N° ISSN 0223-3843

D'après B.N. Ms. Fr. 9186
PS

# LE MEDIEVISTE ET L'ORDINATEUR

N°1 PRINTEMPS 1979

## EDITORIAL

En mai 1975, le colloque franco-italien réuni à Rome pour discuter des problèmes posés par l'utilisation de l'informatique en histoire médiévale concluait ses débats en affirmant que le manque d'information et de coordination était à l'origine d'une mauvaise ou insuffisante utilisation de l'ordinateur pour l'exploitation des sources du Moyen-Age. On décida de procéder à une enquête internationale destinée à repérer les exploitations existantes, à diffuser leurs résultats, à les mettre en relations quand leurs objectifs étaient voisins. Cette enquête s'amorce lentement. Mais, de toutes façons, il ne peut être question de la poursuivre d'une façon satisfaisante si le même travail n'est d'abord mené sur le plan national.

Dès 1976, un Comité formé de quelques utilisateurs de la machine (1) a donc lancé un "questionnaire", distribué à tous les médiévistes, universitaires ou chercheurs, aux conservateurs des dépôts d'archives, également aux historiens du droit, dont l'objet a été de recenser non seulement les traitements en cours, mais aussi les projets, voire les simples velléités d'exploitation : en effet, pour les auteurs du questionnaire, il ne s'agissait pas et il ne s'agit toujours pas d'être et de mettre autrui au courant de ce qui "a marché" mais également de se renseigner sur ce qui a tourné court, sur les raisons de l'échec ou du renoncement, éventuellement sur les moyens d'y remédier.

Un dixième environ des destinataires a fourni une réponse au questionnaire. C'est peu ; mais suffisant pour créer entre les 25 membres de ce premier noyau un réseau d'informations qui peut-être les aidera à mieux orienter leur effort et, qui sait, à persuader avec plus d'assurance les réticents lorsqu'il leur semblera

---

(1) Les signataires de cet éditorial, sous l'égide de l'Institut de Recherche et d'Histoire des Textes.

---

Pour recevoir gratuitement ce bulletin ou le faire parvenir à d'autres, pour tout apport ou demande de renseignement, s'adresser à :

Section Informatique
Institut de Recherche et d'Histoire des Textes
Centre National de la Recherche Scientifique
40, avenue d'Iena
75116 PARIS
(Pour les destinataires C.N.R.S.)

ERA 713
Université de Paris I

17, Rue de la Sorbonne
75005 PARIS
(Pour les universitaires).

que l'utilisation de l'informatique peut porter ses fruits. L'effort, pour parvenir à créer ce réseau, est mené dans trois directions : se rendre sur place pour voir ce qui existe et quels sont les moyens informatiques dont disposent les auteurs de projets ; organiser des rencontres entre ces derniers ; enfin, diffuser périodiquement un bulletin de liaison. La première tâche a déjà été amorcée ; mais il nous semble que la seconde, l'organisation de la rencontre, ne portera ses fruits que si un premier bulletin de liaison a déjà circulé et provoqué des réactions.

Sans qu'elles n'aient rien de définitif, nous avons retenu pour ce journal les rubriques suivantes : un article de fond concernant un type d'exploitation donné, comportant la description détaillée d'un projet s'y rattachant et la liste d'autres projets analogues ; une note technique (description d'un matériel, etc.) ; un compte rendu d'une visite effectuée à un centre de calcul en Sciences Humaines ; un calendrier des rencontres concernant l'informatique et les Sciences Humaines éventuellement assorti de notes bibliographiques ; enfin un "courrier des lecteurs" où nous accueillerons aussi bien les réflexions sur les articles parus ou les suggestions sur la composition du bulletin, son rythme de parution, que les demandes d'information sur les programmes existants et surtout utilisables par d'autres. Nous le souhaitons abondant et varié, car ce sont les échanges de vue et les dialogues que nous aurons su établir entre nos lecteurs qui seront, pour le journal, le gage de réussite le plus concluant.

Caroline Bourlet, Lucie Fossier, Jean-Philippe Genêt,
Christiane Klapisch, Jacques Lefort, Josette Metman, Gian-Piero Zarri

## UN TYPE D'EXPLOITATION : LE TRAITEMENT DE TEXTE.

Le traitement de texte - c'est de ce mot que l'on désigne l'ensemble des opérations que l'ordinateur fait subir aux textes enregistrés en mémoire pour les présenter à l'utilisateur sous une forme différente de celle qu'ils revêtaient à l'entrée - est bien de tous les types d'exploitation appliqués aux Sciences Humaines celui qui connaît le plus de vogue. On ne compte plus les index et les concordances produits dans les officines informatiques du monde entier et considérés, à tort ou à raison, comme les moyens indispensables de connaissance d'un auteur ou d'une oeuvre.

Ce succès est évidemment dû en premier lieu à ce que l'enregistrement de masses considérables de données justifie l'utilisation de l'ordinateur, mais surtout, à ce que le travail accompli par la machine en ce cas satisfait pleinement l'utilisateur parce qu'il a ses limites : tel le texte a été livré par l'homme à l'automate, tel il lui sera restitué, avec seulement une modification de forme qui en facilitera l'étude ; à la machine reviennent manipulations mécaniques, repérages innombrables, comptages fastidieux, à l'homme déductions subtiles et constructions intelligentes.

Mais il est bien évident que le traitement de texte s'avère surtout indispensable pour qui a métier d'étudier la littérature. Qu'en est-il pour l'historien dont les soucis grammaticaux ou stylistiques sont tout de même assez limités (surtout dans la pratique des sources documentaires) et pour lequel la grande affaire sera toujours de saisir le contenu sémantique d'un document, contenu qui, il faut bien le dire, ne transparaît pas toujours clairement au travers du texte ?

Disons tout d'abord que les nouveaux champs d'investigation du médiéviste favorisent le recours aux textes suivis et à leur examen minutieux : l'étude des mentalités, des sociétés ne peut se faire sans connaissance de ce que les contemporains ont pu dire ou penser et qu'on découvrira dans les textes littéraires et narratifs assurément plus que dans les sources documentaires. Ces dernières toutefois sont tout aussi susceptibles d'une exploitation sur ordinateur : l'histoire institutionnelle et sociale des premiers siècles s'appuie sur un examen minutieux du vocabulaire des actes. Qui mieux que la machine pourra repérer la première apparition d'un mot, en suivre l'évolution de sens, en calculer la fréquence, par époques, par texte ou de tout autre manière qu'il plaira à l'historien de demander ?

Pour toutes ces raisons, nombreux sont les historiens médiévistes que l'enregistrement et le traitement des textes en machine intéressent. Nous en aurons tout à l'heure un aperçu ; mais peut-être est-il utile auparavant de donner quelques brèves indications sur la façon dont on peut enregistrer un texte et sur les précautions à prendre pour cette opération ; nous parlerons ensuite du parti qu'en peut tirer le médiéviste.

L'entrée en machine

Le traitement de texte est une appellation globale qui recouvre des opérations très diverses et aboutit à des résultats variés, selon les cas.

Le principe en est simple : le texte est enregistré dans une mémoire d'ordinateur par l'intermédiaire de cartes ou de rubans perforés, ou - de plus en plus fréquemment - par impression directe d'une bande magnétique, d'une cassette, d'un disque. Il en ressort sur listings (ou fiches ou microfilms) et sous des formes diverses : texte suivi, concordances, formes triées alphabétiquement, selon les besoins de l'utilisateur.

Il n'est certes pas plus compliqué d'entrer un texte en machine que de le dactylographier. Cependant, cette première opération exige plus d'attention et de rigueur qu'il n'y paraît de prime abord puisqu'une fois le texte enregistré en mémoire, l'intervention humaine n'est plus possible et le produit de sortie comportera toutes les erreurs faites à l'entrée et portera la marque des négligences et des inattentions qui auront présidé à l'introduction en machine : une simple virgule, un point mal placé risquent de perturber un ordre alphabétique ; des signes spéciaux (guillemets, parenthèses, tirets) imprévus peuvent bloquer les opérations ; etc... Toutes les règles d'écriture doivent être soigneusement déterminées au début des opérations, puis attentivement suivies.

