LES CONCORDANCES ET LISTES DE FREQUENCE DES TEXTES NARRATIFS LATINS COMPOSES EN BELGIQUE AVANT 1200

L. GENICOT, Professeur à l'U.C.L.

Le hasard et la nécessité gouvernent-ils l'univers, comme le veut un livre célèbre dès sa sortie de presse ? Ils ont, en tout cas, présidé à la naissance de l'entreprise dont je voudrais brièvement exposer les origines, l'objet et les limites, les modalités d'exécution, l'état actuel.

Le hasard a joué. La résidence loin de la Belgique d'un collègue bien plus qualifié a amené, en 1961, le "Comité national belge du dictionnaire du latin médiéval", dépendant de l'Académie royale de Belgique et travaillant sous les auspices de l'Union académique internationale, et le "Centre belge d'études de la latinité médiévale" à m'introduire dans leur sein. La maladie d'un autre a déterminé, en 1967, le second de ces organismes à me porter à sa présidence. J'ai pu ainsi, dès la première date, attirer l'attention des deux institutions sur l'intérêt des ordinateurs puis, en 1968, soumettre, au nom du Centre, au Fonds de la Recherche Fondamentale Collective et faire subsidier un projet, esquissé en 1963, de mise sur bandes magnétiques des textes écrits avant 1200 dans les bornes de la Belgique actuelle (1).

Une triple nécessité imposait, à mon sens, le recours aux machines électroniques. Celle d'accélérer le rythme et d'améliorer les méthodes d'élaboration du Dictionnaire du latin médiéval belge mis en chantier il y avait près d'un demi siècle par le Comité mentionné plus haut. Celle de pousser les études de sémantique, fondamentales dans tous les secteurs de l'histoire médiévale, sociale ou politique ou religieuse ou culturelle et, à cet effet, de repérer tous les emplois des termes techniques comme *nobilis, auctoritas* ou *monachus*. Celle de perfectionner nos instruments de travail et notamment nos éditions de sources écrites, qu'il s'agisse de découvrir ou vérifier l'identité ou d'établir le texte (2).

Ce qui précède a déjà annoncé en gros l'objet de l'entreprise. De façon précise, elle entend reporter sur bandes magnétiques, pour en dresser des concordances et des listes de fréquence, les oeuvres narratives composées sur le territoire de la Belgique actuelle avant 1200 et éditées. Elle s'assigne donc des limites, qu'il importe de justifier.

Elle exclut les inscriptions et, masse imposante, les textes juridiques, tant normatifs que de la pratique.

Le traitement de ceux-ci par ordinateurs pose des problèmes spécifiques, dont le Centre de recherches et d'application linguistiques de Nancy et l'Institut de recherche et d'histoire des textes de Paris poursuivent la solution (3). Si, dans trois ans, notre Centre dispose de nouveaux crédits, il pourra, en un second stade, procéder à l'enregistrement et à l'analyse de ces documents en s'inspirant des résultats des expériences menées par nos collègues français.

Le même souci de coordination sur le plan international a dicté le choix du cadre géographique. L'espoir que les pays voisins entament sans tarder un effort analogue au nôtre a fait opter pour les frontières actuelles. La décision nous a coûté : elle obligeait à des sacrifices dont tout médiéviste mesurera le prix, spécialement l'abandon de Saint-Amand. Telle quelle, elle ne suffisait d'ailleurs pas à trancher un bon nombre de cas, comme celui d'auteurs qui ont travaillé en divers endroits, de Sigebert de Gembloux, par exemple, qui a composé plusieurs *Vitae* à Metz. Elle devait être explicitée, en trois points surtout, et elle l'a été de la manière suivante : sont retenues toutes les productions d'un écrivain qui est né en Belgique ou y a passé la plus grande partie de sa vie, toutes les pièces anonymes dont le sujet autorise à croire qu'elles ont été rédigées en Belgique, toutes les refontes ou continuations d'une oeuvre originaire de Belgique, même si elles ont été réalisées par un auteur ou dans un lieu étranger.

