

Le *Corpus Augustinianum Gissense* ou saint Augustin assisté par ordinateur

Présentation générale

Le *Corpus Augustinianum Gissense*¹ (*CAG*) est l'édition électronique sur CD-ROM, préparée par Cornelius Mayer, des œuvres de saint Augustin, y compris les fragments, transmises par la tradition directe ou indirecte², soit un peu plus de cinq millions de mots dans la version actuelle du disque. Celui-ci renferme aussi la liste des œuvres telle que l'a établie Karl Heinz Chelius dans l'*Augustinus-Lexikon*³. Pour chacune d'elles, on trouve le nom abrégé (utilisé dans les champs d'interrogation du logiciel de recherche), le nom complet et le renvoi à la meilleure édition critique disponible. Cette liste peut s'afficher dans l'ordre alphabétique ou chronologique. Le CD-ROM comporte enfin un peu plus de 20 000 références bibliographiques (articles, mélanges, etc.) interrogeables indépendamment du corpus littéraire.

Le CD-ROM⁴ est livré avec deux disquettes d'installation, l'une pour MS-DOS (versions 3.3 et suivantes), l'autre pour Windows (3.1 ou 95). L'installation, qui s'effectue automatiquement, recopie sur le disque dur un dossier, intitulé *CAG*, qui contient deux fichiers : celui du logiciel d'interrogation, intitulé *Cag*, et celui du fichier d'aide, *Cag-Hilfe*. On travaille avec le CD-ROM constamment présent dans le lecteur. Nous avons testé le disque sur un

1. La mise sous forme électronique du corpus a été entreprise à l'université de Giessen. Elle a été poursuivie à Würzburg.

2. Le CD-ROM inclut la plupart des textes retrouvés récemment, par exemple les sermons redécouverts à Mayence en 1990.

3. *Augustinus-Lexikon*, herausgegeben von Cornelius MAYER in Verbindung mit Erich FELDMANN [et al.], Redaktion Karl Heinz CHELIUS, Basel, Schwabe, 1986.

4. Cet article traite de la version 1.0 du *CAG*, testée en février 1998. Pour tout renseignement complémentaire, s'adresser à SCHWABE & CO A.G., Farnburgerstrasse 8, 4132 Muttenz (Suisse). Tél. 061/461.27.61. Fax 061/461.27.51. Prix (en 1996) : 2 950 FF. Interface du logiciel et manuel en allemand. ISBN : 3-7965-0989-4.

compatible PC équipé du système d'exploitation Windows 95 à l'Institut d'Études Augustiniennes⁵ : tout a parfaitement fonctionné. Nous avons également fait tourner le CD-ROM sur Macintosh grâce à l'émulateur⁶ *SoftWindows 95* sans aucun problème.

Avantages de la lemmatisation

Le CAG est fourni avec un logiciel d'interrogation⁷ qui, à l'instar de ses homologues dédiés à l'interrogation de corpus littéraires en texte intégral (*full text*), permet d'effectuer des recherches en proposant des séquences de caractères correspondant parfois à des mots entiers, et le plus souvent à des mots tronqués. Dans ce dernier cas, l'utilisateur dispose de trois métacaractères qu'il place librement en début, en milieu ou en fin de séquence. Le métacaractère <étoile> remplace zéro caractère ou un nombre quelconque de caractères ; le métacaractère <point d'interrogation> remplace un et un seul caractère ; quant au <point d'exclamation>, il se substitue à zéro ou un caractère. Par exemple, *fid** permettra de trouver les occurrences appartenant à la déclinaison de *fides*, mais aussi n'importe quel mot commençant par *fid* ; *?aulus* fera trouver *Paulus* et *Saulus* ; avec la séquence *ne!*, on obtiendra *ne* et *nec*.

