Júlia Pajzs

Réalisation assistée par ordinateur de grands dictionnaires français et hongrois

L'idée d'un grand dictionnaire de la langue française du type "Oxford English Dictionary" fut émise pour la première fois en 1957. Le Trésor de la langue française contenant le vocabulaire utilisé entre 1780 et 1960 et qui constitue la première partie de cette entreprise, est le premier grand dictionnaire réalisé avec l'aide de l'ordinateur. S'inspirant de cet exemple, l'Académie des Sciences de Hongrie envisage d'éditer le dictionnaire de la langue hongroise de Gutenberg à nos jours. Nous nous proposons ici de dégager les principes qui guident ces deux travaux.

C'est en 1961 que commença, sous la direction du professeur Paul Imbs, le travail de collecte qui devait aboutir à la constitution de corpus. Saisi sur ordinateur, le corpus fut classé en listes de concordance comportant les occurrences, les références exactes et le contexte de chaque unité susceptible de figurer comme entrée. Le volume 13 du Trésor vient de paraître; l'ouvrage en comportera 16 ou 17 et devrait être terminé au début des années 90. Le dictionnaire historique contenant le vocabulaire des siècles précédents et qui sera mis en chantier par la suite, utilisera la même méthode. Les travaux assistés par ordinateur du grand dictionnaire hongrois ont commencé en 1985; nous en sommes actuellement à la phase de la collecte des données et la publication ne sera achevée qu'après 2000.

La lexicographie assistée par ordinateur se divise en cinq phases successives; constitution du corpus, saisie sur ordinateur, lemmatisation des mots repérés dans les textes, confection des listes de concordances, rédaction des articles. Pour constituer leur corpus, les lexicographes ayant travaillé avec des fiches lisaient des centaines d'ouvrages, pour y relever les emplois les plus intéressants des mots et pour consigner ceux-ci, avec le contexte dans lequel ils figuraient. Classées - manuellement - par ordre alphabétique, les fiches constituaient les entrées du dictionnaire. Au contraire, l'ordinateur au service du lexicographe saisit des textes cohérents, établit les listes de concordance et parmi les citations ainsi ordonnées, le lexicographe choisit celles qui lui serviront à rédiger les entrées. Autrement dit: alors que la méthode traditionnelle oblige le lexicographe à sélectionner des données à la fois dans la phase de la collecte et dans celle de la rédaction des articles, la lexicographie assistée par ordinateur ne recourt à la sélection que dans la dernière phase; cette sélection s'opère dans les listes de concordance. La méthode informatisée demande au lexicographe des procédés particuliers en ce qui concerne, notamment, l'utilisation des sources permettant la constitution du corpus: le nombre des données à saisir est considérablement plus élevé que dans le cas de la collecte traditionnelle fondée sur une présélection, par ailleurs, il faut veiller avec un soin particulier à ce que les textes saisis soient vraiment

