DOCUMENT RESUME

ED 175 255 FL 010 473

AUTHOR Savard, Jean-Guy
TITLE L'Utilisation de l'ordinateur en lexicometrie (The Use of the Computer in Lexicometry). Series B-1.
INSTITUTION Laval Univ., Quebec (Quebec). International Center for Research on Bilingualism.
PUB DATE 67
NOTE 18p.; Paper presented at "Journees d'etudes sur le vocabulaire des langues de specialites" (Saint-Cloud, France, November 23-30, 1967)
LANGUAGE French

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Bilingualism; *Computational Linguistics; Computer Oriented Programs; *Computer Programs; Computer Science; Definitions; *Dictionaries; Information Processing; Language Research; *Languages for Special Purposes; Lexicography; *Lexicology; Research Methodology; Research Problems; Scientific Literacy; *Vocabulary

ABSTRACT

This report treats some of the technical difficulties encountered in lexicological studies that were undertaken in order to establish a basic vocabulary. Its purpose is to show that the computer can overcome some of these difficulties, and specifically that computer programming can serve to establish a vocabulary common to scientific and technical languages. The discussion of the advantages of computer programming centers on the following topics: (1) availability, including a description of the compilation and correction of data, indexing, and programming; and (2) general vocabulary that is scientifically oriented, including a description of the process of choosing and identifying words, and the choice and structure of dictionaries. (AMH)

Centre international de recherche sur le bilinguisme

L'UTILISATION DE L'ORDINATEUR EN LEXICOMETRIE

JEAN-GUY SAVARD

1967

CIRB
ICRB

L'UTILISATION DE L'ORDINATEUR EN LEXICOMETRIE

Jean-Guy Savard
Centre international de recherches sur le bilinguisme

Communication présentée aux Journées d'études sur le vocabulaire des langues de spécialités tenues à Saint-Cloud du 23 au 30 novembre 1967.

Plan de la Communication

0. Introduction

1. La disponibilité
   1.1 La compilation des données
   1.2 La correction des données
   1.3 Le calcul de l'indice de disponibilité
   1.4 Le programme

2. Le vocabulaire général d'orientation scientifique
   2.1 La séparation des mots
   2.2 L'identification des mots
      2.2.1 Le choix des dictionnaires
      2.2.2 La structure des dictionnaires
         2.2.2.1 Le dictionnaire des mots-forts
         2.2.2.2 Le dictionnaire des mots-fonctionnels
         2.2.2.3 Le dictionnaire morphologique

3. Conclusion

L'UTILISATION DE L'ORDINATEUR EN LEXICOMETRIE

0. INTRODUCTION

Le but de la présente communication est de traiter des difficultés techniques rencontrées lors de nos études lexicologiques en vue d'établir un vocabulaire de base. En prenant comme exemple nos études sur la disponibilité du vocabulaire, nous voulons montrer que l'ordinateur électronique permet de surmonter certaines de ces difficultés. Nous soulignons le fait que les programmes mis au point pour cette recherche ont rendu possible le traitement efficace de ces données, et ont favorisé aussi le développement de techniques générales qui peuvent être utilisées pour apporter des solutions à des problèmes analogues. Nous verrons en particulier, dans quelle mesure ces programmes peuvent servir à l'établissement d'un vocabulaire commun aux langues scientifiques et techniques.

1. LA DISPONIBILITE

Dans le cas de la disponibilité, le travail à accomplir se divise en trois étapes bien distinctes: la compilation des données, la correction des données et le calcul de l'indice.

de disponibilité.

1.1 LA COMPILATION DES DONNEES

Les sujets examinés écrivent les mots de chaque centre d'intérêt sur une feuille. Ces mots sont transcrits sur cartes perforées. Chaque carte porte un numéro de dossier qui permet de retracer la région où est menée l'enquête, le nom de l'école, le niveau scolaire et le nom de l'élève. Le contenu des cartes est enregistré sur bandes magnétiques. A partir des bandes magnétiques, l'ordinateur imprime les listes de données initiales sur lesquelles les mots figurent en ordre alphabétique.

1.2 LA CORRECTION DES DONNEES

Pour corriger les données, un être humain doit annoter à la main les listes brutes de données initiales*. Ou bien la graphie donnée par l'élève est correcte, et le correcteur la conserve; ou bien la graphie est mauvaise, et alors il indique à la main à quelle bonne graphie il faut reporter ce mot. On perfore de nouvelles cartes, dites cartes de

---

* Voir Mackey, W. F. et al. : Le vocabulaire disponible en France et en Acadie, P. U. L., 1968. Le chapitre V, décrit le processus de traitement des données.

correction, contenant les indications fournies par le correcteur. L'ordinateur lit concurremment les cartes de correction et les bandes magnétiques de données initiales. Il imprime les listes de vérification, d'où il faudra repartir si l'on juge nécessaire de recommencer la correction.

