# Préface

Ce numéro est la suite logique de la journée « Traitement automatique des langues et langues anciennes », organisée par l'ATALA en mai 2005. Il s'agissait de faire le point sur l'utilisation des techniques de traitement automatique dans le cadre des langues anciennes. Le présent ouvrage permettra de mesurer les avancées réalisées ces quatre dernières années.

Y a-t-il lieu de faire des « langues anciennes » un objet spécifique pour le TAL ? Il conviendrait d'abord de définir le terme. La tâche n'est pas si simple qu'il y paraît ; on en fournira aisément des prototypes (le latin, le sumérien, l'égyptien ancien...) ; on inclura sans trop d'hésitation le traitement de l'arabe médiéval ou de l'ancien français. Mais où s'arrêter ? Au seizième, dix-septième siècle ? Devions-nous raisonner en termes d'états de langue ? Y a-t-il finalement une unité qui permette de fédérer les travaux présents dans ce volume ?

En fin de compte, le domaine des langues anciennes se délimite sans doute essentiellement par la communauté qui s'y reconnaît. Il y a là un ensemble de pratiques scientifiques similaires, articulées autour de la philologie. Le, ou les textes originaux, sous leur forme physique (stèle, manuscrits, etc.) y sont déjà des objets d'étude[1]. Dans de nombreux cas, la grammaire et le lexique des langues concernées ne sont encore qu'imparfaitement connus. Pour la linguistique en général, l'ancienneté – et souvent la longévité des langues en question –, permet d'entreprendre l'étude de la diachronie sur le long terme. Ce dernier point est explicitement envisagé dans certains des projets présentés ici (Candido et *al.*, par exemple).

Pour de nombreuses langues, il n'existait pas, jusqu'à très récemment, de corpus numérisé un tant soit peu exhaustif. Il semble que de nombreux projets aient vu le jour dans les dernières années pour tenter de combler cette lacune.

Mieux, la communauté des linguistes s'approprie de plus en plus les outils du traitement automatique des langues. Les articles présents dans ce volume montrent que les bases qui se constituent aujourd'hui dans le domaine sont dotées de systèmes d'annotation riches, mettant en œuvre toute la palette des outils possibles. Les textes sont lemmatisés, passent dans des analyseurs syntaxiques ; on étiquette leur structure thématique, etc.

1. Glenisson Jean, Irigoin Jean, éd. La pratique des ordinateurs dans la critique des textes, Paris, 1979.

**Historique du domaine**

Le pionnier en la matière est le père Roberto Busa[2] qui, dès 1948-1949, avec l'aide d'IBM, entreprit la constitution d'un corpus des œuvres de saint Thomas d'Aquin, destiné à créer des concordances et des index. Comme d'autres projets pionniers en la matière, il est toujours d'actualité. Son site Internet est hébergé à l'université de Navarre[3]. L'informatisation du grec ancien est abordée dès 1956 par Leonard Brandwood, en Angleterre, dans une optique de lexicométrie.

Toujours pour les langues classiques, des corpus à visées quasi-exhaustives se développent dans les années 60. À Liège, le *Laboratoire d'analyse statistique des langues anciennes* (LASLA) s'occupe, dès 1961, des textes littéraires latins, puis, à partir de 1965, du grec ancien. Les textes font déjà l'objet d'une analyse automatique, avec le développement d'outils de lemmatisation et d'analyse morphosyntaxique.

Le *Thesaurus Linguae graecae* (TLG) est développé en 1972 par Theodore G. Brunner. Ce corpus est à l'origine du projet *Perseus*, très largement diffusé, sur CDROM d'abord, puis par Internet.

Pour l'égyptien ancien, les premiers travaux sont ceux de W. Schenkel sur le projet M.A.A.T, *Maschinelle Analyse Altägyptischer Texte* (1967) et la création d'une base lemmatisée d'une partie des textes des sarcophages (qui constituent l'un des corpus fondamentaux pour cette langue)[4]. En 1984, le groupe informatique et égyptologie produit à partir des travaux de Jan Buurman, un système de codage informatique des hiéroglyphes. Néanmoins, le codage est encore lent et les bases sont rares. Parmi les projets actuels, citons le *Thesaurus Linguae Aegyptiae* de Berlin, qui fédère un certain nombre de projets plus spécialisés, dans le but d'établir une base lexicale de la langue, et le projet Ramsès de l'université de Liège.

Tout naturellement, les années 80 voient l'extension de ces pratiques et son passage des serveurs universitaires aux micro-ordinateurs ; l'idée de corpus numérisés se popularise. C'est aussi l'époque où l'INaLF lance ce qui deviendra le *Dictionnaire du moyen français*[5], sous la direction de Robert Martin. Ce projet

2. Voir par exemple Roberto Busa, « Rapida e meccanica composizione e pubblicazione di indici e concordanze di parole mediante macchine elettrocontabili », dans Aevum, 25 (1951), 6, p. 479-493). Voir aussi du même auteur « Rapidissima composizione di indici e concordanze di parole mediante schede perforate », dans La documentazione in Italia. Atti del XVIII Congresso mondiale di documentazione, Roma, 1952, p. 95-97.

3. http://www.corpusthomisticum.org/

4. W. Schenkel « Neue Linguistische Methoden », *Textes et langages de l'Égypte pharaonique (1972)*, p. 167-176.

5. Willy Stumpf, L'informatique et la documentation du dictionnaire du moyen français (D.M.F.), *Le médiéviste et l'ordinateur* 25, 1992, p. 2-5.

existe toujours aujourd'hui dans le cadre plus large du laboratoire ATILF (Analyse et traitement informatique de la langue française), à Nancy.