représentatifs. Pour atteindre ce but, l'équipe française et l'équipe hongroise emploient deux méthodes radicalement différentes: la première a constitué un corpus d'environ 90 millions de mots qu'elle complète au fur et à mesure par les mots archaïques des siècles précédents et par les néologismes aujourd'hui, FRANTEXT, le corpus du Trésor compte environ 160 millions de mots. Les textes émanent de grands écrivains et sont toujours reproduits intégralement. Ont été sélectionnés ceux qui sont le plus souvent mentionnés dans les manuels d'histoire de la littérature. Quant aux textes qui ont servi de base pour constituer le corpus du grand dictionnaire de la langue hongroise, ils comptent environ 13-15 millions de mots utilisés de 1533 à nos jours. Les textes sont en général courts, de quelques centaines à quelques milliers de mots, quelquefois extraits d'ouvrages de grande étendue. A titre exceptionnel, certains ouvrages importants (comme la Bible de Vizsoly) sont intégralement saisis. Les textes retenus sont quelquefois dûs à des écrivains mineurs, mais leur étendue dépend de l'importance de l'auteur. Tous les textes ont été sélectionnés par les spécialistes des différentes périodes de la littérature hongroise. Cette méthode vise à obtenir la plus grande diversification avec le plus petit nombre possible de textes. L'une et l'autre équipe ont retenu des textes de spécialité à côté des textes littéraires et utilisent en dehors du corpus saisi sur l'ordinateur, des fiches manuellement établies. L'équipe française puise, à l'occasion, dans les quelques six millions de fiches de l'Inventaire général de la langue française; quant au dictionnaire hongrois, il utilise vraisemblablement les 5 millions de fiches réunies en vue de la réalisation des grands dictionnaires déjà publiés. (Les fiches auraient dû servir à la préparation du Grand dictionnaire de la langue hongroise, mais le matériel ainsi réuni ne paraissant pas assez homogène et fiable, on a préféré recommencer à zéro le travail de collecte, en utilisant, cette fois, les moyens offerts par l'informatique.) En dehors des textes, les ordinateurs ont saisi les entrées des dictionnaires interprétatifs unilingues: l'équipe française a d'ores et déjà enregistré toutes les entrées de tous les dictionnaires français de ce type - quant à l'équipe hongroise, elle a jusqu'à présent informatisé les entrées de deux dictionnaires: celles du Dictionnaire interprétatif et celles du Dictionnaire de fréquence de la prose littéraire que nous complétons à l'heure actuelle par celles du Petit Dictionnaire interprétatif.

Avant d'être saisis, les textes destinés au *Trésor* sont soumis à un certain nombre d'opérations préparatoires: codage, marquage, dans les pièces de théâtre, des fins de scènes, d'actes et de répliques - les textes hongrois ne comportent avant la saisie que les codes des références et les passages à omettre (figures, tableaux, citations en langues étrangères). Le codage (par exemple des titres, des notes, des mots étrangers et des noms propres) est effectué par l'ordinateur. Celui des lettres pourvues d'accents et des particularités dues aux archaïsmes des vieux textes est réalisé grâce à la combinaison de chiffres et de lettres, par exemple le code de l'accent aigu est 1 et la lettre **á** est codée **a1**, la lettre **é** - **e1**, etc. Les textes ainsi transcrits peuvent être, grâce à un programme de conversion, reproduits sous leur forme initiale, mais il nous a paru plus indiqué de la stocker sous cette forme, d'une part, à cause du grand nombre d'archaïsmes, d'autre part,

en raison des possibilités des divers ordinateurs sur lesquels nous travaillons: actuellement, nos collaborateurs travaillent sur des ordinateurs IBM XT/AT compatibles et (à domicile) sur des Commodores 64 - le traitement des données devra être effectué sur des ordinateurs puissants (dont la marque n'a pas encore été précisée).

Après saisie du corpus du *Trésor*, les mots figurant dans les textes ont été répartis selon leur catégorie grammaticale par des procédés semi-automatiques. Toutefois ces procédés n'ont par permis de distinguer entre homographes, en conséquence de quoi 70% seulement de mots ainsi répertoriés se sont retrouvés dans la catégorie qui était la leur. Pas de lemmatisation, mais en ce qui concerne les verbes, un dispositif a été créé permettant d'adjoindre aux radicaux verbaux les suffixes adéquats. Etablis à partir des infinitifs, les paradigmes complets ont été stockés dans une base de données, ce qui permet d'appeler toutes les formes verbales. Les listes de concordance peuvent ainsi être dressés, comprenant toutes les formes verbales suffixées. Les mots sont classés non seulement par ordre alphabétique, mais aussi selon la catégorie grammaticale des éléments qui les précèdent ou suivent immédiatement. (Par exemple, la liste de concordance du verbe *aimer* comportera toutes les occurrences de ce verbe suivi d'un nom, d'un adjectif, etc; ces mêmes occurrences sont ensuite groupées suivant les catégories grammaticales des mots qui précèdent immédiatement le verbe.) Les mots ayant été également classés par ordre alphabétique, le lexicographe dispose de deux listes de contenu analogue mais d'un ordonnancement différent. Dans les listes de concordance, toutes les occurrences d'un mot - à l'exception des fonctionnels - figurent avec leur contexte d'une ligne; les 90 millions de mots ordonnés selon deux principes différents occupent plusieurs salles dans le bâtiment mis à la disposition des rédacteurs du *Trésor*. Les mots les plus fréquents n'ont pas de listes de concordance complètes; les occurrences sont relevées dans un "quasi-corpus" comportant environ 20 millions de mots, soit une quantité de mots supérieure à celle du corpus hongrois intégral.