L'originalité du CAG provient de ce que presque⁸ chaque mot du corpus littéraire a été lemmatisé, c'est-à-dire qu'il est accompagné d'un codage indiquant son analyse grammaticale. Il suffit que le nominatif singulier d'un substantif ou d'un adjectif soit précédé de l'opérateur *l* : pour que toutes les formes de la déclinaison correspondante soient trouvées ; pour les verbes, il convient d'indiquer la première personne du présent de l'indicatif. Grâce à cette facilité, on recherche des mots dans le CD-ROM comme s'ils étaient des entrées dans un dictionnaire traditionnel sur papier. La lemmatisation permet en outre de différencier d'emblée les formes homographes, et d'éliminer, dans l'obtention des résultats, cette part de « déchet » que la recherche habituelle par séquences de caractères est impuissante à éviter. Par exemple, si l'on a demandé *l: lex* et qu'on obtient une occurrence comportant *legis*, on a l'assurance qu'il

5. Nous tenons à remercier Madame Simone Deléani, directrice de la bibliothèque de l'Institut d'Études Augustiniennes, ainsi que Monsieur Stéphane Toulouse, chargé de recherches au C.N.R.S., qui nous ont très cordialement accueilli dans leurs locaux, et sans l'aide desquels ce banc d'essai n'aurait pu être réalisé.

6. *SoftWindows 95* est un émulateur du système d'exploitation des compatibles PC permettant de faire fonctionner sur Macintosh des logiciels et des CD-ROM prévus seulement pour les PC. Il réclame un minimum de 24 Mo de mémoire vive (RAM) et 200 Mo sur le disque dur. Son éditeur est Insignia et son prix d'environ 1 500 FF (TTC). Un autre émulateur est disponible depuis peu : *Virtual PC* (de la firme Connectix) aux environs de 1 500 FF (TTC) également, mais nous n'avons pas eu l'occasion de le tester.

7. Le logiciel d'interrogation est l'œuvre de Makrolog GmbH.

8. 73 formes n'ont pas été lemmatisées parce qu'elles apparaissent trop fréquemment dans le texte (*ad, autem, cuius, cum, etc.*).

s'agit du génitif singulier du substantif et non pas d'une forme conjuguée du verbe *lego*.

Inversement, l'opérateur d'appartenance <^> permet, lorsqu'on cherche une forme déclinée ou conjuguée précise, d'éliminer les formes homographes indésirables. Ainsi, l'interrogation *malo^l:malum*, signifie « trouver les occurrences de *malo* qui appartiennent au lemme *malum*, i, n. », en évitant d'obtenir des occurrences de la forme conjuguée. L'interrogation par lemme présente enfin l'avantage de protéger le chercheur contre les risques induits par les variantes orthographiques.

Deux autres opérateurs permettent d'affiner encore davantage la distinction entre les formes : les noms communs se filtrent à l'aide de *k*: et les noms propres, avec *g*:. Ainsi, *k:maria* permet de ne pas confondre « les mers » avec *g:María*, « la Vierge Marie ». De plus, comme la lemmatisation est indépendante de la mise en forme de la page, les occurrences du nom commun seront trouvées même si, dans le texte, il porte une majuscule (par exemple en début de phrase).

Les mots grecs translittérés sont, eux aussi, marqués d'un code spécial. Quand on affiche l'index alphabétique des formes employées par saint Augustin et qu'on tape *h*:, on obtient la liste des formes grecques. Si l'on tape *h:a* comme séquence à rechercher, on trouve les occurrences de la lettre alpha.

Recherche des citations

Les phrases constituant des citations sont indexées de manière à en identifier la provenance, soit dans la Bible, soit dans un auteur profane. Aussi peut-on interroger le corpus à l'aide d'opérateurs spécifiques pour rechercher les citations selon deux angles d'approche différents.