À cette époque, les démarches de traitement automatique ne sont encore qu'assez marginales, l'urgence étant la création de corpus. Au cours des années 90, les problématiques de structuration des corpus et de représentation des manuscrits se sont développées, autour de SGML et de la TEI tout d'abord, puis d'Internet très rapidement ; citons pour exemple le projet, démarré en 1990 à Princeton par Alfred Foulet et Karl. D. Uitti autour du roman de Chrestien de Troye, *Lancelot ou le chevalier de la charrette*, et placé sur Internet dès 1995[6].

## Présentation des articles

Une comparaison entre le programme de la journée thématique sur le même sujet organisée en 2005 par l'Atala et les articles du présent recueil montre une indéniable évolution. Les techniques classiques de TAL sont de plus en plus utilisées par les équipes qui travaillent sur des langues anciennes. Aux simples corpus en texte intégral se substituent aujourd'hui des bases de textes lemmatisées ou arborées ; leur développement se fait dans le cadre d'une problématique de plus en plus sophistiquée.

La première partie de ce volume se compose de quatre articles qui traitent justement de la constitution de corpus, et sont représentatifs de ces évolutions.

L'article de Haug *et al.*, présente l'informatisation d'un corpus de versions parallèles du *Nouveau Testament* dans diverses langues indo-européennes. La problématique du projet est clairement posée : il s'agit de pouvoir étudier cinq grands types de phénomènes linguistiques dans les langues concernées. Les annotations sont assez riches, puisqu'elles vont de la lemmatisation à la notation de la structure informationnelle en passant par la syntaxe. L'étude de la conception du corpus et l'analyse détaillée des solutions envisagées permettent d'utiliser cet article comme état de l'art.

L'article de Petrova *et al.* expose la création d'un corpus de vieil allemand, en vue d'étudier l'interaction entre le plan énonciatif/hiérarchique et la syntaxe dans une perspective d'évolution diachronique. L'architecture du système et les outils utilisés sont décrits en détail, depuis la saisie des textes et le lien aux manuscrits originaux jusqu'aux mécanismes d'interrogation.

Candido et Aluísio décrivent un corpus du portugais beaucoup plus récent, puisqu'il s'étend du seizième au dix-neuvième siècle. L'ensemble des étapes nous est présenté, de la saisie des textes à la constitution du dictionnaire, avec une discussion des outils existants. Deux points sont plus particulièrement développés :

6. http://www.mshs.univ-poitiers.fr/cescm/lancelot/projet.html

la lemmatisation, particulièrement importante en diachronie, et l'édition des entrées du dictionnaire.

Enfin, l'article de McGillivray *et al.* propose un travail sur le plus ancien corpus électronique, l'index Thomisticus ; il s'agit de créer un corpus arboré ainsi qu'un dictionnaire valenciel. Plusieurs approches de l'analyse syntaxique automatique du latin sont exposées et comparées. Le corpus arboré sert à obtenir des informations plus fines sur la valence des verbes latins.

Nous avons placé ensuite plusieurs travaux focalisés sur la lemmatisation et l'analyse morphologique.

Poudat et Longrée examinent le comportement de différents systèmes de lemmatisation du latin. Il est intéressant, non seulement par ses résultats, mais aussi par le protocole rigoureux qu'il met en œuvre.

Souvay et Pierrel étudient la lemmatisation du moyen français, avec tous les problèmes orthographiques que cela présente. Ils utilisent des règles de réécriture pour représenter, tant les phénomènes morphologiques que les variantes diachroniques.

Enfin, nous changeons de famille de langues avec l'article de Barthélémy, qui porte sur l'analyse morphologique de l'akkadien. Le formalisme utilisé, dénommé par l'auteur « morphologie multi-grain », met en œuvre des techniques à état fini, multi-bandes. Il est décrit en détail, et comparé aux autres systèmes d'analyse morphologique des langues sémitiques.

L'article de Kondrak nous ramène à des problèmes généraux de linguistique historique. Il expose un algorithme permettant d'identifier les cognats et les correspondances phonétiques dans des lexiques, en combinant trois types d'informations : les ressemblances phonétiques, les correspondances phonologiques, et la proximité sémantique.

Enfin l'article de Nederhof qui clôt ce numéro décrit un algorithme pour permettre une mise en page automatique et optimale de données alignées, même dans des cas complexes.

**Remerciements**

Nous tenons à remercier les relecteurs sans qui cet ouvrage n'aurait pu paraître : François Barthélémy, Mahé Ben Hamed, Francesco Citti, Gérard Huet, Wojciech Jaworski, Bastien Kindt, George Kiraz, Christiane Marchello-Nizia, Nicolas Mazziotta, Sylvie Mellet, Remo Mugnaioni, Mark-Jan Nederhof, Mark Olsen, Gerald Penn, Sophie Prevost, Wolfgang Schenkel, Richard Sproat, Achim Stein, Paul Tombeur, Laurence Tuerlinckx, Jerzy Tyszkiewicz et Jean Winand.

Bastien Kindt, Nicolas Mazziotta, Mark-Jan Nederhof et Sophie Prevost nous ont été d'un grand secours ; qu'ils trouvent ici l'expression de notre gratitude.

Joseph Denooz
Laboratoire d'analyse statistique des langues anciennes (L.A.S.L.A.),
Université de Liège
*Joseph.Denooz@ulg.ac.be*

Serge Rosmorduc
Équipe langues et littératures de l'Égypte ancienne,
ÉPHÉ IV$^{e}$ section/IUT de Montreuil,
Université Paris 8
*serge.rosmorduc@qenherkhopeshef.org*