# TRAITEMENTS AUTOMATIQUES DE TEXTES LITTERAIRES EN ANCIEN FRANCAIS. UNE EXPERIENCE: LE CHARROI DE NIMES

G. DE POERCK, Hoogleraar R.U.G.

R. ZWAENENPOEL, Assistente R.U.G.

1. Dans le texte qui a été lu par le premier soussigné à la séance du 26 février, l'accent a été mis sur les problèmes humains que pose l'irruption de la machine dans des domaines qui paraissaient jusqu'ici réservés de droit aux pures démarches de l'esprit. L'expérience acquise par mes collaborateurs gantois et moi-même concerne deux de ces domaines, le T.A. de la *chanson de geste,* et le T.A. de divers *corpus* du français d'aujourd'hui. Après un démarrage assez lent, dont il sera question plus loin, nous nous sommes trouvés soumis, il y a environ un an, à la pression amicale du représentant d'un éditeur français acquis à l'idée de publier une collection de textes en vers en ancien et en moyen français, avec accompagnement de concordances. Comme j'avais commencé depuis quelque temps déjà à établir à l'intention de mes étudiants gantois des textes stencilés reproduisant fidèlement un représentant de la vulgate de quelques chansons faisant partie du petit cycle de la *geste de Guillaume d'Orange,* en l'espèce le texte du ms. Paris B. N.fr.774, à quoi j'avais ajouté un apparat critique, et que d'autre part l'idée même de T.A. avait pris racine dans notre groupe de travail depuis une couple d'années, nous avons marqué à l'éditeur notre accord pour lancer dans les plus brefs délais une collection qui reçut le titre de *Textes et Traitement Automatique.* Le premier volume de cette collection, consacré au *Charroi de Nîmes,* 1487 vers décasyllabiques assonancés, est dans le commerce depuis la fin de juillet 1970(1). A l'heure qu'il est je n'ai connaissance d'aucun compte rendu critique (2). Mais certaines lettres reçues par l'éditeur et par moi-même donnent la température de l'accueil que risquent de réserver à notre initiative des philologues rompus aux techniques traditionnelles, mais peu informés des courants linguistiques d'aujourd'hui, peu enclins à voir dans un texte littéraire aussi un "texte de langue", comme disent les Italiens, et pour le reste peu disposés à admettre le principe même d'une intrusion de la machine dans un domaine où on s'était si bien débrouillé sans elle, et où le succès ne paraissait guère lié à de meilleures façons de traiter l'information que renferme potentiellement tout texte de langue.

Il faut mettre bien sûr au compte d'une réaction épidermique une ironisation un peu lourde sur l'emploi fait par moi de l'adjectif *artisanal,* et de l'adverbe *artisanalement.* Quiconque est tant soit peu familiarisé avec le travail-machine sait que 'traitement artisanal' est le terme dont on se sert dans le jargon du métier, sans la moindre idée péjorative, pour dési-

gner la mise sur fiches des divers genres d'information contenus dans un texte médiéval : lexique, éléments de syntaxe, morphologie, noms de personnes et noms géographiques, assonances, etc. etc. Ce qui est grave, c'est la méconnaissance d'un fait qui crève pourtant les yeux, à savoir que la machine traite toute l'information de routine, lorsque les conditions nécessaires sont réunies par l'analyste et le programmeur, selon les mêmes règles que le philologue de formation traditionnelle, mais que, les difficultés initiales surmontées, elle le fait, et elle seule peut le faire, de façon exhaustive et non-contradictoire, avec une rapidité extraordinaire, et une fidélité absolue à la leçon du texte traité. Si bien que ce qui sort de la machine, c'est-à-dire l'information traitée, est absolument conforme à ce qui y a été introduit, c'est-à-dire l'information brute. Toute information fausse présentée à l'*input* se répercute sur l'ensemble de l'information traitée, mais il en va exactement de même de toute correction apportée tardivement, en cours d'opération.

