QUAND L'ORDINATEUR COMPARE DES TEXTES...

Jean-Louis MALANDAIN

Pourquoi l'ordinateur ? A la première lecture, tout bon lecteur sait, intuitivement, si un texte ressemble à un autre, s'il a déjà vu quelque part le même contenu et, parfois, le même style. De la même manière, un enseignant est en mesure de savoir si un texte sera d'un abord facile ou non pour ses élèves ; il connaît leur niveau de compétence et jauge rapidement le contenu, le lexique et la structure du texte ; il peut donc faire un pronostic plus sûrement qu'un programme informatique puisqu'il prend en compte le SENS du texte, ce qu'aucune machine ne saurait faire car cette compétence dépasse largement les capacités d'un ordinateur.

C'est pourtant un programme informatique qui, par la comparaison systématique du vocabulaire employé, a permis de démontrer que des sonnets restés anonymes étaient bien de la main de Shakespeare. Quand la recherche ne porte que sur le contenu lexical, l'ordinateur offre une aide appréciable : mieux que quiconque, il peut repérer la présence de certains mots et les signaler. Pour vérifier une intuition ou obtenir une première estimation, le lexique est un indicateur stable et objectif, tout à fait satisfaisant.

Pour peu qu'on ne s'abuse pas sur ses capacités réelles, un programme de comparaison rendra de grands services aux enseignants et aux étudiants en signalant, dans un texte quelconque, les mots déjà rencontrés dans un autre texte ou qui figurent dans une liste établie à l'avance, comme celle des mots réputés fréquents de la langue française. Le calcul rapide (autre vertu de l'informatique) du pourcentage de mots signalés donne une information limitée mais cohérente : une forte densité de mots fréquents révèle une écriture "naturelle", délibérée ou non. Inversement, l'observation des mots plus rares est une piste pour dégager la spécificité d'un texte. De même peut-on augurer que deux textes présentant un fort pourcentage de mots communs (non grammaticaux) ont une certaine parenté thématique. On pourrait s'étonner de la capacité à repérer les mots "non grammaticaux" donc plutôt porteurs de sens. Là encore, la procédure utilisée est purement mécanique, ce qui permet

d'analyser des textes qui n'ont fait l'objet d'aucune indexation préalable ou de faire l'économie d'un véritable dictionnaire dont le stockage et la consultation rapide exigeraient des moyens autrement sophistiqués.

Aux particularités du décompte lexical, l'ordinateur ajoute la capacité d'affichage qui permet de mettre en valeur ou, au contraire, de masquer certains mots pour motiver la consultation du texte dans la salle de classe, surtout quand on dispose d'un grand écran. C'est ainsi qu'on peut aborder un texte en distinguant ce qui est fréquent ou déjà connu de ce qui ne l'est pas ; mais on peut aussi ne montrer que les mots "faciles" et masquer les autres... ou réciproquement. Autant de modalités pour varier la présentation, inciter à la lecture et formuler des hypothèses sur le contenu. Pour des adolescents habitués à regarder des écrans et que la télévision captive souvent plus que les livres, ces voies nouvelles pour consulter et observer des textes sont parfois l'occasion de renouer avec la littérature.

Ainsi, par petites touches et très modestement au regard de l'intelligence d'un texte, la machinerie informatique apporte sa contribution à la didactique. C'est toute l'ambition du programme COMPARA, l'un des premiers d'une collection intitulée "ALSO" : Activités Langagières Sur Ordinateur [1].

FONCTIONNEMENT

Le programme permet d'appeler un premier fichier quelconque et d'en extraire un passage limité à 100 lignes afin d'y signaler, sous différentes formes, les mots fréquents (du Français Fondamental) ou les mots présents à la fois dans ce fichier et dans un second appelé à la suite. Au moment de l'affichage du texte, les mots repérés par le programme sont colorés en jaune (ou masqués quand on a choisi cette option). En fin de présentation, l'utilisateur peut franchir une étape supplémentaire et obtenir le pourcentage dit "de compréhension" (rapport entre le nombre de mots communs et le nombre total de mots dans le texte visé) ainsi que