Une des particularités du grand dictionnaire de la langue hongroise consiste en la lemmatisation des mots figurant dans les textes, à l'aide d'un programme d'analyse morphologique automatique. Ce programme d'analyse cherche à séparer les mots figurant dans les textes, en radicaux et en suffixes, marque, chaque fois que cela est possible, les limites morphématiques et pourvoit chaque morphème d'une étiquette comportant, dans le cas des radicaux, la catégorie grammaticale du mot, le cas échéant, le code homonymique distinctif et la forme canonique, (telle qu'elle figure dans l'entrée du dictionnaire) au cas où cette forme diffère de la forme actualisée du radical (telle qu'elle figure dans le texte pris en considération). Dans le cas des affixes, l'étiquette est constituée par le code qui est attribué à chacun d'eux. Les procédés aboutiront à la constitution de listes de concordance ordonnées non seulement selon les mots, mais aussi selon les lexèmes figurant dans les textes. Ils nous permettront en outre d'entreprendre plusieurs types de recherche: les occurrences pourront être groupées suivant leur rection, suivant leurs propriétés syntagmatiques, etc. L'analyse mécanique étant

dans chaque cas, suivie d'une vérification "manuelle", les fautes et omissions pourront être corrigées.

L'analyse morphologique automatique peut être effectuée de différentes façons. Elle peut partir des suffixes (analyse *a tergo*) des radicaux, ou du mot intégral. L'algorithme que nous avons adopté commence l'analyse par le radical, ce qui nous permet d'indiquer, dans chaque cas, le terme "canonique" du mot (tel qu'il figure dans le dictionnaire), même si cette forme est différente de celle du radical. Pour analyser le mot correctement, il est nécessaire de savoir à quelle catégorie grammaticale il appartient; cette information nous sera fournie par l'inventaire des radicaux. Dans le cas des mots dérivés et composés, si le mot lui-même ne figure pas dans l'inventaire des radicaux, mais que ce dernier comporte les éléments qui le constituent, le programme est capable de reconnaître les limites des radicaux et des suffixes dérivatifs. Comme nous commençons toujours par identifier le radical le plus long susceptible d'être élargi à gauche, le programme permettra de retrouver dans l'inventaire des radicaux une partie des mots dérivés et composés. Aussi, leurs éléments ne sont signalés que si l'inventaire de radicaux comporte une limite morphématique.

Nous renonçons provisoirement à l'identification des homonymes ou homographes. Plus tard, un analyseur syntaxique, destiné à compléter notre batterie de programmes, devra nous permettre de les interpréter correctement, mais dans l'immédiat, cette tâche incombe à la post-édition. Le programme lui-même se contente d'en marquer d'un astérisque toute forme lexicale attirant ainsi l'attention du post-éditeur sur les cas qui posent problème.

Voici les principales étapes du programme analyseur.:

- Le mot prélevé dans le texte est-il identique à un lexème? Si oui, le lexème muni de son code indiquant sa catégorie grammaticale, et, éventuellement, de son code d'homonyme, est saisi. Dans le cas des homonymes, nous signalons par un * l'incertitude de leur identification.

- Si l'inventaire des radicaux ne comporte pas de lexème susceptible d'être le radical du mot prélevé dans le texte, nous reproduisons l'intégralité de ce mot, muni d'un *! qui signale que le radical ne figure pas dans l'inventaire ou qu'il s'agit d'une erreur.