La préparation au T.A. est très longue, et hérissée de difficultés de toute espèce; les choix faits au départ, s'ils sont malheureux, sont source de retards considérables. Par ailleurs la collaboration entre philologue, ou linguiste, et analyste programmeur ne s'établit pas sans malentendus. Mais, vaille que vaille, l'entreprise avance, ses progrès représentent autant d'acquis définitifs, et après un gaspillage d'énergie et de temps, faute parfois de quelques minutes de réflexion, et d'un échange de vues au moment opportun, elle devient rentable, car les mêmes programmes de routine restent applicables à des textes nouveaux, et ils sont conçus de façon à fournir toujours une information essentielle, celle dont on ne saurait se passer dans aucun cas. En somme, après une mise en route pleine d'aléas, le T.A. se révèle payant. Sa supériorité sur le traitement artisanal, à tout le moins au niveau de l'information de routine, est dès à présent écrasante, et elle s'imposera aux plus routiniers. Les jeunes, et c'est heureux, comprennent mieux que leurs aînés que le T.A. n'est pas un traitement sui generis, qui sent parfois le soufre, mais la simple mécanisation d'opérations de routine aussi simples qu'elles sont indispensables.

La comptabilité des occurrences ne paraît pas moins suspecte à d'aucuns. Il leur reste à apprendre que le taux de fréquence d'occurrence, qu'il s'agisse d'un mot ou d'une construction, est significatif, et permet d'orienter une recherche.

Dernière querelle : faut-il mettre entre les mains de nos étudiants des textes bruts, accompagnés d'une information de routine reclassée par les méthodes du T.A., qui les obligeront à réfléchir, à se former au travail méthodique, ou faut-il leur imposer des éditions "critiques", avec lexiques "explicatifs", où le travail leur est fourni tout mâché ?

2. Dans le sommaire de notre communication, nous promettions de donner quelques renseignements sur les réalisations et les projets de notre groupe gantois en matière de T.A.

Nous les présenterons de façon succincte, dans une perspective historique.

Un des enseignements du présent colloque, qui aura été pour nombre d'entre nous une révélation, tant sur le plan des réalisations que par ses suggestions, c'est l'explosion informatique, le mot n'est pas trop fort, dans notre pays, au niveau universitaire, au bénéfice des sciences humaines. Et dans cette explosion la linguistique et la philologie ont joué à certains égards le rôle de sciences-pilote. Même si cette explosion a pu paraître anarchique à d'aucuns. Il n'est pas mauvais que chaque chercheur, chaque centre ait fait à cet égard sa maladie infantile : les confrontations prennent de ce fait plus de signification, et rendent possibles des collaborations fructueuses.

La voie a été ouverte en Belgique, M. ENGELS s'est plu à le souligner, et la chose est aussi vraie pour notre groupe gantois, par Mme L. HIRSCHBERG (U.L.B.). Ses recherches personnelles, de nature lexicographique, se sont poursuivies principalement au bénéfice des Communautés Européennes, mais elles ont été une première source d'inspiration. Je lui suis personnellement redevable de mes tout premiers contacts avec l'informatique appliquée à la langue : c'était, j'imagine, en l'été de 1964. Il y a eu ensuite le colloque de Nancy (octobre 1964), qui a eu pour suite la création de l'AILA, et de sa section belge l'ABLA, laquelle malheureusement, pour des raisons qui apparaissent maintenant clairement, n'a pas tenu ses promesses, et a cessé pratiquement toute activité.