Le *terminus ad quem* ne pouvait être que 1200. Le Nouveau Ducange s'arrête à cette date et, après elle, les documents se font trop nombreux pour être traités systématiquement, en tout cas dans une première étape.

Quant à se borner aux textes édités, cela ne porte apparemment guère à conséquence : les sources narratives écrites avant le XIIIe siècle dans ce qui est aujourd'hui la Belgique doivent avoir été presque toutes publiées.

Les limites tracées, le Centre belge a arrêté les directives d'exécution. Il a eu la bonne fortune de disposer, pour leur rédaction comme ultérieurement pour leur mise en oeuvre, de la compétence et du dévouement de son secrétaire, M. Paul Tombeur, chargé de cours associé à l'Université catholique de Louvain, parfaitement secondé par le P. André Stainier, du Saulchoir.
Les premières de ces règles concernent les opérations préparatoires à l'enregistrement, c'est-à-dire le repérage, puis le classement et le codage des oeuvres.

Un *Index scriptorum operumque latino-belgicorum medii aevi* avait été dressé en 1934 et deux fois complété par le médiolatiniste de l'Université de Liège, M. Maurice Hélin (4). Déjà il avait été remis sur le métier par des chercheurs du Centre belge. Il a été entièrement révisé afin de ne pas souffrir de lacunes et de renseigner, pour chaque oeuvre, les éditions et les études de critique externe. Une fois au point, il paraîtra sous les auspices de l'Académie royale de Belgique et guidera dans l'utilisation de nos concordances.

Il donne, en effet, le nom de code des oeuvres. Pour composer ce nom, on a classé celles-ci par siècle, puis par datation certaine ou douteuse, enfin par genre, hagiographique ou non; quand on hésitait entre deux siècles, on a opté pour le plus récent. On a ainsi dressé la grille que voici :

Siècle	EPOQUE CERTAINE		EPOQUE INCERTAINE	
	Oeuvres non hagiogr.	Oeuvres hagiogr.	Oeuvres non hagiogr.	Oeuvres hagiogr.
6e - 7e	A	B	C	D
8e	E	F	G	H
9e	I	J	K	L
10e	M	N	O	P
11e	Q	R	S	T
12e	U	V	W	X
13e	Y	Z	0	1
14e	2	3	4	5
15e	6	7	8	9

Le premier élément du nom de code, lettre ou chiffre, indique, de la sorte, l'époque de rédaction, sûre ou hypothétique, et l'appartenance ou non à un des genres hagiographiques. Deux lettres forment ensuite un sigle qui suggère l'auteur ou le titre : BA signifie, par exemple, un texte intéressant saint Bavon. Une dernière distingue les oeuvres d'un même écrivain ou d'un même groupe, par exemple, *Vita, Miracula, Translatio, Triumphus* d'un saint ou versions successives d'une *Vita,* pour autant évidemment que ces oeuvres et versions appartiennent à un même siècle.

Quatre signes, - quatre colonnes seulement d'une fiche perforée - identifient donc tout écrit et suffisent à en donner les traits essentiels. JHUB, par exemple, désigne une oeuvre hagiographique, certainement composée au IXe siècle,traitant d'un personnage dont le prénom commence par Hu, - on devine qu'il s'agit de saint Hubert - et constituant la seconde d'un cycle, - en l'occurrence une *Translatio* qu'une *Vita* avait précédée.

Dans le classement et donc dans l'*Index scriptorum,* les remaniements figurent en deux endroits : en appendice à l'article consacré à l'oeuvre initiale et à leur place chronologique.

Repérage et codage achevés, l'enregistrement est possible. Il compte une série d'opérations : ce que nous appelons la pré-édition, puis la perforation et son contrôle, le report sur bande avec la référenciation automatique, l'analyse du texte, l'insertion de cette analyse dans la bande initiale.