1.3 LE CALCUL DE L'INDICE DE DISPONIBILITE

Au cours des deux premières étapes, i.e. la compilation et la correction, l'ordinateur conserve pour chaque mot d'un centre d'intérêt, le nombre de fois que les sujets ont fourni ce mot à chaque année du cours, et au total. De même, au moment de la mise sur bandes magnétiques des données initiales, l'ordinateur compte le nombre de sujets examinés. Il peut donc maintenant établir le pourcentage de disponibilité pour chaque mot et imprimer la liste finale contenant les mots classés selon l'indice de disponibilité.

1.4 LE PROGRAMME

Le programme-machine préparé pour cette étude est efficace, puissant et flexible.

Il est efficace, puisqu'il réduit au minimum le temps requis pour effectuer une opération ou une série d'opérations.

Il est puissant, parce qu'il permet d'exécuter automatiquement des opérations aussi complexes que la séparation des mots, la suppression des articles, le classement des mots en ordre alphabétique ou en ordre de fréquence, et même, certaines parties de la correction.

Il est flexible, car il peut facilement être modifié ou utilisé de différentes façons. Ce programme est établi selon une conception relativement nouvelle de la programmation. Suivant la méthode habituelle de programmation, un programme est élaboré pour lire certains paramètres définissant le problème; lire les données du problème; effectuer des opérations; imprimer des résultats intermédiaires; calculer et imprimer le résultat final.

Le nouveau mode de programmation utilisé dans cette étude s'appelle "programmation par blocs d'instructions". Un programme est alors formé d'un assemblage de blocs ou mieux de modules d'instructions, plus ou moins indépendants les uns des autres, et auxquels l'utilisateur réfère à l'aide de cartes de commande qu'on incorpore au programme selon les besoins.

Ce genre de programme est préparé pour lire d'abord une carte de commande. Suivant le contenu de cette carte, l'ordinateur recherche dans le programme complet tel groupe d'instructions, exécute les opérations demandées, puis revient lire une autre carte de commande. Dans le cas de l'étude sur la disponibilité, le programme contient environ 1,000 énoncés en FORTRAN répartis en 15 groupes. On peut, par exemple, demander à l'ordinateur d'enregistrer sur bandes magnétiques, les données initiales du centre d'intérêt numéro 10; d'enlever les articles; de ne garder qu'une entrée pour chaque élément, tout en additionnant les occurrences de cet élément; de classer ces éléments en ordre alphabétique; de les imprimer avec ou sans leur fréquence; puis de revenir lire les cartes de correction du centre numéro 4, et de reprendre tout le processus.

La mise au point de ce programme a contribué au développement de sous-programmes très efficaces pour le traitement non-numérique de l'information. Un certain nombre de ces sous-programmes ont été utilisés pour nos études sur la valence lexicale, par exemple, pour compter le nombre de fois qu'un mot est employé pour en définir un autre; pour

compter le nombre de fois qu'un mot peut entrer en combinaison avec un autre; ou encore pour analyser les méthodes d'enseignement des langues.

Voyons maintenant si ce programme et ces sous-programmes pourraient être utiles dans des études visant à l'établissement du vocabulaire général d'orientation scientifique.

2. LE VOCABULAIRE GENERAL D'ORIENTATION SCIENTIFIQUE

2.1 LA SEPARATION DES MOTS

Comme on l'a dit déjà, dans l'étude sur la disponibilité, il s'agissait de préparer des listes de mots détachés. Au moment de la transcription sur cartes, on a séparé chaque mot par un trait oblique. Ce trait servait de ligne de démarcation entre les mots.