Les missions d'information de M. J. DE KOCK, alors assistant au service de linguistique française de la R.U.G. (Université de l'Etat à Gand), en tout premier lieu chez Mme HIRSCHBERG, ensuite à Nancy, au *Centre d'étude pour un Trésor de la Langue française,* chez M. P. IMBS, et au *Groupe de Traduction Automatique de la Faculté des Lettres,* chez M. G. BOURQUIN, en mars 1965, à Grenoble au *Centre d'Etudes pour la Traduction Automatique,* chez M. B. VAUQUOIS, en juin 1965, à Milan, au *Centro di Cibernetica e di Attività linguistiche,* chez M. S. CECCATO en mai 1966, à Besançon au *Centre d'Etude du Vocabulaire français* et aux *Archives du Français contemporain* chez M. B. QUEMADA dans la première quinzaine d'octobre 1966, et derechef au C.E.T.A. à Grenoble, en mai 1967. Un stage aux U.S.A., d'avril à juin 1970, que j'avais rendu administrativement possible, a profité à d'autres. Que tous les collègues que j'ai nommés veuillent bien trouver ici l'expression de nos sentiments de reconnaissance pour l'hospitalité qu'ils ont accordée si libéralement à l'un des nôtres. Elle a été déterminante dans la vocation "informatique"de notre centre gantois, sous l'impulsion de M. DE KOCK, en liaison avec Mme R. ZWAENEPOEL-DHANIS, licenciée de sciences mathématiques de la R.U.G., alors attachée comme analyste et programmeuse au *Centrum voor de geautomatiseerde behandeling van Cultuurwetenschappelijke problemen te Gent.* Le secours financier accordé pen-

dant quatre ans (1967-1971) à ce centre facultaire a rôdé nombre d'entre nous au travail d'équipe. Une aide financière de la R.U.G. nous a permis d'entreprendre des tâches précises. Lorsque M. DE KOCK eut cessé d'assurer la liaison, dans notre groupe, entre la linguistique française et l'informatique, suite à son départ en 1969 pour la K.U.L., son rôle fut repris par une équipe d'assistantes dans laquelle n'avait pas tardé à s'intégrer, grâce à un recyclage linguistique, Mme ZWAENEPOEL, devenue assistante au Service de linguistique française. Je n'aurai garde d'oublier que nous avons bénéficié dès le départ d'un préjugé favorable de la part du *Centraal digitaal Rekenlaboratorium* (Bureau de Calcul) de notre Université, et de son fondateur, M. C.C.GROSJEAN. Les liens ainsi noués nous ont valu la collaboration spontanée et très précieuse aussi sur le plan de la linguistique française, d'un assistant du Bureau de Calcul, M. W.BOSSAERT, qui s'est agrégé à notre groupe d'étude. Le "staff" linguistique réunit aujourd'hui, outre le soussigné, Mmes R. VAN DEYCK (ancien français) et R. ZWAENEPOEL (analyse et programmation), et Mlles A. BOONE et D. WILLEMS (linguistique du français moderne). Une partie des perforations nécessaires ont été exécutées par M. W.AUTREVE et Mlle E.BAUWENS. Les réunions de discussions, auxquelles participent parfois d'autres personnes, ont été nombreuses et fructueuses. La publication d'un périodique, TRAVAUX DE LINGUISTIQUE, rend possible la communication rapide des résultats obtenus (3).

3. Nos projets pour l'avenir immédiat s'organisent autour de trois centres d'intérêt : (a) Ancien français. L'information ici utilisée puise dans un genre littéraire, la chanson de geste, dont la versification, la syntaxe et le vocabulaire présentent des caractères particuliers, qui facilitent l'extension à d'autres chansons de geste du T.A. Trois *corpus* sont dès maintenant exploitables : le *Charroi de Nîmes,* le *Couronnement de Louis,* et la *Prise d'Orange;* le premier a fait l'objet d'une publication (1); des méthodes de traitement beaucoup plus fines ont été depuis appliquées à l'ensemble des trois textes. Nous espérons ajouter à notre *corpus* global un *Roland* d'Oxford, et un *Saint Alexis,* en vue notamment d'alimenter en matériaux le cours de grammaire historique de la candidature, grâce à une information diachronique s'étendant sur un siècle et demi. Un peu en marge se situe le T.A. de l'oeuvre de *Villon,* d'après les sources *C* et *I*.

(α) Inleiding.

Het doel van deze studie was, uitgaande van de geponste tekst een concordantie op te stellen, die in hoofdzaak aan twee eisen moest voldoen :

Ze moest volledig zijn, d.w.z. alle woorden voorkomend in de te behandelen tekst (of teksten) moesten er in opgenomen worden;

tevens moest ze, *qua* vormgeving, derwijze opgesteld worden, dat ze gemakkelijk consulteerbaar zou zijn.