- Si le radical figure dans l'inventaire (en cas de formes concurrentes, c'est la forme la plus longue qui est retenue), nous l'isolons dans le mot du texte et nous cherchons à identifier ce qui reste en recourant à l'inventaire des affixes. Une fois l'ensemble des affixes identifiés, nous nous demandons s'ils sont susceptibles de se combiner avec le radical en question. Si tel est le cas, nous retenons le radical suivi, entre parenthèses, des codes et, le cas échéant, le lexème auquel il correspond. Le programme prévoit le code des homonymes - si le mot en admet - et des affixes. (Voici, par exemple, la transcription ainsi conçue du mot *ablakot* (fenêtre + Acc.): ablak FN ot "Acc."; celui de *lovaknak* (aux chevaux) sera: lov «ló/FN» ak «PL» nak «DAT».

- Si le bloc des affixes ne figure pas tel quel dans l'inventaire des affixes, nous cherchons le suffixe le plus long figurant dans le bloc et nous l'isolons du reste. Nous cherchons ensuite les suffixes restants dans l'inventaire des affixes, parmi

ceux qui, en principe, peuvent figurer devant le suffixe déjà identifié. Cette opération sera répétée jusqu'à ce que l'étendue du bloc soit égale à zéro. Nous vérifierons ensuite si les codes des suffixes peuvent suivre le radical. Si la réponse est non, nous recommençons l'analyse en procédant à une autre segmentation du bloc. Si cette tentative se révèle infructueuse, nous chercherons dans l'inventaire des radicaux un autre radical possible et nous recommençons encore l'analyse à partir de ce radical. En cas d'insuccès, nous relevons le mot tout entier d'un texte, muni d'un *!

Les textes ainsi traités seront soumis à un nouvel examen: les mots marqués d'un * seront correctement interprétés et pourvus de leurs codes de catégorie grammaticale et d'homonymie. En ce qui concerne les éléments marqués d'un *!, nous vérifierons s'ils sont réellement des lexèmes déjà répertoriés. S'il s'agit de lexèmes "nouveaux", nous les introduisons à la fois dans l'inventaire et dans le texte "brut" traité par le programme. Il peut arriver qu'un lexème "nouveau" ne soit que le résultat d'une erreur de saisie. Il convient alors de corriger l'erreur et d'inscrire à la main la solution correcte.

Les listes de concordance complètes du texte analysé ne seront pas imprimées sur papier. Un programme destiné à interroger le corpus permettra non seulement d'appeler ces listes, mais aussi de repérer les cooccurrences de plusieurs mots ou morphèmes, de modifier, selon les besoins, l'étendue du contexte à retenir et l'ordonnancement de la liste. Il sera également possible d'interroger le corpus sur la fréquence des éléments qu'elle comporte et de fixer, en tenant compte de cette fréquence, le nombre approximatif des occurrences d'un élément quelconque à inclure dans la liste. Le programme d'interrogation et le texte analysé seront stockés sur disquettes optiques: les lexicographes pourront ainsi utiliser les citations - préalablement sélectionnées et ordonnées - nécessaires pour la rédaction des articles. Ces disquettes seront, dans la mesure du possible, diffusées auprès des chercheurs afin de rendre ainsi disponible les matériaux de base de dictionnaire, pendant que celui-ci est en cours de rédaction. Un système de programmes semblables permettra d'interroger FRANTEXT, le SIELLA (Système de Textes en Ligne en Libre Accès): il s'agit d'un système interrogatif et efficace grâce auquel il est possible de calculer rapidement, (en 20 secondes au minimum et en quelques minutes au maximum) le nombre d'ouvrages dans lesquels figure le mot cherché et, par la suite, le nombre de ses occurrences. La totalité ou une partie des concordances peuvent également être appelées et apparaître sur l'écran. Il est également possible de rechercher les cooccurrences de plusieurs éléments, ou les occurrences d'un même mot dans certains ouvrages déterminés (par exemple: dans l'oeuvre de Zola, ou dans les ouvrages parus en 1870, etc.) En dehors des informations concernant les mots, les programmes peuvent fournir des renseignements extrêmement "pointus" du type: "Qui donne la première réplique de la première scène de l'Acte II de Tartuffe? "Qui est le premier à prononcer le mot "aimer"? etc.

Cependant, les rédacteurs utilisent rarement ce programme puissant et préfèrent travailler sur des listes de concordances imprimées sur papier.