Travail décisif, la pré-édition soulève des problèmes nombreux et délicats. Le choix du texte d'abord. Il est dicté par les données de l'*Index scriptorum* et s'arrête donc toujours sur une des éditions existantes. Celles-ci sont rarement exemplaires; elles sont même trop souvent médiocres; quiconque a, -pour prendre un des cas les plus favorables - examiné celles des *Scriptores* en connaît les faiblesses. Mais en établir de meilleures exigerait un temps et des moyens énormes et enfermerait d'ailleurs dans un cercle vicieux : on ne conçoit plus aujourd'hui la critique de restitution, *recensio* ou *emendatio,* sans recours aux machines électroniques. La seule voie à suivre consiste par conséquent à enregistrer le meilleur texte disponible, à le faire analyser par l'ordinateur et à le corriger selon les indications de cette analyse. En principe, on retient la dernière édition complète. On en reprend éventuellement une partielle, plus récente et plus sûre, si elle offre une certaine longueur (5); à charge, il va sans dire, de signaler clairement le caractère "composite" de tels enregistrements.

Les attachés du Centre belge photocopient alors en plusieurs exemplaires le texte sélectionné en même temps que l'apparat critique, les additions et corrections, les données de la ou des préfaces sur la tradition manuscrite. Puis ils procèdent sur un des exemplaires aux aménagements préparatoires à la perforation. Ils trient les variantes, pour ne retenir que celles qui offrent un intérêt lexicographique ou historique. Ils résolvent les abréviations et affectent les mots complétés d'un signe de code qui attirera l'attention de l'utilisateur des concordances : une abréviation peut avoir été mal interprétée. Ils classent par grandes catégories (Bible, Pères, liturgie, sources diplomatiques, etc.) les citations explicites ou repérées par le ou les éditeurs. Ils révisent la ponctuation, pour n'en retenir que les éléments "forts", points, points d'interrogation, points d'exclamation, deux points, point-et-virgule; c'est là une liste *a priori* dont ils peuvent s'écarter, à condition de le signaler dans la notice de pré-édition; de toute façon, ils suppriment les point-et-virgule et les deux points qui ne font que séparer une subordonnée de sa principale et ils ramènent les éléments "forts" à une seule forme, celle du point. Ils transposent dans l'alphabet latin les mots tracés en caractères étrangers, séparent les enclitiques (-que, -ve, -ne, -cum) des mots auxquels ils sont liés, changent l'e "caudé" en ae ou pa. fois en oe, placent le signe N devant les chiffres, unifient les mots composés, corrigent les erreurs typographiques. Ils décident enfin, sur la foi des études existantes, si titre, *incipit, explicit* sont originaux ou anciens et méritent ainsi d'être enregistrés.

Les principales conclusions de toutes ces recherches et options sont consignées dans une "notice de pré-édition". Cette carte d'identité, si l'on veut, fait connaître le titre de l'oeuvre, l'auteur, l'édition (ou les éditions) retenue avec la justification du choix par renvoi aux travaux critiques, le nom de code, les adaptations apportées à cette édition en dérogation aux règles générales qu'on vient de définir.

L'enregistrement sur cartes perforées suit la pré-édition. Il commence par la perforation d'une "carte initiale", qui indique les références du premier mot du texte et le nom de code de l'oeuvre. Il se poursuit par la perforation du texte en continu, à raison d'une ligne de l'édition par carte; un signe codé achève celle-ci et indique que la césure correspond à une fin de ligne ou une fin de page ou une fin de livre ou de chapitre. On retient en effet pour la référenciation les divisions en livres et chapitres si elles sont originales ou anciennes ou aujourd'hui courantes.

Le contrôle de la perforation s'opère selon les normes que chacun sait : un *listing* imprimé par une tabulatrice est collationné avec l'édition et les fautes repérées sont corrigées sur les cartes.