Il est rare qu'un texte soit enregistré en machine tel quel, sans la moindre indication supplémentaire. Au moins doit-il dans tous les cas être accompagné d'une référence. S'il s'agit d'un texte littéraire déjà édité, la référence reproduit le plus souvent celle de la publication. Mais l'essentiel est que, pour l'utilisateur, elle permette de situer le mieux possible un mot ou une phrase qui lui seront fournis, au sortir de la machine, isolés de l'ensemble du texte dont ils émanent : si par exemple l'utilisateur est historien toujours désireux de situer son information dans le temps et l'espace et s'il a affaire à un corpus reconstitué de textes disséminés, il peut préférer voir figurer dans sa référence une indication même approximative de date et de lieu plutôt qu'une simple cote de conservation (encore que cette dernière lui soit également indispensable).

Les autres indications entrées en mémoire en même temps que le texte (1) - et qui constituent ce que l'on appelle le "pré-codage" - sont uniquement fonction des produits de sortie que l'utilisateur souhaite obtenir : par exemple, des codes précédant des noms propres permettent l'édition de ces derniers à l'exclusion des autres mots et par conséquent la constitution immédiate d'index de noms de lieu ou de personne. Les codes peuvent également servir à repérer une phrase, un membre de phrase, une partie du discours (citation, formule, etc...) sur lesquels l'historien souhaite porter particulièrement son attention et dont il préfère disposer sans que l'ensemble du texte vienne encombrer son champ de vision.

Codage ou pré-codage ne sont pas les seuls moyens utilisés pour limiter les produits de sortie. Il est également possible, en entrant le texte tel quel, de donner à l'ordinateur des indications pour que tels mots ne figurent pas sur les listings d'index : au lieu de les coder à l'entrée, manuellement, on en constitue des listes - "anti-dictionnaires", dictionnaires de "mots vides" enregistrés en mémoire et que la machine consultera au fur et à mesure qu'elle élaborera l'index pour savoir si tel ou tel mot doit ou non figurer dans ce dernier .

Mais déjà les anti-dictionnaires procèdent d'un autre type d'instruments : ceux dont on pourvoit la mémoire de l'ordinateur pour qu'il puisse mener à bien les opérations commandées par les programmes de sorties du texte.

---

1) Signalons que ces indications peuvent aussi être entrées dans un second temps s'il semble plus aisé d'enregistrer d'abord le texte simple ; dans ce cas les codes reportés à la main sur le premier listing de sortie font l'objet d'une insertion automatique au lieu où ils doivent intervenir pour influer sur les produits de sortie.

Le texte en effet peut être produit sous l'aspect d'une liste alphabétique des formes, réclamée par exemple par tous les travaux de lexicographie. Mais en réalité cette liste sera plus aisément consultable si les formes d'un mot sont regroupées sous une rubrique ; pratique à peu près indispensable pour les langues à flexions, comme le latin. La "lemmatisation" qui consiste précisément à regrouper les formes flexionnelles sous un même "lemme" peut être en partie automatisée si l'on pourvoit l'ordinateur d'une grammaire à l'aide de laquelle il puisse décliner et conjuguer, ou bien plus simplement d'un dictionnaire lui fournissant pour un lemme toutes les flexions qui s'y rattachent. Mais il faut aussi procurer à la machine une autre liste, celle des homographes, dont on réglera le sort manuellement à moins que l'on utilise des programmes très sophistiqués permettant à la machine d'établir elle-même des discriminations entre homographes.

Pour le médiéviste enfin, qui se trouvera toujours confronté à des textes où abondent les diversités orthographiques, il sera bon de disposer d'une liste de regroupement des formes variées sous une forme normalisée : cette liste pourra au choix être distincte ou non de la liste de lemmatisation.

Les résultats

Munie de ces instruments de recherche, la machine s'efforcera de fournir les résultats voulus et qui, bien entendu, varient selon les besoins.

Les produits primaires, ceux qui en tout état de cause seront les premiers à être produits par la machine, sont l'index et la concordance : l'index donne la liste alphabétique des formes, regroupées ou non sous un lemme, selon ce que l'on aura demandé, chacune d'entre elles étant munie d'une référence. La concordance donnera chacune de ces formes dans le contexte choisi (nombre défini de mots, groupe de mots compris entre deux signes de ponctuation, etc..). Pour les études lexicographiques, pour la recherche d'un mot, les index et concordances seront parfois suffisants. Cependant, même sous cette forme, l'historien ne pourra mener une étude de ce type que si formes et contextes sont tout de suite replacés dans un environnement géographique et chronologique précis qui peut, ainsi que nous l'avons indiqué plus haut, se trouver sous une forme codée dans la référence.

Mais l'historien aura souvent besoin de bien plus : si c'est l'étude d'un mot qui l'intéresse, il voudra en connaître au moins la fréquence -absolue et relative - en fonction des dates et des lieux. Si c'est l'étude d'une pensée ce n'est plus le mot isolé qui l'intéressera, mais le discours dont il pourra percevoir l'orientation le sens, l'évolution au travers de l'étude de groupes de mots, et à ce moment les concordances simples ne suffisant pas toujours ; interviennent les recherches de co-occurrences, de coefficients de voisinage, d'environnement, de classement des proximités. On entre à ce moment dans le domaine de la statistique lexicale, voire de l'analyse factorielle et de correspondances si l'on tient à visualiser les résultats.

Ce n'est pas le lieu ici d'insister davantage sur les produits obtenus qui, encore une fois, varient avec les besoins des utilisateurs. Nous préférons laisser la parole aux historiens médiévistes qui ont adopté (et souvent mis au point eux-mêmes) ce type d'exploitation, et qui permettront aux lecteurs, au vu d'un exemple précis, de mieux évaluer les possibilités offertes par le traitement de texte.

Lucie Fossier

Un exemple de programme de traitement de texte : ALINE

Il existe à l'heure actuelle un assez grand nombre de programmes de traitement de texte. A l'étranger, signalons des packages comme COCOA (1), JEUDEMO (2), GENDEX (3), LEXICO (4), ou les programmes du CETEDOC de Louvain ou du L.A.S.L.A. de Liège (5). En France, plusieurs équipes de recherche ont mis au point des bibliothèques de programmes plus ou moins importantes : celle de Madame Gallais-Hamono pour l'étude de la langue des économistes anglosaxons (Université Paris I), celle du Centre de Lexicologie Politique de l'Ecole Normale Supérieure de St Cloud ( programmes de René Pellen à Poitiers.

Mais chacun de ces programmes a été conçu pour répondre à des besoins précis. Mon but est aujourd'hui de présenter un de ces programmes, le package ALINE, mis au point dans le cadre de l'E.R.A. 713 (C.N.R.S. -Paris I), équipe de traitement automatique des sources du bas Moyen Age dirigée par le Professeur Guenée.

A

L'idée de réaliser ce programme est née d'un besoin spécifique : la difficulté d'utiliser les programmes existants pour des textes médiévaux dépourvus d'orthograph régulière et rédigés dans des langues variées (les membres de l'équipe travaillant sur des textes en français, en anglais et en latin) nous obligeait en effet à concevoir un outil nouveau. Les circonstances dans lesquelles nous avons dû réaliser le programme, en outre, nous ont dicté plus précisément nos objectifs : le manque de moyens financiers nous a imposé la recherche systématique de l'économie de fonctionnement ; la certitude de ne pouvoir disposer à moyen terme des services réguliers d'un informaticien nous a poussé à rechercher la plus grande simplicité possible dans le maniement du programme. ALINE est donc un programme qui a opté pour une lemmatisation semi-automatique, qui est peu sophistiqué et ne fournit que les produits de base absolument nécessaires à l'historien (7) : c'est avant tout un générateur de dictionnaires, de référenciations, et de concordances.

Peu diversifié dans ses produits, ALINE est donc d'abord un programme économique En premier lieu, il est économique par le support choisi, la carte perforée : des perforatrices sont disponibles dans beaucoup de centres de recherche, et la perforation peut être effectuée soit par le chercheur, soit par une personne connaissant simplement les rudiments de la dactylographie.