Uitgaande van een eenvoudige woordenlijst, die in de loop van de studie herhaaldelijk werd aangepast en verbeterd, resulteerden uiteindelijk twee concordantievormen.

(β) Concordanties.

De eerste concordantie bevat alle woorden, alfabetisch gerangschikt, met vermelding van hun frekwentie en van de referenties die verwijzen naar de tekst, nl. tekstkode, bladzijde en versnummer.

Daar de analyse van deze listing voor de filologen zeer tijdrovend was, bleek het nuttig, naast de referenties, ook een gedeelte van de kontekstvan het woord op te nemen. Eerst kozen wij een half vers, later een volledig vers. De tweede concordantie noemen we "een morfologische concordantie".

Hierin worden, in tegenstelling met de eerste concordantievorm, alle varianten van een woord onder één vorm gegroepeerd. De woorden worden niet meer louter alfabetisch gerangschikt : de verbuigingsvormen van een woord worden niet meer op hun eigen alfabetische plaats ondergebracht, maar wel onmiddellijk na de basisvorm van dat woord, d.i. de infinitief voor de werkwoorden, en het mannelijk enkelvoud voor de naamwoorden (of vrouwelijk enkelvoud voor de vrouwelijke substantieven).

De basisvormen van de woorden vormen evenwel altijd een alfabetische reeks.

Bovendien zijn in deze woordenlijst al de woorden, benevens hun referenties en hun kontekst, vergezeld van een kode : *inf., part. prés., cond. 3, s.m., s.f.pl., art.,* enz. Sommige woorden worden meerdere keren in de woordenlijst opgenomen : samentrekkingen worden gerangschikt onder hun samenstellende delen (vb. *au* = *à* + *le*) ; dubbelzinnige vormen worden gerangschikt volgens hun respektievelijke kodes (vb. *as* = *as ind. prés. 2* + *as* [= *à prép.* + *les art. pl.* ]). Van deze laatste groep worden echter de meest frekwente reeds in de tekst gedifferentieerd door het aanbrengen van aksenten (vb. *le* lidwoord, *lé* persoonlijk voornaamwoord, *lè* < *latum).*

(γ) Voorbereidend werk.

Teneinde de hierboven besproken morfologische woordenlijst te bekomen, moest benevens een voorbereide tekst ook nog bijkomende informatie in de machine ingevoerd worden. Uitgaande van de gepubliceerde woordenlijst van *Charroi de Nîmes,* werden alle woorden uit die tekst geponst met hun kode en met hun basisvorm, waaronder ze moesten geklasseerd worden.

Deze reeks ponskaarten noemen we het "morfologisch woordenboek".
Dit woordenboek werd reeds uitgebreid met de woordenschat van de te behandelen teksten van *Villon*. Daartoe werd een programma geschreven, dat in de tekst alle woorden opzoekt, die niet werden opgenomen in het morfologisch woordenboek. Op dezelfde wijze werd ook een lijst van varianten opgesteld.

(δ) Programmatie.

De programma's werden geschreven in Assembler en in P1/1, en uitgevoerd op de computer van het centraal digitaal rekenlaboratorium van de R.U.G., een IBM-machine 360/30 (64 K.), uitgerust met 3 schijfgeheugens type 2311 en 2 magneetbanden type 2401.
De programma's zijn zeer algemeen en kunnen toegepast worden op elke tekst in om het even welke taal.
De programmatie van het project berust hoofdzakelijk op het herhaaldelijk sorteren en vergelijken van gegevens, komende uit verschillende bestanden, nl. de tekst, het morfologisch woordenboek en de varianten.
Daar bij dergelijke projecten de input omvangrijk is, moeten de bestanden degelijk georganiseerd worden, zodanig dat ze gemakkelijk en vlug kunnen geraadpleegd worden. Wij gebuikten hiervoor disk en tape.
Bij het sorteren en het uitprinten stelde de aanwezigheid van letters, voorzien van aksenten, een probleem. Immers, voor de machine is een letter, voorzien van een aksent, een andere eenheid, en komt dus bij de alfabetische rangschikking niet op de plaats waar hij, naar traditionele normen, verwacht wordt. Deze moeilijkheid werd omzeild door een sorteerwoord in te voeren, d.i. een fictief woord, bestaande uit dezelfde opeenvolging van letters, maar zonder aksenten.
Het uitprinten van letters met aksenten werd opgevangen door een speciale schrijfstaaf, die voorzien is van alle letters, die voorkomen in de franse taal.