Le report sur bandes magnétiques, destiné à accélérer les travaux d'analyse et d'exploitation, se réalise également de la manière habituelle et n'appelle pas de commentaires. On notera seulement qu'il s'accompagne de la référenciation intégrale de chaque mot. Des *listings* peuvent alors être établis, où les mots sont classés soit dans l'ordre du texte, soit alphabétiquement et flanqués de la référence à l'oeuvre, à la page, à la ligne, au rang dans le texte.

Le stade suivant, l'analyse du texte a, par contre, posé un problème qui a été longuement discuté au sein du Centre belge... et ailleurs : celui de la lemmatisation. Confier celle-ci à des personnes ? Travail fastidieux, lent, coûteux ! Aux machines ? Danger des homographes ! Finalement on a opté pour une semi-automatisation. L'ordinateur lemmatise tout ce qu'il sait et il accompagne les mots ambigus d'un signe de code qui avertira le chercheur. Celui-ci vérifie le travail, le complète et le corrige.

Point particulier : les noms propres. Les toponymes qui ne figurent pas dans l'*Onomasticon* de Forcellini sont ramenés à leur forme actuelle dans l'idiome du pays mais on en dressera des concordances non lemmatisées. Les anthroponymes latins ou latinisés dans le texte reçoivent un lemme latin, les formes en langue vulgaire sont reprises au lemme, éventuellement avec une certaine uniformisation.

C'est au cours de cette phase également que chaque mot des citations est doté du signe de code approprié, P, par exemple, pour les passages des Pères ou D pour les emprunts à une source diplomatique. La même chose se fait pour chaque mot qui est l'objet d'une note de l'apparat critique et pour les variantes d'intérêt lexicographique ou historique fournies par cette note. Il s'agit d'attirer l'attention ici sur le caractère hypothétique, là sur l'absence d'originalité du texte et de sa langue.

Reste à inscrire les résultats de l'analyse sur la bande magnétique par forme de mise à jour (6).

Le Centre belge tiendra ses bandes et ses *listings* à la disposition de tous. Mais il ne peut en rester à ce stade qui suffirait à la rigueur aux études de sémantique historique. Il doit faciliter d'autres recherches et surtout préparer le dictionnaire dont la publication est sa mission première. Cela suppose qu'il pousse jusqu'à l'établissement d'une concordance.

Une concordance. Une seule. Générale. Le danger guette aujourd'hui l'érudition de multiplier les emplois partiels des ordinateurs, enregistrements partiels, *listings* ou concordances partiels, réalisés dans le cadre d'une étude limitée et pratiquement inutiles lorsque celle-ci est achevée. Ce que le Centre belge veut forger, c'est un instrument de travail qui aide à répondre au maximum de questions.

Mais quel type de concordance ? Après débats et expériences, on s'est décidé pour une concordance générale lemmatisée. Qui relève tous les mots de chaque oeuvre. Qui les accompagne d'un contexte de 100 signes maximum, contexte "optimalisé", c'est-à-dire appartenant tout entier à une seule phrase ou à un seul membre de phrase limité par deux points "forts" de la ponctuation. Qui les classe par lemme, pour chaque lemme, par ordre alphabétique des formes, pour chaque forme, par la place dans le texte. Qui indique enfin la fréquence. Ainsi une première ligne donne le lemme et sa fréquence; une seconde contient la forme unique ou la première forme alphabétique du mot avec référence et contexte, une troisième, éventuellement le second emploi de cette forme avec référence et contexte, etc. Ce système borne le travail à la lemmatisation et évite une longue analyse morphologique. Il a, pour les "grammairiens", l'inconvénient de classer les formes dans l'ordre alphabétique et non morphologique.