Il y a aussi économie dans la préparation du texte : le programme ne suppose aucun précodage, puisqu'il reconnaît le mot à la seule présence d'un blanc avant et aprés une séquence de caractères. Au reste, dans la mesure où l'utilisateur définit lui-même les signes alphabétiques et les séparateurs qu'il retient, rien ne lui interdit d'introduire un précodage, par exemple en plaçant le signe / devant les noms propres et en déclarant / comme lettre, ce qui fera sortir les noms propres en bloc en queue de dictionnaire.

Il y a enfin, du moins le pensons-nous, économie dans l'exploitation. Voici, à titre d'exemple, une table des coûts de l'exploitation du programme aux mois de décembre 1978 et janvier 1979, sur quatre textes de tailles différentes. Les coûts sont exprimés en francs "CIRCE", si l'on peut dire, puisque tout dépend bien sûr des tarifs pratiqués par chaque centre de calcul.

| | Guibert de Nogent *De Sanctis et Pignoribus Sanctorum* (8) | George Asby *Active Policy of a Prince* (9) | Anonyme *The III Considerations* (10) | Anonyme *Tractatus de Regibus* (11) |
|---|---|---|---|---|
| | (prose latine) | poesie anglaise | prose anglaise | prose anglaise |
| Nombre de cartes perforées (1 carte = 1 ligne imprimée) | 3.468 | 999 | 1.154 | 579 |
| Nombre de formes | 8.660 | 2.008 | 1.936 | 1.268 |
| Nombre de mots | 26.136 | 6.547 | 12.439 | 6.369 |
| Fréquence moyenne des mots | 3,01 | 3,26 | 6,43 | 5,07 |
| Nombre de séparateurs (signes de ponctuation etc...) | 4.586 | 1.349 | 1.065 | 628 |
| Coût de l'enregistrement du texte et de l'édition du texte de contrôle | 32 F 42 | 11 F 02 | 12 F 56 | - |
| Dictionnaire alphabétique | 37 F 13 | 18 F 69 | 18 F 12 | 32 F 28 |
| Dictionnaire par fréquences | 38 F 30 | | | |
| Réalisation des formes | 169 F 14 | 32 F. 73 | 37 F. 43 | |

Ce tableau donne cependant une idée un peu trompeuse, dans la mesure où la phase la plus coûteuse n'est pas l'exploitation, mais la perforation : les tarifs pratiqués dans notre équipe permettent de faire perforer et corriger 1 500 cartes par un vacataire pour une somme de 1 000 F. Mais le chercheur individuel dépourvu de crédits peut perforer lui-même, comme je l'ai indiqué tout à l'heure : c'est d'ailleurs ce qu'a fait Madame Mireux pour le texte de Guibert de Nogent, c'est ce que j'ai fait moi-même pour le Governance of England de Sir John Fortescue. De plus, même si le vacataire a vérifié son travail, une seconde vérification s'impose : à l'E.R.A. 713, ce travail est accompli par Madame Millet.

B

Quelles sont les possibilités offertes par ALINE ? Prenons l'exemple le plus simple, c'est-à-dire celui d'un texte sans précodage. Voici tout d'abord la façon dont se présente le texte :

Exemple 1

ANALYSE DU TEXTE GUIBERT DE NOGENT

| N.CAR. | REFERENCE | TEXTE |
|---|---|---|
| 1 | 1 1 1 | *DOMNO ET PATRI SANCTI SIMPHORIANI ABBATI ODONI , GUITBERTUS |
| 2 | 1 1 2 | DEI GENITRICIS UTINAM MINISTER AC SERVUS , LETOS IN DEI |
| 3 | 1 1 3 | SERVITIUM HABERE PROCESSUS . * |
| ... | | ... |
| 14 | 1 1 14 | QUAM ADORSUS FUERAM ORATIONE LIBELLULUM. IN SEQUENTI RE QUAM |
| 15 | 1 1 15 | DICERE DEVOVERAM JAM LIBATA , DUM DE DOMINI CORPORE SERMO |
| 16 | 1 1 16 | INCIDIT , TOTUM ETIAM NUNC QUOD LOQUI SUPER DENTE CEPERAMUS |

On voit ici comment se présente le texte : chaque ligne correspond à une carte, qui consiste donc en une zone de référenciation et en une zone texte. On note aussi que certains signes spéciaux ont été introduits - en l'occurrence par Madame Mireux - pour délimiter des titres ou des extensions du texte (étoile,< >). Pour que la machine puisse correctement analyser le texte, il lui faut donc préalablement lire un certain nombre de cartes de commandes ; par exemple

```
O    E
S    . : ; ? '
F    03 10 17 80
T    GUIBERT DE NOGENT
```

La carte O est une carte d'option : l'option E est celle qui demande un comptage des cartes, des formes, des mots et des séparateurs ; l'option A permettrait de redéfinir l'alphabet. La carte S donne la liste des séparateurs. La carte F indique le format selon lequel le texte a été perforé : par exemple, le texte de Guibert de Nogent est perforé sur les cartes de la colonne 17 à la colonne 80, les références occupant les colonnes 3 à 10. La carte T est la carte titre.

Second type de produits : les dictionnaires. Voici ici deux exemples tirés d'un même texte, The III Consideracions : tout d'abord un dictionnaire alphabétique, puis un dictionnaire par ordre de fréquence.

Exemple 2

TEXTE THE III CONSIDERACIONS

DICTIONNAIRE DES FORMES SELECTIONNEES
**********

| FORME | FREQUENCE | FREQUENCE RELATIVE | FREQUENCE CORRIGEE | STATUT |
|---|---|---|---|---|
| ***** | ********* | ********** | ********** | ****** |
| A | 128 | .0103 | .0103 | 0000 |
| AEASHT | 1 | .0001 | .0001 | 0000 |
| AEAYSHT | 1 | .0001 | .0001 | 0000 |
| AEIDE | 4 | .0003 | .0003 | 0000 |
| AEIDING | 1 | .0001 | .0001 | 0000 |
| AELAME | 1 | .0001 | .0001 | 0000 |
| AELE | 1 | .0001 | .0001 | 0000 |
| AELEMENTIS | 1 | .0001 | .0001 | 0000 |
| AECUTE | 14 | .0011 | .0011 | 0000 |
| AECVE | 22 | .0018 | .0018 | 0000 |

Exemple 3

TEXTE THE III CONSIDERACIONS

DICTIONNAIRE DES FORMES SELECTIONNEES

| FORME | FREQUENCE | FREQUENCE RELATIVE | FREQUENCE CORRIGEE | STATUT |
|---|---|---|---|---|
| AND | 1089 | .0875 | .0875 | 0000 |
| THE | 605 | .0487 | .0487 | 0000 |
| OF | 507 | .0408 | .0408 | 0000 |
| TO | 320 | .0257 | .0257 | 0000 |
| THAT | 310 | .0249 | .0249 | 0000 |
| IN | 275 | .0221 | .0221 | 0000 |
| HIS | 267 | .0215 | .0215 | 0000 |
| HE | 247 | .0199 | .0199 | 0000 |
| IS | 216 | .0174 | .0174 | 0000 |
| AS | 154 | .0124 | .0124 | 0000 |

Ces dictionnaires donnent non seulement une liste, mais aussi des indications quantitatives. Outre la fréquence (nombre d'occurrences), on a ici la fréquence relative (en fonction de la longueur du texte). Mais il est possible d'obtenir une fréquence relative corrigée, en fonction du statut du mot. Dans l'exemple donné, le statut de tous les mots étant identique, la fréquence relative et la fréquence corrigée sont bien sûr semblables ; cependant, il peut être intéressant de distinguer entre les mots grammaticaux et les autres, la fréquence des premiers étant si importante qu'elle peut introduire des distorsions dans les indices, distorsions qui peuvent avoir une signification plus stylistique que sémantique ; de même, chez certains auteurs, le recours à des citations latines est si important qu'il peut être utile de recalculer un indice de fréquence corrigé qui ne tienne pas compte des mots latins mais concerne uniquement les mots français, ou, le cas échéant, anglais.