[1] Le programme COMPARA, élaboré sous DOS, est utilisable directement sur n'importe quel ordinateur PC, à partir de fichiers Ascii. Il est fourni avec un "lanceur" WINCOMPA qui en facile l'emploi sous Windows et qui gère la plupart des problèmes liés au format des fichiers textes (Ascii ou Ansi, longueur des lignes, TXT ou HTML). En attendant la rétribution automatique des productions informatiques, l'auteur laisse aux utilisateurs le soin d'apprécier une contribution en fonction de l'usage et du service rendu. Le programme COMPARA est en libre circulation et disponible chez l'auteur en envoyant une participation pour la disquette et son expédition

la liste des mots repérés. Les résultats de la comparaison peuvent être imprimés ou sauvegardés et, au besoin, réutilisés ultérieurement afin d'étendre l'observation à d'autres textes, par cumuls successifs.

Bien que le programme ait été conçu pour le français, rien n'empêche de l'appliquer à d'autres langues gérées par le clavier en usage. Les mêmes comparaisons seront conduites à partir de deux textes ou en constituant des listes de référence (par exemple, les mots fréquents de telle ou telle langue). On peut même envisager une comparaison entre des textes de langues différentes pour déceler les mots transparents ou les "faux amis".

OBJECTIFS ET APPLICATIONS DIDACTIQUES

Comment savoir rapidement si un texte est difficile pour un niveau d'apprentissage ? Comment repérer avec certitude les mots nouveaux par rapport à d'autres textes étudiés précédemment ? Comment vérifier si une liste de mots est présente dans un texte en les visualisant dans leur contexte ? La réponse à ces questions intéresse autant l'enseignant soucieux d'organiser une progression que l'étudiant à la recherche d'aides en situation d'autonomie.

A ces objectifs que permet la capacité de la machine à repérer les mots s'ajoutent des fonctions qu'on pourrait assimiler au feuilletage d'un ouvrage. Cette opération, tellement naturelle dans une librairie - c'est le premier contact décisif avant la lecture -, est souvent invoquée pour marquer l'avantage du livre sur l'écran. Pourtant, puiser au hasard une centaine de lignes dans un roman et les afficher immédiatement pour toute la classe est une économie appréciable. C'est aussi une approche motivante procédant par coups de sondes dans les eaux profondes d'une oeuvre. On connaît bien les exploitations des débuts et des fins de romans dont on peut dresser une typologie. Le programme COMPARA propose l'observation des "coeurs" de romans... En voici un exemple (abrégé) que la machine a puisé en quelques secondes au milieu d'un fichier d'environ 1000 lignes pour donner envie de lire ce qui précède ou ce qui suit :

Or, l'arquebusier, revenant ce soir-là, comme de coutume, trouva la porte close et les lumières éteintes. Cela l'étonna beaucoup, la guette n'étant pas sonnée au Châtelet et, comme il ne rentrait point d'ordinaire sans être un peu animé par le vin, sa contrariété se produisit par un gros jurement qui fit tressaillir Eustache dans son

Si la motivation a fonctionné, le lecteur trouvera ce qu'il cherche dans *La main enchantée* de Gérard de Nerval...

EVALUER LA DIFFICULTÉ D'UN TEXTE

En prenant pour base de référence la liste des mots fréquents établie par l'enquête sur le Français Fondamental (Gougenheim, 1964), le programme COMPARA permet de repérer la présence des 1000 mots les plus fréquents, considérés comme le noyau dur de la langue, lexique normalement connu par tous les débutants après une année d'étude. Le pourcentage de compréhension est une indication fiable sur la facilité relative du texte. La liste du Français Fondamental, telle qu'elle a été publiée, correspond aux entrées d'un dictionnaire ; elle a été affinée en reprenant les résultats d'une recherche sur les formes verbales (cf. Le Français dans le Monde, n° 83, sept. 1971). Ce sont les formes conjuguées réellement fréquentes qui sont utilisées dans le programme : le verbe "avoir" est fréquent, la forme "avait" l'est aussi mais pas la forme "eut". On obtient ainsi un résultat précis sans recourir à des outils spécialisés pour "conjuguer" les infinitifs.