Si l'on emploie l'opérateur *q*: (« Quelle⁹ »), on veut obtenir les occurrences d'une source précise, qui peut être un auteur ou une œuvre. Ainsi, *q:Rom* détectera les citations faites à partir de l'*Épître aux Romains*, *q:Cic* celles de Cicéron (quelle que soit l'œuvre) et *q:doctr. chr.* celles où saint Augustin renvoie à son propre ouvrage.

Si maintenant on utilise l'opérateur *z*: (« Zeugnis »), on peut chercher des citations dont on ignore la source exacte, soit dans la Bible (en étant dispensé de spécifier de quel livre il s'agit), soit dans la littérature profane (sans avoir à fournir un nom d'auteur ou d'œuvre). Très organisé, le logiciel prévoit neuf catégories de citations, parmi lesquelles les citations bibliques, celles d'un autre auteur, celles où saint Augustin se cite lui-même, ou encore les citations d'un autre auteur citant la Bible : il existe, à cet effet, neuf opérateurs intitulés *z1*:, *z2*:, ... *z9*:. Cependant, si l'on souhaite éviter de restreindre abusivement son champ d'exploration, on a le droit d'employer l'opérateur de troncature et d'écrire *z**:

9. Les métacaractères ont été choisis pour leur valeur mnémotechnique. On les manipulera plus aisément si l'on se souvient du mot allemand auquel ils correspondent.

Il importe enfin de se souvenir que les opérateurs relevant de la syntaxe de recherche des mots et ceux de la recherche des citations sont combinables. Par exemple, la séquence *l:deus^q:mt* fera trouver les occurrences du lemme *deus* dans les citations tirées de l'Évangile selon saint Matthieu.

Recherche de mots multiples

Le logiciel d'interrogation donne accès aux opérateurs booléens devenus traditionnels : la syntaxe *séquence1 oder séquence2* détecte l'occurrence quand le contexte comporte soit *séquence1*, soit *séquence2*, soit les deux séquences dans n'importe quel ordre et séparées par un nombre quelconque de mots à l'intérieur de ce contexte ; *séquence1 und séquence2* retient l'occurrence si et seulement si les deux séquences se trouvent dans le même contexte, peu importe dans quel ordre et à quelle distance l'une de l'autre ; enfin, *séquence1 nicht séquence2* examine si *séquence1* est présente et *séquence2* absente.

D'autres opérateurs permettent de spécifier avec précision la proximité et l'ordre de succession obligatoire entre deux séquences. La syntaxe */wx* (où *x* représente un nombre entier) indique dans quel intervalle, en nombre de mots (« Wort »), doivent se succéder les séquences, tandis que */sx* (« Satz ») effectue la même opération, mais en nombre de phrases. Ainsi, l'instruction *l:Stephanus /w5 l:Paulus* fera trouver les formes déclinées des lemmes *Stephanus* et *Paulus*, à condition qu'elles se trouvent éloignées l'une de l'autre de cinq mots au maximum, et dans n'importe quel ordre.

Si l'ordre de succession est une condition essentielle de l'interrogation, on emploie la syntaxe *séquence1 /+wx séquence2* pour indiquer que *séquence1* doit obligatoirement précéder *séquence2* du nombre *x* de mots, et *séquence1 /+sx séquence2* quand l'intervalle est exprimé en nombre de phrases. Le cas inverse est également prévu, et la syntaxe *séquence1 /-wx séquence2* signifie que *séquence1* doit suivre *séquence2* à une distance maximale de *x* mots, alors que *séquence1 /-sx séquence2* indique la succession à une distance de *x* phrases.

À l'intention de ceux qui éprouvent quelque réticence à rédiger eux-mêmes la formule d'interrogation avec ses subtilités, le logiciel propose un dialogue d'aide dans lequel il suffit de taper les deux séquences, puis de préciser, en cochant des cases, quels opérateurs on souhaite utiliser, puis, le cas échéant, de taper la valeur numérique de l'intervalle entre les mots ou les phrases : après quoi, la formule d'interrogation, mise en forme par le logiciel et syntaxiquement correcte, apparaît, prête à être exécutée.