L'étude du vocabulaire général d'orientation scientifique se fera à l'aide de textes suivis. En vue de pouvoir reconstituer le texte, il faut le codifier au fur et à mesure que se fait la transcription sur cartes perforées. On peut par exemple, réserver soixante-dix positions de la carte IBM pour écrire les mots, alors que les dix dernières colonnes

contiennent de l'information semblable à celle que contenait le numéro de dossier lors de l'enquête sur le vocabulaire disponible. Ici le numéro réfère au titre du volume, à la page, à la ligne et au numéro d'ordre du mot. Toute cette information constitue les données initiales dont nous disposons à l'entrée de l'ordinateur. Les cartes sont lues par l'ordinateur qui en enregistre le contenu sur bandes magnétiques. Le contenu d'une carte constitue une unité d'enregistrement, ou si l'on veut un champ de lecture sur la bande magnétique. A l'aide d'un sous-programme on commence le traitement, i.e. la séparation des mots. Evidemment, cela suppose qu'on a donné à la machine une définition opérationnelle de ce que l'on considère comme un mot. Disons qu'un mot, dans ce cas, c'est une série de caractères précédés ou suivis d'un espace.

L'ordinateur lit une position, un caractère sur ruban. Là, il vérifie s'il s'agit

1) d'un espace blanc,
2) d'un caractère alphabétique ou d'un chiffre,
3) d'un signe de ponctuation.

En somme, le rôle de ce sous-programme se résume ainsi. Il lit le texte fourni à l'entrée, il fait disparaître les espaces blancs, et il écrit sur la bande magnétique, dans une autre unité d'enregistrement, chaque mot, chaque chiffre ou chaque signe de ponctuation, suivi l'un ou l'autre, de l'information nécessaire à la reconstitution du texte. On a donc sur bande magnétique, le mot à mot du texte que l'ordinateur peut imprimer à volonté. A la sortie, sur papier, on réserve, disons, 30 espaces pour le mot et 10 espaces pour l'information.

Il est possible encore ici, à l'aide d'autres sous-programmes, de classer les mots en ordre alphabétique, de compter leur fréquence, et même leur distribution dans le texte, e.g. ce mot est revenu tant de fois au total, dont tant de fois à la page 10, 30 ou 50.

Un autre sous-programme utilisé pour publier des index analytiques ou pour mesurer, en valence lexicale, la puissance de combinaison des mots, pourrait être utile à cette étape-ci de la recherche. On a pu séparer les mots, on peut maintenant les réunir. L'ordinateur peut facilement imprimer tel mot accompagné du mot qui le précède et du

mot qui le suit. De même, on peut faire imprimer des groupes de 4, 5 ou 10 mots. À ce moment-là, les techniques de correction semi-automatique utilisées dans l'étude du vocabulaire disponible pourraient être adaptées à ces nouvelles fins.

Plus simplement encore, on peut fournir à l'ordinateur un dictionnaire de syntagmes, qui lui permettrait d'identifier les groupes de mots relevés dans le texte. Voilà posée la question de l'identification des mots ou des syntagmes d'après un dictionnaire pré-établi.

2.2 L'IDENTIFICATION DES MOTS

Il ne suffit pas en effet de savoir que tel volume contient 5,000 ou 10,000 éléments différents. Il est intéressant de savoir aussi à quelle catégorie grammaticale appartiennent ces mots; de savoir si ces mots ont un haut degré de valence, de disponibilité ou de fréquence dans la langue commune. Pour obtenir ces renseignements, il faut avoir recours à un dictionnaire pré-établi selon certains critères.

2.2.1 LE CHOIX DES DICTIONNAIRES

Comment choisir ce ou ces dictionnaires ? Devons-nous utiliser un dictionnaire à entrées invariables et contenant tous

les mots que l'on puisse rencontrer dans tel texte ? Ce genre de dictionnaire serait de consultation facile. Par contre, il serait très volumineux, et il faudrait beaucoup de temps et d'efforts pour en fixer le contenu.

Il vaut mieux se servir de plusieurs dictionnaires: un dictionnaire de radicaux pour les mots-forts, un dictionnaire des mots-fonctionnels et un dictionnaire morphologique. Le dictionnaire des mots-forts ne contient que la partie invariable des mots. Le dictionnaire des mots-fonctionnels est à peu près complet. Le dictionnaire morphologique n'est pas autre chose qu'une liste des terminaisons possibles. De l'organisation de ces dictionnaires dépend en grande partie, l'efficacité de la recherche.

2.2.2 LA STRUCTURE DES DICTIONNAIRES

2.2.2.1 LE DICTIONNAIRE DES MOTS-FORTS

Les décisions à prendre quant à la structure des divers dictionnaires dépendent d'un certain nombre de facteurs comme par exemple: la puissance de l'ordinateur dont on dispose, le matériel à étudier, ou encore le type de recherche que l'on veut faire dans le dictionnaire. Cette recherche peut être du type linéaire ou du type binaire.