Troisième sortie, la localisation des formes "sélectionnées". Effectivement, en fonction du statut attribué aux mots, il est possible, par exemple, de laisser de côté les références des mots grammaticaux, peu utiles à l'historien, et qui, étant donné les fréquences élevées de ce type de mot, reviennnent cher. Dans le cas présent, tous les mots ayant le même statut, ils sont tous sélectionnés.

Exemple 4

TEXTE ACTIVE POLICY OF A PRINCE

LOCALISATION DES FORMES SELECTIONNEES

| | |
|---|---|
| AB | /024 08/024 11/ |
| ABHOMINABLE | /034 15/ |
| ABIDE | /016 35/018 15/030 30/ |
| ABIDYNG | /026 07/ |
| ABOUTE | /026 03/ |
| ABOVE | /037 14/ |
| ABYDY | /017 27/ |
| ACCEPTABLE | /034 17/ |
| ACCEPTYNG | /014 21/ |
| ACCOMPLICE | /021 14/ |

La localisation des formes, dans la mesure où elle donne la référence de toutes les formes dans le texte est déjà utile "en soi" à l'historien. Mais elle est surtout à la base du processus qui permet de procéder à la lemmatisation, donc de passer des formes au mot : cette lemmatisation peut se faire par regroupement, par différenciation ou par hyphénation. Dans le cas d'ALINE elle est semi-automatique, c'est-à-dire qu'il faut réintroduire des cartes indiquant au programme que, pour trois formes identiques contenues dans un texte donné, la première et la

troisième sont des réalisations d'un verbe, la seconde est celle d'un substantif (par exemple : je porte, la porte, je porte). La reconnaissance n'est donc pas ici automatique : une des raisons en est que les membres de notre équipe travaillent sur des textes en plusieurs langues, et que toute reconnaissance automatique passant par l'introduction en mémoire de tests fondés sur les processus grammaticau propres à chaque langue, l'alourdissement du programme aurait été considérable (12)

Le dernier produit fourni par ALINE est le dictionnaire de concordances : voici un extrait de celui tiré de la première version du sermon parlementaire de John Russel (13).

Exemple 5

* BELYE

SLOWTHFULLE PARTE OF THE BODYE , AND DENYE THE PROUYSYON OF SYCHE NECESSARYE FOODE AS THE STOMAKE CALLETH FOR , HYT MIGHT SONE HAPPE , THAT FAYLYNGE THE BELYE FOR LAKE , THE GUTTES AND INTESTINES COMPRESSED AND SHUT BY DRYNESSE , ALLE THE OTHER MEMBRES SHOLLD NEDES PERYSHE TOGEDYR .

*** BEST

* POLITICORUM QUOD-INGENUITAS-EST-VIRTUS-ET-DIUITIE-ANTIQUE , NOBELESSE I VERTU AND AUNCIENNE RICHESSE , HIT SEMETHE THAT CONUENIENTLY SYCHE FERME NESSE OF GROUND AS MAY BE BEST FOUND IN THYS WORLD , OUGHTE TO BE APPROVE VN TO THE LORDES AND NOBYLLE MEN , AND SOO THEY TO BE SYGNIFYED AND VNDER STONDED BY THAT SOO

. HYT YS COMENLY SEYD , THE BEST WOLLE SAUF HYT SELF .

On remarquera que les concordances sont ici données d'une ponctuation forte à l'autre. Les sorties de concordances coûtant cependant cher, il est possible de moduler la taille des concordances : par exemple, cinq mots avant le mot vedette, et cinq mots après. C'est évidemment ce dictionnaire de concordances qui est pour l'historien le meilleur outil d'exploitation sémantique d'un texte.

En conclusion, je signalerai qu'ALINE est un package : par conséquent, une fois connue la procédure d'appel de la bande sur laquelle est catalogué ALINE, une fois enregistré le texte sur lequel l'utilisateur compte travailler, il suffit d'introduire en machine outre les cartes JOB usuelles, quelques cartes commandes. Nous espérons que d'ici trois mois ALINE sera accessible librement à tous les utilisateurs qui le désireraient. Un manuel d'utilisation sera ensuite préparé et il sera normalement disponible à la fin de l'année.

J.-Ph. Genet

Notes

(1) Voir Association for Literary and Linguistic Computing. Bulletin, I(1) et III(2
(2) Voir P. BRATLEY, S. LUSIGNAN et F. OUELETTE, JEUDEMO : A Text-Handling System, dans J.L. MITCHELL, Computers in the Humanities, Edinburgh, 1974, p.234-249.
(3) O.R. SMITH, GENDEX : GENeral DEXer of Words with Context : A Concordance Generator, Computer Studies, III, 1970-2, p. 50-53.
(4) R.L. VENEZKY, N. RELLES, L. PRICE, LEXICO : a system for Lexicographic Processing, Computer and Humanities, 11, 3, Mai-Juin 1977, p. 127-137.
(5) Les travaux du L.A.S.L.A. sont présentés dans la Revue de l'Organisation Internationale pour l'Etude des Langues Anciennes par Ordinateur ; pour ceux du CETEDOC, on trouvera une bibliographie dans Computers and Medieval Data Processing, VIII (2), Oct. 1978, p. 38-58.

(6) Une excellente démonstration des possibilités de ces programmes est fournie dans Tracts de Mai 1968. Mesures de vocabulaire et de contenu, Paris, 1975.
(7) Un algorithme d'étude sémantique dont la programmation était initialement prévue a du être provisoirement laissé de côté : cf. J.-Ph. GENET, Ordinateur, Lexique, Contexte, dans L. FOSSIER, A. VAUCHEZ et C. VIOLANTE, Informatique et Histoire Médiévale, Rome, 1977, p. 297-317.
(8) Texte en cours d'édition par Madame Marie-Danielle MIREUX.
(9) ed. M. BATESON, George Ashby's Poems (Early English Text Society, E.S.,vol.76),1899.
(10) ed. J.-Ph. GENET, Four English Political Tracts of the Later Middle Ages, Camdem 4th. Series, vol. 18, Londres, 1977, p. 180-209.
(11) ibidem, p. 5-19.
(12) Les procédures de lemmatisation semi-automatique propres à ALINE sont exposées dans J.-Ph. GENET, F. HUCHER, J. MONDELLI, E. VALENSI, Un programme de traitement automatique des textes : ALINE, Bulletin du Centre d'Analyse du Discours de l'Université de Lille III, 1974, p. 96-121.
(13) Ed. S.B. CHRIMES, Constitutional Ideas in Fifteenth Century England, Cambridge, 1936.
(14) Ce sera chose faite dans deux à trois mois ; le programme sera accessible de Paris I, et peut-être même directement du C.I.R.C.E.

## Appendice

## La Cartothèque de l'E.R.A. 713

a) Textes anglais

1. J. WYCLIF : on the seven deadly sins (J.W., Select English Works, ed. ARNOLD, III, p. 119-167).
2. An. : Tractatus de Regibus (Four English Political Tracts, ed. GENET, Camden 4th. Series, 18, 1977, p. 5-19).
3. Th. HOCCLEVE : The Regement of Princes (ed. F. FURNIVALL, E.E.T.S., Extra Ser. 61, 1892).
4. An. : The III Consideracions (Four English Political Tracts, p.180-219).
5. J. FORTESCUE : The Governance of England (ed. Ch. PLUMMER, Oxford, 1896).
6. W. WORCESTER : The boke of noblesse (ed. J.G. NICHOLS, Roxburghe Club, 1860).
7. J. RUSSEL : Three versions of the Parliamentary Sermon (ed. S.B. CHRIMES, English Constitutional Ideas..., Cambridge, 1936).
8. G. ASHBY : The Active Policy of a Prince (ed. M. BATESON, E.E.T.S., Extra Ser. 76, 1879).
9. E. DUDLEY : The Tree of Commonwealth (ed. D.M. BRODIE, Cambridge, 1948).
10. An. : Mum and the Sothsegger (ed. M. DAY et R. STEELE, Orig. Series, 199, 1936).
11. J. FORTESCUE : diverses petites oeuvres en anglais (ed. PLUMMER et ed. CLERMONT).

