors de la troisième étape, les résultats des calculs précédemment effectués sont mis sous forme de fichiers : les classes obtenues, les formes les plus fréquentes de chacune d'elles…

L'étape D

Au cours de l'étape D, des calculs complémentaires sont effectués :

- deux types de tris croisés (une partie du texte est croisée avec une variable ou un mot particulier) ;
- une analyse factorielle des correspondances (croisement du vocabulaire et des classes), visualisable à l'aide de représentations graphiques très parlantes ;
- une classification ascendante hiérarchique qui montre les liens plus ou moins proches que les mots entretiennent entre eux.

Ces analyses constituent une aide à l'interprétation des résultats statistiques et à la description des classes.

QUELS TYPES DE RESULATS OBTIENT-ON GRACE A ALCESTE ?

En précédant à des regroupements de formes, Alceste restitue le corpus en classes qui dégagent des sortes de « vision du monde ». Ces classes de mots doivent ensuite faire l'objet d'une interprétation qui dépend des objectifs de l'analyse.

Chaque classe peut être examinée grâce à un « profil » : pour chacune d'elle, Alceste donne accès à la liste des mots les plus significatifs, aux unités de contexte les plus significatives, aux cotextes caractéristiques des classes et aux concordances. Le Chi^2 ($\chi^{2)}$ permet de déterminer la forte ou la faible appartenance d'un mot à une classe, et de mettre ainsi en évidence les termes les plus représentatifs d'une classe donnée.

Divers « documents lexicométriques » sont obtenus à partir des formes segmentées :

- un dictionnaire des formes analysées ;
- un dictionnaire des formes réduites avec leur affectation à chaque classe et leur distribution ;

- une liste des formes réduites les plus fréquentes ;
- une liste des hapax[3] du corpus ;
- des dendrogrammes, figures qui visualisent les liens que les formes d'une même classe entretiennent entre elles ;
- un profil des classes en fonction des absences significatives ;
- un profil des classes en fonction des présences significatives ;
- une carte correspondant à l'analyse factorielle des correspondances ;
- une liste des segments répétés par fréquence décroissante sur l'ensemble du corpus et par classe…

Ces documents peuvent être obtenus sur le corpus entier ou sur plusieurs fragments du corpus que l'on peut ensuite comparer entre eux. Ces résultats n'ont bien évidemment de valeur qu'en tant qu'éléments de comparaison, soit à l'intérieur d'un même corpus, soit entre corpus comparables.

L'ensemble des documents lexicométriques permet une analyse de contenu du corpus. Cependant, dans la mesure où le logiciel dénombre des formes lexicales et en montre les cooccurrences, ces documents présentent également un intérêt certain pour une analyse de discours et une étude du fonctionnement d'unités particulières. C'est donc une aide informatique utilisable pour une description d'usage des formes linguistiques dans les corpus.

Même si Alceste ne peut suffire à une analyse réellement linguistique, l'avantage des documents lexicométriques obtenus tient au fait, qu'en sortant de la linéarité du texte, le regard porté sur le corpus est différent.

QUELQUES LIMITES DU LOGICIEL

L'exploitation des résultats nécessite de se familiariser avec un certain nombre de méthodes statistiques. Malgré la transparence affirmée par l'auteur, le logiciel reste une « boîte noire » qui, de fait, « donne à voir sa complexité » (Alceste, 1995 : 5)…

De façon générale, chaque étape de construction d'une analyse statistique pose des problèmes spécifiques. Il est essentiel d'avoir un regard informé sur les limites des modèles statistiques mis en œuvre lors des analyses. Il ne faut pas oublier que malgré leur attrait, les analyses informatiques doivent être considérées comme des outils, certes extrêmement fiables et performants quant aux résultats proposés, mais qui ne sont pas neutres sur le plan des méthodes. Celles-ci impliquent une certaine vision du texte, considéré comme un « sac de mots » (Lebart et Salem, 1994 : 146), négligeant sa syntaxe et son organisation propre. Sans remettre en cause le caractère heuristique de l'analyse informatique, il convient donc, dans la mesure du possible, de garder un regard critique sur ces méthodes.

Tout d'abord le continu des discours doit être transformé en discontinu, créer des unités discrètes comptabilisables. Le choix des unités de décompte implique des options quant à la segmentation du texte. Par exemple, la lemmatisation, en regroupant sous une même forme les différentes flexions des occurrences, génère des problèmes d'ambiguïté parfois difficiles à résoudre, mais aussi des non-sens d'un point de vue linguistique : utiliser un terme au singulier peut ne pas être la même chose que de l'utiliser au pluriel.

[3] Mot ou expression qui n'apparaît qu'une seule fois dans un corpus donné.