La limitation pour l'essentiel à des textes décasyllabiques assonancés répond à la préoccupation de ne pas empiéter sur les recherches similaires d'autres centres, consacrées soit aux textes rimés octosyllabiques, soit aux textes en prose. De tels centres existent, fonctionnent, et produisent. Une collaboration sous la forme d'échange d'information à partir de listings de routine est certainement souhaitable, et nous espérons la voir se réaliser. Toutes nos réalisations en ancien français se trouvent placées sous la responsabilité immédiate de Mmes R. VAN DEYCK et R. ZWAENEPOEL.

(b) Français contemporain. Nous sommes surtout intéressés par la syntaxe, d'où nos besoins considérables en matière d'information semi-traitée. Nos sources sont de deux espèces :

(a) La mobilisation des 40.000 exemples de l'*Essai de grammaire de la langue française* de Damourette et Pichon, à quoi pourraient venir s'ajouter plus tard les *corpus* sélectifs de diverses grammaires et de divers dictionnaires. Pour rendre possible l'utilisation des exemples, et pour être en mesure, à partir de là, de poser à la machine des questions pertinentes et de sélectionner l'information adéquate, nous avons conçu un certain nombre d'instruments de travail dits de routine :

- une liste numérotée de tous les exemples de l'E.G.L.F. avec pour chaque exemple la spécification de l'auteur et du genre littéraire; la numérotation est destinée à assurer les renvois en circuit fermé;
- une liste dans l'ordre alphabétique de tous les auteurs cités avec leur date de naissance, les titres de leurs oeuvres, ainsi que les références bibliographiques complètes.

Ces instruments de travail de routine doivent permettre par la suite une sélection des exemples sur les bases suivantes (l'énumération n'est pas limitative) : quels sont les exemples des conjonctions *comme, quand, si* représentés dans les sept volumes de l'E.G.L.F. : pour l'ensemble du *corpus,* ou seulement pour une partie du *corpus :* exemples d'A.France, exemples en vers seulement, exemples de dialogues seulement, en limitant, si on veut, à une période bien précise, par exemple entre 1890 et 1910 ?

L'établissement de ce dernier instrument de travail, conçu et commencé par M. J.DE KOCK, mais qui n'avait guère progressé au moment de son départ, présente des difficultés pratiques considérables qui tiennent notamment aux variations dans la façon de citer les sources dans l'E.G.L.F. Aussi avons-nous décidé, pour tester cet instrument de travail, de nous limiter provisoirement au tome III, avec ses 6700 exemples; cette limitation doit nous permettre d'aboutir dans les prochaines semaines.

Au point de vue statistique, il est évident que l'exploitation de l'E.G.L.F. ne peut donner aucune indication sur les fréquences d'occurrences, mais seulement sur la façon dont Damourette et Pichon ont choisi leurs sources et les ont exploitées.

- des *corpus* particuliers de textes modernes comprenant le texte avec ses phrases numérotées présenté à l'*input,* et la liste complète des mots occurrents avec comptage et contexte