Déroulement d'une recherche de formes

Avant d'effectuer sa recherche, il ne faut pas oublier de définir la manière dont s'afficheront les résultats. Quand on déroule le menu *Anzeige*, on choisit entre plusieurs options : en sélectionnant l'article *Kurztitel*, on obtiendra la visualisation des résultats selon la syntaxe Titre abrégé, Livre, Paragraphe (ex : *Acad. 2, 12*) ; sinon, les occurrences trouvées seront affichées à l'intérieur d'un

contexte réglable de une à quatre lignes (articles *1-Zeilen KWIC* à *4-Zeilen KWIC*).

Par défaut, la recherche porte sur l'ensemble du corpus. Si l'on souhaite la restreindre à une seule œuvre, on clique dans le champ *Werkauswahl* et l'on spécifie l'œuvre à examiner, soit en tapant directement le titre abrégé, soit en cliquant sur le bouton *Index* et en sélectionnant le nom requis dans une liste déroulante.

Ces précautions préalables prises, on clique dans le champ *CAG-Suche* et l'on tape la formule d'interrogation conformément aux règles syntaxiques exposées dans les paragraphes précédents. Si l'on appuie sur la touche *Entrée*, une fenêtre s'ouvre, listant toutes les formes complètes, présentes dans la base de textes, et correspondant à la séquence de caractères demandée. Toutes les formes sont cochées par défaut. Si l'on considère certaines d'entre elles comme indésirables, on les exclut de la recherche en enlevant la coche qui les précède.

Lorsqu'on appuie sur la touche *Retour*, un message signale que la recherche s'effectue : celle-ci peut être interrompue à tout moment en cliquant sur le bouton *Abbrechen*. Une fois la recherche terminée, le nombre d'occurrences s'affiche dans la fenêtre de recherche, tandis que s'ouvre une seconde fenêtre qui visualise la liste des résultats conformément à l'option d'affichage définie précédemment.

On examine une occurrence dans son contexte en cliquant sur la ligne voulue de la liste. Une nouvelle fenêtre s'ouvre, dans laquelle l'occurrence apparaît en inversion vidéo. Un ascenseur permet de dérouler toute la page de l'édition critique qui contient l'occurrence. Grâce au bouton *Blättern*, on peut, si nécessaire, continuer l'examen en remontant jusqu'à la première page de l'œuvre, ou au contraire, en descendant jusqu'à la dernière. L'identification du passage en cours d'examen est faite simultanément suivant la syntaxe du *Kurztitel* et, dans un cartouche au bas de l'écran, par un renvoi aux numéros de volume, page et ligne d'une édition critique (appartenant, par exemple, au *Corpus Christianorum Latinorum*). On examine une occurrence après l'autre, soit en remontant dans la liste par le bouton *Flèche gauche*, soit en descendant par *Flèche droite*.

Les résultats peuvent être envoyés à l'imprimante ou exportés dans un fichier ASCII (ou Texte seul) qui est lisible sur n'importe quel micro-ordinateur PC ou Macintosh et dans n'importe quel traitement de texte. Dans les deux cas, les options sont identiques : on récupère soit la seule occurrence actuellement visible à l'écran, soit la totalité des occurrences, soit une sélection d'occurrences définie en laissant ou en ôtant des coches dans une liste, soit la seule liste des œuvres.

En cliquant sur le bouton *Neue Suche*, on efface d'un coup le contenu de tous les champs et l'on est prêt à effectuer une nouvelle recherche.

Recherche bibliographique

Par défaut, c'est la fenêtre de recherche des formes qui se présente (*CAG-Suche*). Pour effectuer une recherche bibliographique, il faut dérouler le menu *Suchen* et sélectionner l'article *Literatur*.