En préparation :

12. Th. ELYOT, The Boke Named the Governour (ed. H.S.C. CROFT, Londres, 1883).
13. J. YONGE, The Governaunce of Princes (ed. R. STEELE, Extra Ser.74,Londres, 1898).
14. An., /Version anglaise du Secreta Secretorum de University College MS. 85/, (ed. MANZALAOUI, E.E.T.S.).
15. An., The Libelle of Englyshe Polycye (ed. G. WARNER, Oxford, 1926).

b) Texte français (réalisations supervisées par Madame Claude GAUVARD)

16. J. GERSON : Discours Vivat Rex (in J. GERSON, Oeuvres complètes, ed. GLORIEUX VII, p. 1137-1185).
17. J. GERSON : Discours sur l'unité de l'Eglise (ibid., p. 1093-1100).
18. J. GERSON : Discours au roi pour la réconciliation (ibid., p. 1100-1123).
19. Ch. de PISAN : Le livre de la paix (ed. WILLARD).
20. 30 lettres de rémission du registre JJ 143 des Archives Nationales (transcrites par Mademoiselle Verduron).
21. 30 lettres de rémission du registre JJ 150 des Archives Nationales (transcrites par Mademoiselle Brès).

c) Textes latins

22. De Sanctis et Pignoribus Sanctorum (en dépôt de Madame Mireux).

## UN PEU DE TECHNIQUE ....
## PLAIDOYER POUR LA CARTE PERFOREE.

Les quelques remarques que je me propose de présenter ici n'ont pas l'ambition de constituer l'étude approfondie d'un moyen de saisir les données en Sciences humaines : sur ce point, il convient de se reporter à l'article de M. Gian Piero ZARRI, intitulé "Quelques aspects techniques de l'exploitation informatique des documents textuels : saisie des données et problèmes de sortie" (1), qui fournit une étude comparative des différents matériels existant sur le marché (ou, du moins, sur le marché tel qu'il se présentait en 1975, car ce marché, remarquait l'auteur, est en "état d'ébullition perpétuel"). Ce témoignage est uniquement fondé sur une certaine pratique, acquise au sein du Centre d'Histoire Juridique, qui utilise la carte perforée, à l'exclusion de tout autre moyen ; il ne peut donc être le fruit de comparaisons.

Dominée par la contrainte spécifique aux Sciences humaines - la grande quantité de données -, la saisie telle que nous l'effectuons (saisie des données et leur correction avant l'entrée en machine), présente deux traits originaux qui méritent d'être soulignés :

- la saisie est faite à partir d'un bordereau qui est utilisé à une double fin : le bordereau vierge sert de grille à l'analyse juridique d'une décision judiciaire (arrêt du Parlement de Paris au XIVe siècle) ; le bordereau rempli sert à la perforation des cartes. Ces deux étapes sont effectuées par la même personne, en l'occurrence l'analyste, car le bordereau utilisé ne correspond pas aux normes du bordereau de perforation traditionnel et ne peut donc être remis à une perforeuse professionnelle.
- Pour préciser l'importance de la quantité des données, on peut dire qu'en moyenne chaque document est représenté par une quarantaine de cartes perforées (une quinzaine de cartes pour les actes les plus courts, 70 à 80 pour les plus longs).

Qualifiée support "primaire" de transcription (G.P. Zarri, p. 402), la carte perforée est l'ancêtre des moyens de saisie, puisqu'elle existait avant l'ordinateur Loin de lui conférer des lettres de noblesse, cette longévité lui a attiré, depuis longtemps déjà, de sévères critiques qui l'ont conduite à un déclin bien amorcé, que d'aucuns disent inéluctable. En dépit des inconvénients qu'elle présente, quelques avantages - et des plus décisifs - sont à mettre à son actif et justifient, pour nous, son emploi contre vents de la mode et marées de nouveaux matériels.

Les inconvénients :

- Considérons directement le résultat de la perforation, car c'est là le moindre défaut. Ce résultat, c'est du poids et de l'encombrement : les cartes sont lourdes ; les boîtes remplies de ces cartes sont volumineuses ; ce double handicap rend leur transport difficile et onéreux.
- Un inconvénient technique, qui peut paraître dérisoire, mais dont seuls ceux qui ont passé des heures devant une perforatrice mesureront toute l'importance, tient à l'impossibilité dans laquelle on se trouve de voir les caractères que l'on vient de perforer jusqu'à ce que les six suivants l'aient été à leur tour. Ce défaut de conception impose de fournir une attention soutenue.
- Enfin, et c'est là le plus grave, soulignons la lenteur des opérations de correction après contrôle des cartes éditées sur listing, lenteur qui résulte des manipulations imposées alors à l'utilisateur : il faut, soit sortir la mauvaise carte et la reproduire en modifiant le caractère erroné, soit perforer une nouvelle carte et la mettre à la place de la carte fautive. L'opération se rèvèle encore plus lourde s'il s'agit d'un texte continu et si la correction ou l'addition de mots entraînent des décalages en chaîne qui vont imposer une seconde perforation de plusieurs cartes.

---

(1) Voir L. FOSSIER, A. VAUCHEZ et C. VIOLANTE, Informatique et Histoire Médiévale, Rome, 1977, p. 399-413.

Les avantages :

- La perforation de cartes ne nécessite pratiquement aucun apprentissage et, dans notre cas (qui n'est pas isolé en Sciences humaines), l'analyste peut s'en charger sans difficulté. Précisons, cependant, que la vitesse ne sera acquise que par une longue pratique.
- Ce système de saisie est une façon simple et commode de charger un fichier pour la première fois.
- Les cartes constituent un bon système d'archivage, permettant de conserver le fichier initial sans frais (à condition toutefois de disposer de place pour le stockage) et d'y accèder sans recourir à un moyen technique.
- La correction ponctuelle est facile quand l'erreur est décelée immédiatement après la perforation de la carte, la possibilité de reproduire une zone à conserver permettant une modification rapide.
- Enfin, l'avantage le plus déterminant est le faible prix de revient. A la différence de l'emploi de cartes ou bandes magnétiques, disques souples, voire lecteurs optiques, terminaux conversationnels et autres matériels sophistiqués, l'utilisation de la carte perforée aboutit à des prix défiant toute concurrence.

Bernadette Auzary

## VOYAGE AU CENTRE DE.... POITIERS.

Le groupe de recherche de Poitiers sur l'analyse informatique des textes, a commencé à fonctionner en 1974. Il regroupe les enseignants des 3 U.E.R. de Lettres et Sciences Humaines de l'Université de Poitiers qui poursuivent des travaux de recherche sur les textes à partir de l'informatique.

I - PRESENTATION DES RECHERCHES ENTREPRISES AU SEIN DE L'EQUIPE

- Epigraphie médiévale : Monsieur Robert Favreau, professeur d'Histoire médiévale à l'Université de Poitiers et Monsieur Jean Michaud, attaché à l'ERA 441 ont mis sur ordinateur les inscriptions métriques médiévales pour l'ensemble du territoire. Le même travail a été mené pour les épitaphes du VIIIe au XIIIe siècle. Ces fichiers ont abouti à l'établissement de concordances permettant l'étude du formulaire, du vocabulaire et, dans certains cas, à la reconstitution d'inscriptions métriques lacunaires. Ce travail a été largement utilisé pour l'établissement des notices publiées dans les fascicules I, 3 et II du "Corpus des Inscriptions Médiévales". Sont actuellement en cours de traitement les inscriptions en prose des VIIIe-XIIIe siècles, autres que les épitaphes. En outre, un fichier des sources "littéraires" épigraphiques vient d'être mis en chantier, à partir des textes recensés dans l'index épigraphique de la Patrologie latine.
- Etude de la langue espagnole médiévale : Réalisé par Monsieur René Pellen, maître-assistant à l'Université de Poitiers, l'enregistrement du poème espagnol du Cid a permis, grâce aux sorties d'index et de concordances et à l'exploitation des fichiers créés par PECI grâce à de nombreux programmes complémentaires, d'étudier plusieurs aspects de la langue du document. Une analyse graphémique et morphologique est en préparation en vue de le dater (la date estimée est irrecevable) et de le localiser (l'origine castillane est contestable).