La limitation volontaire à des formes fréquentes permet cependant le repérage des variations morphologiques régulières telles que les féminins et les pluriels. De plus, elle fait apparaître dans les textes des expressions ou des mots composés qu'on ne prend pas toujours en compte et qui sont pourtant constitués de mots fondamentaux : c'est le cas d'un mot comme "après-midi" qui n'est pas dans la liste du Français Fondamental. Le système adopté offre donc une bonne rentabilité malgré la grande économie des moyens mis en oeuvre.

Dans l'exemple qui suit, un même texte (tiré de "Ô vous, frères humains" d'Albert Cohen) est présenté selon deux modalités : ce qui est fréquent et ce qui ne l'est pas, en conservant la même "image" du texte en partie masqué. En un seul regard, on apprécie la densité relative des éléments signalés :

A En ce ¤¤¤¤¤¤¤¤ jour du mois d'août,
à trois heures cinq de l'après-midi, ¤¤¤¤¤¤¤ du lycée
où j'étais allé suivre un cours de vacances pour
¤¤¤¤¤¤¤ en ¤¤¤¤¤¤¤¤¤¤¤¤¤, je ¤¤¤ un ¤¤¤¤¤¤¤¤¤¤¤¤.
A l'¤¤¤¤¤ de m'intéresser et de ¤¤¤¤¤ de la vie, de ma
vie qui venait de commencer, je m'¤¤¤¤¤¤¤¤¤. C'était un

¤¤¤¤¤¤¤ qui devant sa table ¤¤¤¤¤¤¤ ¤¤¤¤¤¤¤¤¤¤ avec ¤¤¤ les ¤¤¤¤¤¤¤ de son ¤¤¤¤¤¤¤¤¤ ¤¤¤¤¤¤¤¤¤. Fort ¤¤¤¤¤ et me ¤¤¤¤¤¤¤¤ déjà mille ¤¤¤¤¤¤¤¤¤ de connaissances nouvelles, je me ¤¤¤¤¤¤¤¤ au premier ¤¤¤¤ pour mieux entendre et ¤¤¤¤¤¤¤ le ¤¤¤¤¤ ¤¤¤¤¤¤¤ aux fines ¤¤¤¤¤¤¤¤¤¤.

B ¤¤ ¤¤ seizième ¤¤¤¤ ¤¤ ¤¤¤¤ ¤'¤¤¤¤,
¤ ¤¤¤¤¤ ¤¤¤¤¤¤ ¤¤¤¤ ¤¤ ¤'¤¤¤¤¤-¤¤¤¤, sortant ¤¤ ¤¤¤¤¤ ¤¤ ¤'¤¤¤¤¤ ¤¤¤¤ ¤¤¤¤¤¤ ¤¤ ¤¤¤¤¤ ¤¤ ¤¤¤¤¤¤¤¤ ¤¤¤¤ cancres ¤¤ arithmétique, ¤¤ vis ¤¤ attroupement.
¤ ¤'affût ¤¤ ¤'¤¤¤¤¤¤¤¤¤¤ ¤¤ ¤¤ jouir ¤¤ ¤¤ ¤¤¤, ¤¤ ¤¤ ¤¤¤ ¤¤¤ ¤¤¤¤¤¤ ¤¤ ¤¤¤¤¤¤¤¤¤, ¤¤ ¤'approchai. ¤'¤¤¤¤¤ ¤¤ camelot ¤¤¤ ¤¤¤¤¤¤ ¤¤ ¤¤¤¤¤ pliante démontrait ¤¤¤¤ feu ¤¤¤ mérites ¤¤ ¤¤¤ détacheur universel. ¤¤¤¤ animé ¤¤ ¤¤ forgeant ¤¤¤¤ ¤¤¤¤¤ félicités ¤¤ ¤¤¤¤¤¤¤¤¤¤¤¤¤ ¤¤¤¤¤¤¤¤¤, ¤¤ ¤¤ faufilai ¤¤ ¤¤¤¤¤¤¤ rang ¤¤¤¤ ¤¤¤¤¤ ¤¤¤¤¤¤¤¤ ¤¤ admirer ¤¤ blond camelot ¤¤¤ ¤¤¤¤¤ moustaches.

