

The Académie Sample Database

Russon Wooldridge¹ and Isabelle Leroy-Turcan²

¹ Trinity College, 6 Hoskin Avenue, Toronto M5S 1H8, Canada.

² Ferme de la Roche, Romans, 01400 Chatillon-sur-Chalaronne, France.

KEYWORDS: early dictionary base, source text base, dictionary-text base

AFFILIATION: ¹ University of Toronto,

² Université de Lyon III

E-MAIL: ¹ wulftric@epas.utoronto.ca

FAX NUMBER: ¹ 1-416-978-4949

PHONE NUMBER: ¹ 1-416-978-2885

The Académie Sample Database

1. Description of the research project

The Académie Sample Database (ASD) forms part of the international *Dictionnaire de l'Académie française* Computerization Project, which has as its object the creation of a database of the eight complete editions of the *Dictionnaire de l'Académie* (1694–1935). The three main components of the ASD are the Dictionary Base, the Text Base and the Critical Base, the last comprising expert notes written by members of the project team and theoretical texts written by contemporaries of the various editions of the dictionary; others include a bibliographical base, an image base and a metalinguistic keyword base (Wooldridge 1994; Wooldridge & Leroy-Turcan 1995; Leroy-Turcan 1996a and 1996b). In the present paper we wish to concentrate on the Dictionary Base and the Text Base of the Sample Database.

The Dictionary Base (DB) and the Text Base (TB) are complementary, the one representing a description of the language (*langue*), the other the discursive usage on which the description is based (*discours*). The ASD is modelled on the Renaissance dictionary/source-text bases RenDico and RenTexte comprising the dictionaries of Estienne and Nicot and the texts of some of their 16th-century French sources (Wooldridge 1995). In the case of the *Dictionnaire de l'Académie*, the sources are in principle the Academicians themselves: the *Dictionnaire* proudly declares that it has no need to use quotations since the best writers of French are those engaged in the writing of the dictionary!

The purpose of creating the three sample databases is, on the one hand, to test the model before

committing ourselves to a fixed methodology for the global project, and, on the other, to provide usable data for the study of dictionary methodology and the history of the language – the *Dictionnaire de l'Académie* is unique in that it gives eight synchronic descriptions of the language, encompassing 240 years, and constitutes the linguistic norm of French.

The ASD-DB contains a selection of articles, the same for each edition, representing approximately 1% of the whole dictionary. The selection criteria are that the sampling contain both semantic and function words, that it be representative of the alphabetical divisions of the text (beginning-middle-end), that it include sequential entries (blocks), that it contain words of cultural significance, and that it cater to some extent to the particular interests of the database authors (academic researchers and students). The chosen entries, all entered and on-line (see section 3), are the following: *acanthé*, *âme*, *cloche* to *clochette*, *douaire* to *douzil*, *gagner*, *gras*, *gros*, *loin* to *loisir*, *loup* to *louvre*, *que*, *queue*, *tige* to *tintouin*, *vent*, *vin*, *voler*.

The ASD-TB comprises short texts or extracts from the writings of a number of major and minor writers of French prose and poetry, all of them members of the Academy. The choice of texts is based on several criteria: diachronic coverage (comparable volume for each edition of the dictionary); historical representativity of usage (based on the role that various Academicians played in the preparation of each edition of the dictionary); the occurrence of a majority of the words included in the Sample Dictionary Base; availability. Among the better-known names of the several dozen it is hoped to include are: Balzac (Guez de), Bossuet, Buffon, Chateaubriand, Condorcet, Corneille, Cuvier, France, Hugo, La Fontaine, Lamartine, Marivaux, Mauriac, Mérimée, Montesquieu, Musset, Perrault, Racine, Renan, Romain, Sainte-Beuve, Tocqueville, Valéry, Voltaire.