Pour la rédaction de leurs articles, les collaborateurs du *Trésor* peuvent utiliser trois sortes de données: celles provenant de dictionnaires déjà existants, celles contenues dans des études consacrées aux divers mots et enfin, mais non en dernier lieu, celle fournies par les listes de concordances tirées du corpus informatisé. Chaque article est confié à deux lexicographes dont l'un est chargé de rédiger les vérifications et commentaires du mot et de choisir les citations illustrant son emploi et l'autre d'indiquer les renseignements historiques et étymologiques le concernant. Chaque rédacteur reçoit un dossier dont la première page est constituée par un tableau établi par l'ordinateur sur les données statistiques les plus importantes du mot: le nombre de ces occurrences dans le corpus, et plus spécialement, dans les ouvrages littéraires et dans les ouvrages scientifiques de ce corpus, dans quels dictionnaires généraux ou techniques on le trouve; se fondant sur le nombre d'occurrences et sur la distribution du mot, l'ordinateur propose le nombre de citations à retenir (à titre indicatif, le nombre de citations retenues est en général inférieur à cette proposition). Les pages suivantes contiennent la bibliographie complète du mot et des néologismes auxquels il a donné éventuellement lieu. On y trouve ensuite les photocopies des articles que les divers dictionnaires (généraux ou techniques) consacrent au mot. Après lecture attentive de ces documents, le lexicographe utilise surtout les entrées des dictionnaires généraux, parcourt celles des dictionnaires techniques et se met ensuite à étudier les concordances. Il lit toutes les citations et relève celles qui lui semblent particulièrement illustratives, chacune des acceptions d'un mot devant être illustrée par des citations extraites d'ouvrages parus dans différents siècles. La liste des concordances ne contenant que des contextes d'une ligne, elle ne permet pas de se prononcer sur le caractère approprié de la citation. Celles qui paraissent utilisables seront portées avec un large contexte (8 lignes avant et 8 lignes après le mot) sur des fiches qui permettront d'établir le texte définitif de l'article. Si les listes de concordance ne lui fournissent pas tous les renseignements souhaités, le lexicographe consulte sa collection de fiches et peut, en dernier recours, chercher lui-même les citations nécessaires (en lisant la littérature). Les lexicographes chargés de la partie historique s'appuient avant tout sur les dictionnaires anciens et historiques et sur les recherches étymologiques concernant tel ou tel mot. Bien entendu, les listes de concordance peuvent leur être utiles, mais, chose curieuse, ils reçoivent les mêmes que leurs collègues synchronistes; or, ces listes n'indiquent aucune datation et sont ordonnées suivant la catégorie grammaticale du contexte immédiat. Les articles déjà rédigés du *Trésor* sont clairement articulés. Alors que dans Oxford English Dictionary, la présentation des différents sens du mot est quelquefois trop compliquée, le *Trésor* sépare nettement sens apparentés et sens lointains. Pour la première fois dans un dictionnaire historique, l'entrée comporte également la fréquence du mot grâce au dictionnaire de fréquence, un sous-produit du travail. Les définitions sont brèves et ne contiennent que les renseignements indispensables: ceux sur les citations qui permettent de cerner avec précision le sens du mot. Ces dernières, destinées à illustrer les différentes nuances du sens, sont chronologiquement ordonnées, suivant l'ancienneté de la

source à laquelle elles sont empruntées. L'article indique également la valeur stylistique et la sphère d'emploi du mot. Dans ce dictionnaire normatif, l'histoire du mot est traitée séparément de son étymologie proprement dite. Chaque article comporte une bibliographie.

Pour ce qui est du grand dictionnaire de la langue hongroise, la phase de la rédaction sera abordée lorsque 80% au moins du corpus sera accessible sur ordinateur et que les programmes d'interrogation seront élaborés. Il semble d'ores et déjà certain que les rédacteurs préféreront l'ordinateur au papier et au stylo. Les articles seront rédigés directement sur ordinateur et un système de menus permettra d'appeler simultanément les concordances. Les articles déjà rédigés, stockés dans un autre fichier, seront également consultables. L'ensemble du dictionnaire étant directement enregistré sur disquettes, les travaux d'édition seront accélérés et les corrections pourront être effectuées plus facilement. La base de données réunies pour le dictionnaire pourra être utilisée pour la réalisation d'autres projets.