La lemmatisation a bien sûr une forte incidence sur les décomptes lexicométriques. C'est la raison pour laquelle elle a été l'objet de nombreux débats entre « lemmatiseurs » et « non-lemmatiseurs ». L'option originale prise par Reinert, qui regroupe l'ensemble des formes susceptibles d'appartenir à une même famille morphologique, n'est pas sans poser de problème. Classer sous la même forme *arme+* : *arme, armes, armés, armées, armement, armements,* ou sous *incid+ent* : *incidence, incidences, incident* et *incidents* est bien évidemment contestable.

Ensuite, dans la mesure où Alceste ne prend pas le sens en compte, il ne peut bien évidemment départager les cotextes dans lesquels une forme est utilisée au sens propre et ceux dans lesquels elle prend un sens métaphorique.

Par ailleurs, le traitement auquel procède Alceste est sous-tendu par l'hypothèse selon laquelle que les structures sémantiques sont liées à la distribution des mots dans le texte et que cette distribution est pertinente. C'est une hypothèse forte qui doit être gardée à l'esprit.

Comme le montrait Jean-Baptiste Marcellesi (1971) à propos du vocabulaire politique, un type d'énonciateurs n'est pas forcément repérable au vocabulaire qu'il emploie : il arrive que le discours d'un adversaire soit repris, mis en scène, contesté, brouillant ainsi les pistes. D'où l'importance de la prise en compte du contexte.

Si Alceste repère par des méthodes statistiques ce qu'il y a de commun entre les différents points de vue sur un objet de discours particulier, rappelons que l'analyste a une part essentielle en ce qui concerne *l'interprétation* des résultats. La construction de classes peut laisser croire que le logiciel livre une « vérité intrinsèque » sur le corpus, mais il s'avère que, dès lors que l'on change quelques paramètres (modification des variables par exemple), ces classes peuvent changer.

Il existe donc un risque de dérapage interprétatif qui nécessite de se poser la question de la fiabilité des résultats de ce type d'analyse. Une bonne connaissance du corpus est nécessaire avant l'utilisation d'Alceste et une méthodologie d'analyse des réponses doit être élaborée afin de minimiser ce risque. Les résultats fournis par le logiciel sont des *pistes* qui réclament un retour à la globalité et à la linéarité des textes et qui doivent être croisés avec d'autres types de faits pour une analyse complète.

CONCLUSION

L'analyse de textes est un problème d'une telle complexité que le seul recours à un logiciel d'analyse ne saurait l'épuiser. Il serait illusoire d'imaginer que l'analyse informatique se suffit à elle-même. C'est à l'analyste qu'incombe l'étude des résultats et leur interprétation, et le travail n'est pas là des moindres. Cette interprétation varie en fonction du corpus construit, du questionnement posé et des problématiques retenues.

Si le regard doit rester critique par rapport à l'outil, les résultats obtenus grâce à Alceste permettent néanmoins une analyse de contenu très convaincante et fournissent d'intéressantes pistes de réflexion. Des indicateurs forts peuvent en ressortir quand bien même le logiciel ne relève qu'un type de phénomène, le vocabulaire, et occulte les autres faits linguistiques.

Alceste, auxquelles les dernières versions confèrent une interface conviviale, est actuellement distribué par une société toulousaine, la société Image[4].

BIBLIOGRAPHIE INDICATIVE

Alceste version 3.0, Mode d'emploi, 1995.
BARDIN Laurence, 1991, *L'analyse de contenu*, Presses Universitaires de France, Paris.
HARRIS Zellig, 1969, « Analyse de discours », *L'analyse du discours*. *Langages* n°13, pp. 8-45.
LEBART Ludovic et SALEM André, 1994, *Statistique textuelle*, Dunod, Paris.
MARCELLESI Jean-Baptiste, 1971, *Le congrès de Tours (Décembre 1920). Études sociolinguistiques*, Le pavillon-Roger Maria Editeur, Paris.
PECHEUX Michel, 1969, *Analyse automatique du discours*, Dunod, Paris.
RASTIER François *et alii.*, 1994, *Sémantique pour l'analyse*, Masson, Paris.
REINERT Max, 1983, « Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte », *Les Cahiers de l'analyse des données* vol VIII n° 2, pp 187-198.
REINERT Max, 1990, « ALCESTE, une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval », *Bulletin de méthodologie sociologique* n°26, pp. 24-54.
REINERT Max, 1993, « Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars », *Langage et Société* n° 66, pp. 5-39.
REINERT Max, 1999, Quelques interrogations à propos de l'"objet" d'une analyse de discours de type statistique et de la réponse "Alceste", *Langage et Société* n° 90, pp. 57-70.
TOURNIER Maurice, 1980, « D'où viennent les fréquences de vocabulaires ? », *Mots* n°1, Presses de la fondation nationale des sciences politiques, Paris, pp. 189-209.

Valérie DELAVIGNE
Le Bourg
76190 Ecretteville les Baons
valerie.delavigne@normandnet.fr

[4] www.image.cict.fr