Des concordances d'autre type ne sont pas exclues. Notamment une concordance par formes, qui place celles-ci au milieu de la ligne. Elle présente le défaut de réduire souvent le contexte utilisable, puisqu'elle ne tient pas compte de la ponctuation : une partie des 100 signes disponibles appartient fréquemment à une phrase ou à un membre de phrase antérieur ou postérieur. Mais elle offre l'avantage de mettre en relief, dans sa colonne médiane que l'oeil parcourt aisément de haut en bas, des associations de mots qui peuvent être instructives. Elle révélerait, par exemple, que *vir* va presque toujours de pair avec *nobilis* et *homo* avec *liber.*

Des listes de fréquence sont également prévues, qui ramassent les indications des concordances sur ce point et donnent une vue d'ensemble du vocabulaire d'un auteur ou d'une oeuvre.
m
Les érudits et les organismes scientifiques disposeront à leur gré de ces concordances et listes de fréquence. Ils pourront aussi forger d'autres instruments de travail à partir des matériaux enregistrés par le Centre. Et encore enrichir ces matériaux par leurs recherches personnelles.

Au moment où ce rapport est rédigé, à la mi-février 1971, le travail est déjà bien avancé. L'*Index scriptorum operumque* est révisé pour les 9e et 10e siècles. La pré-édition, la perforation et son contrôle, le report sur bandes sont terminés pour le 9e siècle et pour la moitié des oeuvres du 10e. La lemmatisation est faite pour les oeuvres hagiographiques du 9e et partiellement pour Christian de Stavelot et Sedulius Scottus. La concordance de l'ensemble des oeuvres hagiographiques du 9e siècle est dressée. Les délais prévus au départ sont ainsi respectés et tout permet d'augurer que l'entreprise s'achèvera en 1973. Elle sera alors d'un appoint décisif pour l'étude de notre moyen âge dans tous les domaines.
Un schéma, sans phrases et sans alinéas, aidera sans doute à suivre la marche des opérations analysées et commentées dans ce rapport.

OPERATIONS PREALABLES A L'ENREGISTREMENT
Repérage des oeuvres et établissement d'un *Index scriptorum operumque*
Attribution à chaque oeuvre d'un nom de code

ENREGISTREMENT

Pré-édition :

- choix de l'édition
- photocopie en plusieurs exemplaires et aménagement du texte sur un de ceux-ci : variantes, abréviations, citations, ponctuation, mots étrangers, enclitiques, nombres, mots composés, erreurs typographiques, titre, incipit, explicit
- établissement de la notice de pré-édition

Enregistrement :

- perforation
- contrôle de la perforation : collationnement du *listing* imprimé par une tabulatrice avec l'édition
- report sur bandes magnétiques et référenciation automatique
- lemmatisation semi-automatique et analyse du texte
- enregistrement de l'analyse sur la bande initiale

EXPLOITATION

Etablissement d'une concordance générale lemmatisée

Etablissement éventuel d'autres concordances, notamment par formes

Etablissement de listes de fréquences.

(1) *Ordinateurs électroniques et études médiévales,* dans *Académie royale de Belgique, Bulletin de la Classe des Lettres,* 5e s., t. XLIX, 1963, p. 66-76.

(2) J'ai évoqué les principaux problèmes que soulèvent nos instruments de travail dans les premières pages de *Pour une organisation de la recherche en histoire médiévale,* dans *Francia,* t. I, 1971.

(3) L. FOSSIER et M. CREHANGE, *Un essai de traitement sur ordinateurs des documents diplomatiques du Moyen Age,* dans *Annales. Economies. Sociétés. Civilisations,* 25e année, 1970, p. 249-284.

(4) Cet *Index* a paru dans l'*Archivum latinitatis medii aevi,* t. VIII, 1934, p. 77-163 et les suppléments, dans la même revue, t. XVI, 1942, p. 65-75 et t. XVIII, 1945, p. 31-34.

(5) Pour les *Acta Sanctorum,* on s'en réfère à la 3e édition, non qu'elle l'emporte sur ses devancières, mais parce qu'elle est plus répandue et plus accessible : le traitement par ordinateur ne dispense pas de voir l'oeuvre elle-même.

(6) Sur cette opération, voir P. TOMBEUR et A. STAINIER, *Les méthodes et les travaux du Centre de traitement électronique des documents,* dans *Bulletin de philosophie médiévale,* nr. 10-12, 1968-1970, pp. 141-174.