Le formulaire d'interrogation se compose de six champs : AUTor, TITel, BIBliographie, QUElle, GATtung et REFerenz. On peut taper librement dans chaque champ, mais il est préférable, pour s'assurer que les informations entrées sont pertinentes, de cliquer dans le champ voulu, puis de cliquer sur le bouton *Index* afin de dérouler une liste dans laquelle on sélectionnera, suivant le cas, un nom, un titre, un mot-clé, etc., qui viendront se coller automatiquement et sans risque d'erreur.

Le champ AUTor reçoit le nom de l'auteur d'un article ou d'un ouvrage, et TITel, son titre. Dans le champ BIBliographie, on saisit soit le numéro d'une année pour obtenir toute la littérature parue cette année-là, soit le nom d'une revue, éventuellement accompagné de la tomaison. Le champ GATtung offre de restreindre la recherche à l'un des cinq supports suivants : *Festschrift*, *Lexikon*, *Monographie*, *Sammelband* et *Zeitschrift*. Quant au champ REFerenz, on l'alimente avec un ou plusieurs mots-clés répertoriés dans l'*Augustinus-Lexikon* (ex. : *deus*, *historia*, *Plato*, etc.). Si l'on renseigne un seul champ, on risque de définir une recherche trop générale et donc d'obtenir des résultats trop nombreux et insuffisamment précis ; aussi a-t-on intérêt à mieux cibler sa recherche en remplissant plusieurs champs, ce qui équivaut à lier le contenu de chacun d'eux par un opérateur booléen ET. Par exemple, en utilisant simultanément les champs AUTor, GATtung et REFerenz, on peut obtenir le relevé de toutes les monographies dans lesquelles tel auteur a examiné la notion de *deus* chez saint Augustin.

Statistiques

Le menu *Suchen* possède un troisième article, *Frequenzanzeige*, dont la sélection provoque l'affichage d'une fenêtre pourvue du champ *Begriff*, destiné à être rempli soit avec un lemme, soit avec une forme entière ou tronquée, exactement comme on le fait pour une recherche d'occurrences, soit avec une forme sélectionnée à partir de l'index.

Dès qu'on a cliqué sur le bouton *Frequenz anzeigen*, une liste affiche la fréquence d'emploi des formes demandées, avec les écarts types, et en présentant le décompte œuvre par œuvre. Cette liste ne peut être ni imprimée ni exportée dans un fichier.

Conclusion

Cette version 1.0, nous a-t-il semblé, ne comporte pas la totalité des textes augustiniens récemment redécouverts, et le nombre des références biblio-

graphiques, bien que considérable, reste encore inférieur à celui qui est disponible dans les fichiers manuels de centres spécialisés comme l'Institut d'Études Augustiniennes.

Ces quelques réserves faites, qui ne concernent que l'aspect quantitatif des données introduites dans le CD-ROM, il convient de retenir que le CAG représente, pour le chercheur, un outil d'excellente qualité, agréable à manipuler et d'une grande sûreté de fonctionnement.

Daniel BÉGUIN
École normale supérieure

RÉSUMÉ : Le CAG est un CD-ROM qui donne accès à l'ensemble presque complet des œuvres de saint Augustin et à une base de données bibliographiques d'environ 20 000 références. Le logiciel d'interrogation permet de retrouver les mots en spécifiant soit des séquences de caractères, soit des lemmes. Les recherches complexes s'effectuent grâce aux opérateurs booléens et à d'autres qui définissent l'ordre de succession et l'intervalle maximal entre les mots. Cet outil bien conçu semble capable de rendre de grands services aux chercheurs.

ABSTRACT : The CAG CD-ROM is constituted of an almost complete database of the Augustinian works and a bibliographical database including about 20,000 items. The retrieval software lets the scholar look up words when typing either search patterns or lemmas. Complex search formulations use the Boolean operators and ones more sophisticated that specify a fixed order or a proximity of words in a sentence. This well-designed tool will probably be of great use to the researchers.