  Dans la perspective beaucoup plus large d'établir une banque de données de l'Espagnol médiéval, le rassemblement et la confrontation des vocabulaires de trois textes du XIIIe siècle : El Poema de mio Cid, Los Milagros de nuestra Senora de Berceo, la General Estoria, permettront d'établir une partie manuellement un "dictionnaire-noyau" de l'ancien Espagnol qui servira de base pour une lemmatisation automatique des textes étudiés ultérieurement.
- Lexicologie du vieux-russe : Dans le cadre d'une étude du vocabulaire politique et social à partir des sources narratives, Monsieur J. P. Arrignon a procédé à l'enregistrement sur l'ordinateur du Sermon d'Hilarion et s'attache plus particulièrement à montrer la filiation de l'idéologie politique entre les textes byzantins d'Eusèbe de Césarée (Discours en l'honneur du 20e anniversaire du règne de l'Empereur Constantin) et de Photius (Introduction à l'Epanagogé), ainsi qu'à l'étude des co-occurrences.

- Sources narratives et lexicologie médiévale : Dans le cadre d'une étude lexicologique du vocabulaire politique et social des biographies royales et princières d'Eginhard à Rigord, Madame Elisabeth Carpentier, maître de Conférence d'Histoire médiévale à l'Université de Poitiers, a procédé à l'enregistrement de la Vie de Louis-VI le Gros de Suger, et de la Vie de Robert le Pieux d'Helgaud de Fleury. A partir des concordances et index établis automatiquement, elle a pu faire une analyse de fréquences sémantiques, menée dans le cadre d'un séminaire de 3e Cycle. Par ailleurs, suivant les mêmes méthodes, elle a entrepris l'étude de la chronique locale de Pierre de Maillezais.

  D'autres entreprises ont vu le jour dans ce centre, mais elles relèvent de l'époque moderne ou contemporaine. (Lexicologie de la prose russe au XVIIIe siècle, étude de la poésie russe contemporaine, du roman anglais moderne).

II - ASPECTS TECHNIQUES

- Les programmes PECI permettent d'éditer divers index et des concordances, partielles ou complètes, avec des contextes plus ou moins larges, à partir de textes dont toutes les formes et occurrences sont répertoriées.

  Ils éditent également des bordereaux de lemmatisation pour préparer la procédure automatique de lemmatisation (cf. infra). Le chercheur lemmatise manuellement chaque nouvelle forme et la décrit.

  Les programmes PECI sont conçus comme une suite de modules indépendants et s'articulent en blocs (1, 2 ou 4 programmes) ; chaque bloc, sauf le premier, constitue une option ; un même programme peut comporter, en outre, plusieurs options.

- Les programmes LEMAN (ou de "LEMmatisation et ANalyse") sont un prolongement de PECI. A partir d'un enregistrement des bordereaux de lemmatisation, ils éditent actuellement le Dictionnaire lemmatisé des formes et des références d'un texte. Ils éditent éventuellement le manuscrit de ce Dictionnaire pour publication en format 21 x 29,5. Les programmes sont en cours de normalisation pour que les formats standards soient les mêmes que ceux de PECI.

- Le Centre de Calcul : Les traitements informatiques sont effectués au C.I.C.U.P., service commun de l'Université, implanté dans les locaux universitaires : 40 av. du Recteur Pineau, 86022 POITIERS CEDEX. Tél. (49) 46-26-54 et 46-27-80.

  ◦ Le centre dispose d'un matériel de moyenne puissance :

  . 1 ordinateur IRIS 45
  . 1 unité centrale 256 K Octets
  . 3 unités disque : 50 millions d'octets/unité
  . 2 unités bandes magnétiques : 9 pistes, 1600 BPI
  . 2 imprimantes : 900 lignes/mn, 600 l/mn
  . 1 lecteur de cartes : 600 cartes/mn
  . 2 perforatrices et 1 vérificatrice dont 9 sont à la disposition des utilisateurs.

  ◦ Les tarifs appliqués par le centre diffèrent selon que les utilisateurs sont universitaires ou non. Pour les universitaires :

  . Coût d'une heure-calcul : 150,00 F
  . Coût d'une carte lue : 0,0030 F
  . Coût d'une ligne imprimée : 0,0030 F
  . Chargement d'une unité bande ou disque : 10,00 F

- Traitements effectués pour des chercheurs extérieurs. Les programmes ne peuvent actuellement pas être implantés ailleurs qu'à Poitiers et le centre de calcul n'accepte pas de faire des traitements sur mesure. Les chercheurs extérieurs qui désireraient établir des index et des concordances peuvent prendre contact, soit avec M.C. SIREDEY, Ingénieur au C.I.C.U.P. (49. 46-27-80), soit avec M. R. PELLEN (U.E.R. des Lettres et des Langues, 49. 46-25-75, poste 33 ou au C.I.C.U.P.). Pour chaque traitement, une étude préalable avec le chercheur essaie de dégager les meilleures options de PECI en fonction des recherches envisagées. Quand le projet est défini, le C.I.C.U.P. établit un devis et ouvre un compte au chercheur, s'il veut donner suite à son projet. L'exploitation a lieu dès que le compte est alimenté. Délai de réalisation : selon l'importance des données et des options, de quelques jours à deux ou trois semaines.

  Deux chercheurs extérieurs à l'université ont déjà utilisé les programmes PECI et tous les traitements demandés seront effectués dans la mesure du possible.

C. Bourlet

## L'INFORMATION.

A/ Nous proposons ici une liste de bulletins, revues, périodiques susceptibles de contenir des articles (ou, plus généralement, des renseignements) utiles à l'historien utilisateur de l'informatique. Nous avons volontairement exclu de ce panorama toute publication de caractère exclusivement théorique ; pour chaque élèment de la liste nous donnons l'adresse du responsable (ou de l'un des responsables) de la publication.

1) Signalons en tout premier lieu, parce que la revue intéresse spécialement le médiéviste :
Computers and Medieval Data Processing (CAMDAP)
c/o M. le Professeur Serge Lusignan - Institut d'études médiévales - Université de Montréal - C.P. 6128, Succursale "A" - MONTREAL, P.Q. H3C 3J7 - Canada.
Ce bulletin contient, entre autres, un recensement périodique à l'échelle mondiale de tous les projets en cours dans le domaine "Informatique et Moyen-Age", et une bibliographie très détaillée sur le même sujet. Gratuit.

2) ACM SIGLASH Newsletter
c/o Mrs. Christine Montgomery - Operating Systems, Inc., 21031 Ventura Bvld. - WOODLAND HILLS, Ca. 93164 - U.S.A.
Le SIGLASH est le "Special Interest Group for Language Analysis and Studies in the Humanities" de l'ACM ("Association for Computing Machinery").

3) ARITHMOI
c/o Professor Richard E. Whitaker - Central College - PELLA, Iowa 50219 U.S.A.
Lancé par le Prof. Whitaker comme équivalent pour les études bibliques de Calculi ou de CAMDAP (voir plus bas), il semble avoir cessé de paraître ces derniers temps. Gratuit.

4) Bulletin of the ALLC
c/o Mrs. Susan M. Hockey - Oxford University Computing Laboratory - 19 Banbury Road - OXFORD - England.
Revue éditée par l'ALLC ("Association for Literary and Linguistic Computing").

5) CALCULI
c/o Professor Stephen V.F. Waite - Kiewit Computation Center - Dartmouth College - HANOVER, New Hampshire 03755 - U.S.A.
Bulletin d'information dédié principalement aux applications des ordinateurs dans les études classiques. Gratuit.