Simple information pour l'enseignant qui prépare son cours (le texte comporte 76 % de mots relativement faciles), cette présentation devient un support didactique intéressant quand on se demande, à un niveau avancé, quelle information passe dans chacun des cas : le premier bloc où n'apparaissent que les mots faciles est proposé à un groupe et le second (mots difficiles) à un autre groupe, avec pour consigne commune de formuler des hypothèses et d'échanger des informations afin de raconter l'événement, avant de vérifier en affichant le texte sans masquage mais en deux couleurs, jaune pour les mots faciles et bleu pour les autres. En d'autres circonstances, l'une ou l'autre de ces images masquées peuvent guider une reconstitution, une reformulation ou même l'invention d'un nouveau texte avec pour seule obligation de conserver les mots affichés dans le même ordre.

COMPARER DEUX TEXTES

Les mots communs peuvent être signalés ou masqués selon qu'on souhaite les mettre en valeur ou les faire deviner. Un cas très simple consiste à mettre en évidence le vocabulaire rencontré précédemment : faire apparaître dans la leçon 2 ce qui a été vu dans la leçon 1, repérer dans un récit les mots d'un épisode précédent... Dans cette situation, les mots communs sont déjà connus et les mots spécifiques sont nouveaux ; les deux listes ainsi obtenues constituent une base utile pour préparer des exercices. D'autres cas de comparaison sont particulièrement intéressants quand une même information est diffusée par des voies ou des sources parallèles, avec les variantes ou les transformations qu'on imagine : relation d'un événement oralement et par écrit, bulletin à la radio et dans la presse, articles traitant le même sujet dans des organes de tendances opposées.

Afin d'éviter les mots courts qui sont aussi parmi les plus fréquents, il est possible de sélectionner la longueur prise en compte (par exemple, commencer les repérages à partir de 3 ou 4 caractères). C'est un filtrage intéressant puisque la plupart des déterminants et des anaphoriques ne seront pas pris en compte. Une sélection plus rigoureuse consiste à exclure les mots fréquents de la comparaison, hors de la liste des 1 000 premiers mots du Français Fondamental dont la présence est "normale" dans tous les textes. Grâce à ces deux procédés, on isole des mots réellement porteurs de sens, sans tenir compte des outils grammaticaux ; il est alors possible de mettre en évidence de fortes convergences lexico-sémantiques entre des textes relevant d'une même thématique, à la façon des méthodes de recherche documentaire. S'il reste 20 à 30 % de mots communs, la parenté entre deux textes est très forte. Mais il en faut moins pour établir des relations et suggérer des commentaires. Ainsi, à propos des notices biographiques de Mitterrand et de Gaulle (diffusées par l'AFP), on pourrait se demander pourquoi les mots communs aux deux sont : **de Gaulle** (mais pas Mitterrand), **juin, secrétaire d'état, président, mai, gouvernement, résistance, propre, conseil, assurances, acquis, Etats, échec, rejoint, en selle...**

Un cas particulier consiste à signaler dans un texte quelconque la présence de mots dont la liste a été établie à l'avance (par le professeur ou les élèves) afin de repérer une particularité du texte. Par exemple, une liste de conjonctions pour visualiser la structure, une liste de mots rendant compte d'un thème. On peut ainsi imaginer que toute une classe recherchera les mots en relation avec les saisons puis en vérifiera la pré

sence dans une poésie. C'est également possible pour des listes concernant des notions (le temps, l'espace, les modalités...) ou encore des thèmes comme le feu... ou l'optimisme. C'est affaire de niveau et l'intérêt de la démarche est autant dans la réflexion qui préside à la constitution d'une liste fiable que dans les commentaires qui accompagneront le résultat de la recherche.

Jean-Louis MALANDAIN
19, rue du Docteur Renou
F-49620 La Pommeraye
Tél. et Fax 02 41 77 75 79

Ce logiciel est disponible dans la bourse d'échanges de l'EPI sous la référence 9411-FR.