minimum, suffisant pour opérer une première sélection, ce qui permet d'interroger ces *corpus* sur des problèmes particuliers. Voici la liste des textes dépouillés : prose très littéraire : Le Cléziot (J.M.G.), *L'Extase matérielle,* Gallimard, 1967 (financé en partie par le Rijksuniversitair Centrum Antwerpen); prose scientifique de technicité réduite (biologie) : Rostand (J.), *L'état présent du transformisme,* Stock, 1931, 187 pp., *Le droit d'être naturaliste,* Stock, 1963, 413 pp., et *Biologie et Humanisme,* Gallimard, 1964, 301 pp.; français (relativement spontané) parlé par des locuteurs instruits, en table ronde : W. Zwanenburg, *Recherches sur la prosodie de la phrase française,* Leiden, Universitaire Pers, 1965; français parlé de forme oratoire : allocutions radio-diffusées du Général de Gaulle : Moreau (R.) et Cotteret (J.M.), *Recherches sur le vocabulaire du général de Gaulle,* Analyse statistique des allocutions radiodiffusées (1958-1965), Paris, Colin, 1969. Ces deux derniers *corpus* ont été retenus pour une raison bien précise : pour le premier il existe une transcription phonétique et prosodique, pour le second, les enregistrements sur disque sont dans le commerce. La responsabilité immédiate de l'établissement de ces différents instruments de travail incombe à Mlle A. BOONE et D. WILLEMS pour l'analyse linguistique et l'utilisation, à Mme ZWAENEPOEL pour le T.A., à Mlle BOONE pour la comparaison de l'écrit et du parlé.

Aussi, en ce qui concerne la langue moderne, notre centre gantois de T.A. du français est prêt à examiner tout projet de collaboration sous forme d'échange d'information et de programmes, selon des normes à fixer.

(c) Apprentissage de la langue.

Un groupe de travail dirigé par A.BOONE et W.BOSSAERT, auxquels se sont associés MM. WUYTS et Y. DECANNIERES, s'intéresse à une réévaluation critique des matériaux contenus dans les travaux classiques d' A.GREGOIRE, *L'apprentissage du langage,* Les deux premières années, Liège, 1937, et La troisième année et les années suivantes, 1947, réévaluation poursuivie en T.A. et en traitement artisanal, avec apport de données nouvelles, et notamment d'enregistrements.

G. DE POERCK (en français) et R. ZWAENEPOEL (en néerlandais).

(1) DE POERCK (G.), VAN DEYCK (R.), ZWAENEPOEL (R.), *Le Charroi de Nîmes,* chanson de geste, Librairie-Editions Mallier, 27-Saint-Aquilin-de-Pacy (Eure), 1970, 119 et 231 pp., dans *TEXTES ET TRAITEMENT AUTOMATIQUE,* 1.

(2) Deux annonces, d'ailleurs bienveillantes, ont paru dans le *Bulletin critique du Livre français,* n 301, janvier 1971, n 80672, et dans la *Revue de linguistique romane,* t. 34, 1970, pp. 425-26 (P. Gardette).

(3) *Publikaties van de Dienst voor Franse Linguïstiek bij de Rijksuniversiteit te Gent,* Wetenschappelijke Uitgeverij E. Story-Scientia, tome 1, Gent, 1969, 204 pp.; v. en particulier pp. 60-69, *La programmation de la recherche automatisée des formes pronominales,* de J. DE KOCK et R. ZWAENEPOEL-DHANIS. Le tome 2, qui paraîtra dans le cours de l'année, sera publié par notre *Service de linguistique française,* Blandijnberg 2, 9000 Gent, Belgique. La part faite au T.A. deviendra plus importante au cours des années.

*SUMMARY*

The paper gives a survey of three main points :

1. The human problem due to some misunderstanding about the real nature and extent of the routine work in matters of selection and classification of lexical, semantic, syntactic etc. items done by computers. There is no essential difference between the computational processing and the traditional way, but only a difference in speed and accuracy. The real philological problems are in no way affected by their approach. Neither are statistics in every philologist's good graces : their bearing is usually ignored. The usefulness of scholarly editions with classified raw material, but lacking the usual apparatus of emendations, translations and interpretations is denied. This hostile attitude towards computational processing applied to philology must be exposed and contended with.

2. Technicalities about the working out of elaborate concordances of some French *chansons de geste* belonging to the cycle of William of Orange.

3. A mainly historical account of the activities of the Department of French linguistics at the State University of Ghent in connexion with computational processing, chiefly in the field of contemporary French.