On voit que les travaux de préparation des deux dictionnaires présentent de nombreuses analogies, ce qui n'est pas un hasard: le dictionnaire hongrois se modèle à de nombreux égards sur le *Trésor* et les entretiens que nous avons eus avec ses réalisateurs nous ont été d'un grand secours. Par ailleurs, nos conceptions divergent quelquefois en ce qui concerne l'étendue du corpus, la sélection des sources et l'utilisation de l'informatique. Nous comptons utiliser l'ordinateur non seulement pour faciliter la constitution des fichiers, mais aussi pour exploiter au maximum ses capacités et cela dans toutes les phases des travaux. En effet, pendant les 24 années qui se sont écoulées depuis le démarrage des travaux du *Trésor*, l'informatique a réalisé des progrès énormes, ce qui nous permet de procéder d'une façon différente pour la solution de nos problèmes. De plus, nous avons eu, dès la phase initiale préparatoire, la possibilité de profiter des expériences recueillies au cours de l'élaboration d'autres dictionnaires réalisés à l'aide d'ordinateurs, avant tout du New Oxford English Dictionary et du Dictionary of Old English.

Bibliographie

AITKEN, A.J. (1971), Historical dictionaries and the computer. In: *The computer in Literary and Linguistic Research.* R.A. WISBEY (ed.), Cambridge. UP.

AMOS, A. (1984), *Computers and Lexicography: The Dictionary of Old English.* Status Report on the Dictionary of Old English Project. Manuscrit

BERG, D.L. és GONNET G.H. és TOMPA, T.Wh (1988), *The New Oxford English Dictionary Project at the University of Waterloo* OED-88-01

DENDIEN, J. és GORCY, G. és MARTIN, E. (1986), *Le Trésor général des langues et parlers français de l'Institut national de la langue française.* Manuscrit. INALF

GORCY, G. (1983), Les dérivés d'esprit en français moderne: Méthodologie et esquisses d'articles à paraître dans le Trésor de la langue française. *Lessico Intelletuale Europeo IV Colloquio Internazionale.* Roma 7-9 gennaio 1983.

GORCY, G. (1984), L'ordinateur au service de la lexicographie: une expérience et ses problèmes. (HAMESSE 1985).

JOHANSSON, S. (1988), The New Oxford English Dictionary project: A presentation. *ICAME Journal* No. 12. Norwegian Computing Centre for the Humanitites.

LANDAU, S. I. (1984), Computer Use and the Future of Dictionary Making. *Dictionaries. The Art and Craft of Lexicography.* Charles Scribner's sons. Nex York. pp. 272-294.

MARTIN, E. (1984), Une banque de données sur la langue française. *BRISES. Bulletin de recherches sur l'information en sciences économiques, humaines et sociales.* La Linguistique dans les systèmes documentaires. Avril 1984. No.4.

MERKIN, R. (1983), The historical/ academic dictionary. In: R.R.K. HARTMANN (ed): *Lexicography: Principles and Practice.* Academic Press.London - New-York - Paris - stb. pp. 123-133.

PAJZS, J. (1987), *Javaslat a Nagyszótár számítógépes megvalósítására. A gyűjtés és feldolgozás rendszerterve.* Manuscrit

PAJZS, J. (1988a), Félmillió szó számítógépen. *Computerworld - Számítástechnika* II. évf. 5. p.24-25.

PAJZS, J. (1988b), Számítógépes szótárak. *NyK.* (A paraître).

PAJZS, J. (1988c), Dictionary digitalisiert: Oxford English per Knopfdruck. *Computerwelt Österreich* No.9.

PAPP, F. és HEXENDORF, E. (1985), Magyar szókincs a könyvnyomtatástól napjainkig - számítógépre tervezve. *Magyar Tudomány,* 1985/1.