6) Computers and the Humanities
c/o Professor Joseph Raben - Queens college of the City University of New York - FLUSHING, New York 11367 - U.S.A.
Il s'agit de la revue la plus connue et, peut-être, la plus prestigieuse dans le domaine "Informatique et Science de l'Homme".

7) Historical Methods
c/o Professor Reginald Baker - University of Pittsburgh - Faculty of Arts and Sciences - Department of History - PITTSBURGH, P.a. 15260 - U.S.A.
Ce bulletin s'est transformé de simple bulletin de liaison en véritable revue largement ouverte aux problèmes et aux méthodes concernant l'emploi de l'informatique en histoire.

8) Informatica e diritto
c/o Dr. Costantino Ciampi - Istituto per la Documentazione Giuridica del C.N.R. - Via Panciatichi, 56/16 - 50127 FIRENZE - Italia.
La revue accepte toute contribution du type "Informatique et Traitement des textes".

9) Informatie Nederlandse Lexicologie
c/o Dr. Félicien de Tollenaere - Beatrixlaan 7 - WARMOND - Pays-Bas.

10) Informatique et Sciences Humaines
c/o M. Philippe Cibois - CNRS - LISH - 54, boulevard Raspail - 75270 PARIS Cedex 06

11) Journal of the Association for Computational Linguistics
c/o Dr. Donald E. Walker - SRI International - MENLO PARK, Ca. 94025 U.S.A.
Même si cette revue a un caractère plus "théorique" que, par exemple, "Computer and the Humanities", elle n'exclut pas la publication d'articles concernant des applications pratiques dans le domaine de l'emploi des ordinateurs pour le traitement des textes. A partir de cette année, le "Journal" va abandonner l'ancien système, très malcommode, de publication sur microfiche, pour se transformer en revue "normale", sur papier.

12) Meroitic Newsletter
c/o M. le Professeur Jean Leclant - 77, rue Georges Lardennois - 75019 PARIS.
Ce bulletin contient parfois des articles concernant l'emploi des ordinateurs pour les études d'épigraphie méroïtique. Gratuit.

13) Newsletter of the ALLC Software Specialist Group
c/o Dr. P.J. Wolfangel - Institut für Deutsche Sprache - Abt. LDV/SuC Postfach 5409 - 6800 MANNHEIM - Allemagne Fédérale.

14) Programmation et Sciences Humaines (PSH)
c/o M. Michael Hainsworth - CNRS - LISH - 54, boulevard Raspail - 75270 PARIS cedex 06.
Le titre de cette revue veut souligner, par rapport par exemple à Informatique et Sciences Humaines", un intérêt plus pragmatique pour les problèmes liès à l'utilisation concrète des ordinateurs dans les Sciences de l'Homme.

15) Revue du LASLA
c/o Monsieur le Professeur Louis Delatte - L.A.S.L.A. - 110, boulevard de la Sauvenière - 4000 LIEGE - Belgique.
Revue éditée par L.A.S.L.A. ("Organisation Internationale pour l'étude des langues anciennes par ordinateur").

16) Sprache und Datenverarbeitung
c/o Professor Winfried Lenders - Institut für Kommunikationsforschung und Phonetik - Universität Bonn - Adenauerallee 98a - BONN - Allemagne Fédérale.
Voir note à propos de "Informatica e diritto".

17) T.A. Informations
c/o M. André Deweze - St Vincent de Mercuze - 38660 LE TOUVET.
Revue internationale du traitement automatique du language - bulletin semestriel de l'A.T.A.L.A. ("Association pour le traitement automatique du language").

18) Travaux de Lexicometrie et de Lexicologie Politique
c/o Mme Gabrielle Drigeard - "Lexicologie et textes politiques" - E.N.S. de Saint-Cloud - 2, avenue du Palais - 92211 SAINT-CLOUD.
Bulletin de l'U.R.L. "Lexicologie et textes politiques" de l'Institut de la Langue Française du C.N.R.S.

B/ **D'un colloque à l'autre. Quelques rencontres passées...**

Colloque International du C.N.R.S. sur "La pratique des ordinateurs dans la critique des textes".

Ce Colloque, organisé par MM. les Professeurs Glénisson, Irigoin, Marichal et Monfrin dans le cadre de l'Institut de Recherche et d'Histoire des Textes, s'est tenu à Paris du 29 au 31 mars 1978. Le secrétariat scientifique a été assuré par Mme Fossier et M. Zarri.

Le but du Colloque - après plus de dix ans d'utilisation de l'ordinateur dans le domaine de la critique des textes - était de faire le point des diverses méthodes mises en oeuvre dans plusieurs pays, de confronter les résultats obtenus et de dégager les grandes lignes des développements futurs de ces recherches. Ce souci semble avoir été partagé par un bon nombre de philologues et d'informaticiens car le Colloque qui réunissait vingt-deux conférenciers appartenant aux organismes scientifiques de sept nations différentes a été suivi par quatre-vingt chercheurs environ représentant douze pays. Deux Tables Rondes "Problèmes de sélection et d'utilisation des variantes" et "Le classement des manuscrits et son approche formelle" ont conclu les travaux. Les Actes paraîtront dans le courant de l'année dans la collection "Colloques Internationaux" des Editions du C.N.R.S.

Sans prétendre donner des jugements de caractère définitif, nous nous limiterons ici à quelques remarques qui semblent avoir été partagées par la majorité des assistants.

A propos de la polémique entre "méthodes statistiques" (voir communications de Mme Galloway et de MM. Berghaus et Griffith) et "méthodes non-statistiques" ("généalogiques", pour reprendre le terme proposé par M. Irigoin - voir communications de Dom Froger, MM. Dearing, Poole, Zarri etc.), il semble qu'un certain accord se soit dégagé sur l'utilité potentielle des méthodes statistiques pour une "répartition" préliminaire des manuscrits en familles dans le cas de traditions très riches. Une représentation suffisamment précise des relations réciproques entre les manuscrits ne pourra toutefois être obtenue que par le recours aux méthodes du deuxième groupe. A remarquer que certaines méthodes "généalogiques" (voir communication de MM. Najock et Zarri) semblent désormais permettre le traitement automatique, du moins partiel, des phénomènes de contamination.

En ce qui concerne l'utilisation de techniques automatiques pour le problème de la reconnaissance d'abord, et de la sélection ensuite, des variantes à utiliser comme base des opérations critiques, nous ne ferons ici que deux observations. Les techniques de collation automatique (dont les limites sont bien connues : voir, par exemple, l'inutilité de leur emploi pour les traditions romanes, où très souvent chaque manuscrit présente une version pratiquement différente du même texte) sont désormais utilisables par les philologues pour une opération de routine (la communication de Mme Gilbert a été extrêmement convaincante à ce propos). En revanche, la formalisation des critères implicites que le philologue utilise pour organiser les "lieux variants", pour choisir entre variantes "significatives" et "non-significatives" etc. ne semble pas encore suffisamment mûre (voir communications de MM. Duplacy et Heinemann), et le colloque n'a pas apporté de réponse à ce sujet.

Fifth International Symposium on "Computers in Literary and Linguistic Research".

Ce Colloque, parrainé par l'ALLC ("Association for Literary and Linguistic Computing") s'est déroulé à l'Université de Aston in Birmingham du 3 au 7 avril 1978 ; il s'agissait de la cinquième rencontre biennale de ce type tenue en Angleterre - la première avait été organisée à Cambridge en 1970.

Les participants ont assisté à quatorze sessions comprennant chacune trois/quatre communications ; nous indiquons ici les titres de ces sessions : "The Variety of Work I", "The Variety of Work II", "Input/Output", "Textual Criticism", "Authorship Attribution", "Software I", "Linguistic Ambiguity", "Software II", "Information Science", "Literary Statistics", "Lexicography", "Stylistc Analysis", "Structures", "New Departures". Les "Proceedings", comprenant un choix des communications les plus significatives, vont paraître avant l'été 1979 ; les personnes intéressées peuvent s'adresser à M. le Professeur D.E. Ager - Department of Modern Languages - The University of Aston in Birmingham - Gosta Green - BIRMINGHAM B4 7ET - England.