Supposons pour les besoins de la cause que tous les mots-forts du texte suivi sont maintenant en ordre alphabétique. Tentons une recherche linéaire dans un dictionnaire alphabétique. C'est très facile. Chaque mot à identifier se trouve dans le dictionnaire, très près du mot cherché précédemment. Il n'est donc pas nécessaire de lire tout le dictionnaire pour trouver un mot. Point n'est besoin non plus, de garder continuellement tout le dictionnaire. L'expérience démontre qu'il suffit amplement de garder en mémoire une quinzaine de radicaux. Seulement, selon cette façon de procéder, il vaut mieux, parfois, avoir dans le dictionnaire, plusieurs radicaux pour le même élément. Pensons aux verbes faire et aller. Pour chaque forme différente, il faut d'abord comparer le mot à identifier au radical du mot identifié immédiatement auparavant. Si le radical est le même, on cherche seulement la terminaison dans le dictionnaire morphologique. Sinon, il s'agit d'un mot différent et alors on cherche un autre radical.

Le laps de temps nécessaire à la recherche dans de telles conditions dépend évidemment de l'étendue du dictionnaire. Si l'on a un dictionnaire de 10,000 mots et un

texte contenant 1,000 éléments différents, il faudra en moyenne, 10 comparaisons par mot.

Imaginons maintenant qu'on veuille faire une recherche linéaire dans un dictionnaire alphabétique, alors que les mots du texte sont restés en ordre textuel; ou vice-versa, une recherche linéaire dans un dictionnaire de fréquence et un texte en ordre alphabétique. Cette fois, théoriquement, une recherche linéaire pour identifier les 1,000 éléments dans un dictionnaire de 10,000 mots supposerait 5,000 comparaisons par mot, en moyenne. Par ailleurs, une recherche binaire, dans les mêmes conditions, nécessiterait, en moyenne, trois cents fois moins de consultations pour identifier chaque mot.

Jusqu'à présent nous avons utilisé une méthode combinée de recherche linéaire et binaire dans un dictionnaire alphabétique pour identifier les mots d'un texte classés en ordre alphabétique. Ce choix est dû, en grande partie, au fait qu'au moment de l'élaboration des programmes, nous ne disposions que d'un ordinateur moyen, i.e. la machine IBM-1410. Maintenant que nous pouvons utiliser la machine IBM-360 modèle 40, munie de disques magnétiques, il est

plus facile de consulter un dictionnaire plus étendu. L'enregistrement sur disques permet un accès plus rapide à toutes les entrées dans un dictionnaire, même s'il est assez volumineux. C'est pourquoi, on tentera sûrement de laisser les mots à identifier en ordre textuel. Quant au dictionnaire, il pourra être constitué du vocabulaire du français fondamental que nous avons déjà, et auquel on ajoutera le dictionnaire du vocabulaire scientifique que possède maintenant l'équipe du C.R.E.D.I.F.

2.2.2.2 LE DICTIONNAIRE DES MOTS-FONCTIONNELS

La liste complète des mots-fonctionnels est tellement restreinte qu'elle est facile à établir. Et il ne faudra jamais beaucoup de temps pour retrouver dans une liste alphabétique ou dans une liste de fréquence, l'un ou l'autre des 270 mots-fonctionnels du français.

2.2.2.3 LE DICTIONNAIRE MORPHOLOGIQUE

Les terminaisons des mots dans un texte suivi se présentent à peu près au hasard. En gros, on peut dire que si l'on fait une recherche linéaire dans un dictionnaire alphabétique des terminaisons, contenant 200 éléments, il faudra, en moyenne, 100 comparaisons pour identifier chaque mot.

Si l'on fait une recherche du type binaire, dans les mêmes conditions, il faudra en moyenne, 8 comparaisons par mot.

Dans le dictionnaire morphologique, les terminaisons seront en ordre alphabétique, tout simplement parce qu'on ne connaît pas leur fréquence.

3. CONCLUSION

Voilà un exposé, beaucoup trop schématique à mon gré, des principaux programmes élaborés pour mener à bien nos études lexicologiques.

Quant aux résultats de ces recherches, on ne m'en voudra pas, j'en suis sûr, de mentionner que la revue IRAL a publié en juillet dernier un résumé de nos études sur la valence lexicale*. D'autre part, les Presses de l'Université Laval pourront distribuer dès 1968, deux volumes intitulés: LE VOCABULAIRE FRANCAIS DISPONIBLE EN ACADIE ET EN FRANCE.

---

* W. F. Mackey et Jean-Guy Savard: The Indices of Coverage, dans IRAL, volume V, numéros 2 et 3, Heidelberg, 1967.