Nous nous limiterons à remarquer ici qu'un certain nombre de nouveautés intéressantes ont été évoquées alors que les rencontres précédentes s'étaient déroulées de façon très traditionnelle. Les techniques dites d'"Intelligence Artificielle" ont fait une apparition timide par le biais des communications de M. Cercone sur les réseaux sémantiques, de Mme Galloway sur l'analyse automatique des

textes narratifs, et de M. Zarri sur le projet RESEDA. D'autres communications "inhabituelles" ont traité de traduction automatique (M. Loh), de sémantique (MM. Clark et Rieger), de "Rhetorical Punctuation by Machine" (Mme Mastermann), etc. Les applications de techniques "statistiques" ont fait naturellement l'objet d'un grand nombre d'exposés (nous voulons signaler ici celui de l'équipe Saint-Cloud sur "Les co-occurrences : une nouvelle approche statistique"). Par contre, rares ont été les communications qui ont fait allusion aux techniques lexicographiques les plus habituelles (index et concordances), envisagées d'ailleurs surtout sous des aspects strictement techniques (constitution de "packages" par exemple, voir exposé de Mme Hockey).

7th International Conference on Computational Linguistics (COLING/78)

Le Congrès COLING/78 - organisé par l'"International Committee on Computationa Linguistics", dont le président est M. le Professeur B. Vauquois de l'Université Scientifique et Médicale de Grenoble - s'est tenu à l'Université de Bergen, en Norvège, du 14 au 18 août 1978. Il a été suivi par deux cents chercheurs environ qui ont écouté soixante et onze communications. Un certain nombre de Tables Rondes "officielles" (à propos de langages de programmation pour les humanités, de traitement des textes, de problèmes de terminologie etc.), et de réunions informelles, ont complété un programme déjà très riche.

Si le Symposium ALLC du mois d'avril permettait de percevoir certaines tendances nouvelles dans le domaine "Informatique et Sciences de l'Homme", ces tendances ont profondément marqué le COLING/78, de manière presque caricaturale. En écoutant les communications en effet, on avait parfois l'impression de se trouver à un congrès d'Intelligence Artificielle plutôt que de Linguistique. L'impression était évidemment renforcée par la présence au Congrès d'un groupe nombreux de vedettes américaines de cette discipline, l'"A.I.", maintenant tellement à la mode (MM. Carbonnell Jr., Joshi, Schubert, Steels etc.) ; du côté européen, nous n'oublierons pas de citer les communications très appréciées de Mme Schwind et de M. Pétöfi. Les exposés de type statistique - ou portant sur les techniques lexicographiques traditionnelles - ont été très rares, et une rencontre organisée hors-programme sur la préparation d'index pour les grands corpus textuels n'a eu qu'un succès mitigé.

Nous soulignerons enfin un dernier aspect intéressant de ce Congrès, la quantité (et la qualité) inhabituelle des communications ayant comme objet la traduction automatique (voir les exposés de Mmes Hauenschild et Kulagina, de MM. Boitet, Isabelle, Thouin etc.). Ce problème est en train de connaître un regain d'intérêt au niveau international, sous l'effet de plusieurs facteurs : succès commercial du système américain SYSTRAN ou, sur une plan plus "politique" lancement par la Communauté Européenne du projet EUROTRA.

... et à venir

Ecole Pluridisciplinaire de l'Institut de Recherche d'Informatique et d'Automatique (I.R.I.A.) sur "Informatique et Histoire" (Maison des Sciences de l'Homme, Paris, deuxième quinzaine d'octobre 1979). Pour informations, écrire à : M. Gian Piero Zarri - CNRS-LISH - 54 boulevard Raspail - 75270 PARIS Cedex 06.

G.-P. Zarri

et à ce propos

---

Seriez-vous interessé par une rencontre informelle d'historiens médiévistes qui pourrait être :

- Pour les utilisateurs, une confrontation de vues.
- Pour les "non-initiés", une information sur les applications en cours.

Cette réunion pourrait être réalisée, selon le point de vue adopté, soit dans le cadre du L.I.S.H. - Marseille, soit à l'occasion du stage pluridisciplinaire de l'I.R.I.A. (Cf. plus haut).

## COURRIER DES LECTEURS.

Cette rubrique ne peut évidemment être alimentée dès le premier numéro; les réflexions d'un utilisateur "chevronné" en tiendront lieu :

### UN ASPECT PARTICULIER DE L'INFORMATIQUE EN SCIENCES HUMAINES

### LA DOCUMENTATION HISTORIQUE

Les services rendus aux historiens par l'informatique sont aussi divers que riches de possibilités actuelles et futures. L'un d'entre eux mérite peut-être d'être signalé dès ce premier Courrier, car, s'il répond à des ambitions modestes, il a l'incontestable avantage d'une simplicité d'accès qui le met à la portée de bon nombre de chercheurs isolés ou d'équipes aux moyens limités : il s'agit de la constitution et de la gestion automatisées de fichiers documentaires.

Ces fichiers, destinés à fournir aux chercheurs les matériaux indispensables à la synthèse qu'ils projettent, sont le fruit de l'analyse patiente et aussi exhaustive que possible d'un ensemble de sources - imprimées ou manuscrites -, auxquelles peut s'adjoindre éventuellement la bibliographie relative à la masse documentaire exploitée. Le traitement soit des noms de personnes et de lieux, soit des données thématiques pose, certes, quelques problèmes méthodologiques spécifiques, mais comparables à ceux que maîtrise actuellement fort bien la science documentaire classique, dont les solutions pourront être facilement adaptées. La seule exigence nouvelle est liée à la nécessité de réduire et de codifier le vocabulaire conceptuel : le codifier, pour qu'analystes et utilisateurs puissent les uns, choisir, les autres, trouver aisément les mots-clefs qui, décrivant au mieux le document, permettront d'y recourir à bon escient ; le réduire, pour des raisons évidentes de rigueur scientifique (ce qui se conçoit bien s'énonce clairement ; un trop grand affinement dans l'analyse a souvent pour conséquence des inexactitudes, pour ne pas dire des erreurs) et de coût de traitement. Il faut donc établir des dictionnaires, ou, mieux, des thésaurus élaborés en fonction du fonds à exploiter.

Ces instruments de travail seront, pour l'analyste, le plus sûr des guides. Insistons sur le rôle, qui doit être bien défini, de cet analyste : il a pour mission de donner à d'autres "travailleurs scientifiques" les élèments de leur recherche personnelle. C'est à lui d'extraire des documents signalés ces élèments, de les expliciter, de les trier, de les organiser. Il ne convient donc pas qu'il reste passif devant un texte, dans l'espoir que le traitement automatique de la source considérée ("full text", établissement de concordances variées...) résoudra tous les problèmes. La masse, souvent écrasante, de la documentation à exploiter, les difficultés de son interprétation ne peuvent être ainsi dominées ; rappelons aussi que, dans la quasi-totalité des cas (et il ne s'agit pas que des sources médiévales), la notion qui sera qualifiée par un descripteur précis doit être dégagée du contexte, celui-ci devant s'entendre parfois de l'ensemble du document. (1)

Il va de soi que, au stade de l'exploitation des documents ainsi fournis au chercheur, celui-ci peut avoir le plus grand intérêt à enregistrer totalement, puis à traiter par telle ou telle de ces méthodes dont la mise à disposition du public est de plus en plus répandue, les textes diplomatiques, doctrinaux, narratifs dont le choix lui aura été rendu possible par le labeur de ceux qui ont trié et, en quelque sorte, catalogué la masse initiale. Ce travail préliminaire à toute synthèse valable, l'historien en sait l'importance aussi bien qu'il en ressent le côté ingrat ; les moyens informatiques actuels permettent de l'effectuer dans des conditions de sûreté et d'exhaustivité inégalables par les méthodes classiques.

Josette Metman

---

(1) "Le seigneur de.... avait promis de donner à sa fille, lorsqu'elle se maria..." ; deux descripteurs : "contrat de mariage" et - en fonction de l'ensemble du texte (détroit coutumier ; conventions spéciales...) soit "dot", soit "avancement d'hoirie".