2. Database structure and search typology

The dictionary is tagged for headword, co-headword, headword variant, main part of speech, paragraphing, typography, edition, page and column. The unpredictability and ambiguity of microstructure fields has led us to prefer the use of a list of lemmatized metalinguistic keywords – e.g. *masculin* for references to masculine gender, *signifie* for definition copulas, *familier* for colloquial usage labels – to a systematic, and subjective, tagging of information fields that would distort the text, particularly in the early editions. A complement to metalinguistic keywords is provided by typographical discrimination: definitions are always in roman, examples in italic. Links are made

for each headword to occurrences in the text base, and other links are made for headwords or sub-entries to the critical base and to images (e.g. the history of the word *feuille d'acanthé* or graphical representations of the acanthus leaf in architecture).

Dictionary data retrieval can be either full-text searching, with optional filtering by tagged fields (edition, headword, typography, etc.), or entry look-up – the indexed word list contains word occurrences in the first part and headwords in the second (thus tokens *doux* 722, *douce* 353, *douces* 59, headword @*doux* 8).

The texts are tagged for structural division – title, section, paragraph etc. –, book division – page –, and typography. Data retrieval is classical full-text search with optional tag-field filtering.

Concurrent searching of dictionary and texts is achieved simply by combining both types in one global database. The global base constitutes the default corpus; the user can create sub-corpora by restricting particular searches: for example, to dictionaries only, to texts only, to 18th-century dictionary editions and texts, to dictionary edition A and texts M and N, etc.

3. The ASD on-line

The ASD is currently using the World Wide Web as a design tool. For the moment, searching is simulated by links from selected items to occurrences; these latter are preformatted in KWIC, extended context and distribution displays. It is planned to use a version of PAT as a search engine for the on-line version, and to distribute the finished ASD both on-line and on CD-ROM. The WWW version – currently including all of the selected dictionary entries and lists of meta-linguistic keywords linked to preformatted displays of occurrences – can be accessed at <http://www.epas.utoronto.ca:8080/~wulftric/academie/>.

4. The complementarity of the Dictionary Base and the Text Base

The principal significance of the combined dictionary-text database is the comparison it allows between codified usage (the dictionary) and natural usage (the texts). Since the *Dictionnaire de l'Académie* is both normative and conservative, one can expect to find in text bases such as Fran-text and ARTFL many examples of usage either condemned or ignored by the *Dictionnaire*. One can also expect that for a number of lexical items the Academicians themselves, like all speakers, who have the two basic registers of formal and informal use, will say one thing in the dictionary and do another in their writings.

For example, the adjective *timoré* “timorous” is treated in the dictionary from 1694 to 1878 as

applying almost exclusively to the fear of offending God. From 1694 to 1762 the two collocates given by the examples are *âme* “soul” and *conscience*, both feminine. The edition of 1762 adds the remark that the word is used almost exclusively in the feminine form. From 1798 to 1878, the masculine collocate *il* “he”; is added. The text bases offer examples of usage that conform to the pronouncements of the dictionary, and others that do not. Bossuet (1685) gives *conscience timorée*; Montesquieu (1755) uses the masculine *timoré* to qualify the pronoun *vous* “you”; Voltaire (1776), writing about the Bible, gives two occurrences of *âme(s) timorée(s)*. In all of the preceding cases *timoré* is used in reference to the fear of God. In an earlier text (1755), Voltaire gives an example in which, as will become increasingly the case, *timoré* is used simply in reference to a person's character or behaviour: *main timorée* “hand”. Similarly, Sainte-Beuve (1834) *quelque chose de timoré* “something”; Chateaubriand (1848) *corruption timorée*.

In the 6th edition (1835), the *Dictionnaire* states that *tillac* “upper deck”; is almost always used in referring to merchant vessels. Chateaubriand uses the word 11 times in his *Memoirs* (1848) in reference to merchant ships, passengers ships and naval vessels.

The word *timbre* acquires new senses with each edition. The meaning “postage stamp” is expressed by *timbre-poste* in the 7th (1878), with the 8th (1935) adding the elliptical *timbre*. Obviously the dictionary is recording established usage that can be observed in earlier texts. The earliest attestation of *timbre-poste* in the 1,880 texts of the ARTFL database is 1863 (Goncourt brothers); Hugo uses it several times in the volume of his correspondence published in 1866. In the same volume he uses the elliptical form *timbre* once (69 years before the Académie); by the following volume (1873), the shortened form has become more frequent than the full one.

5.

The computerization of early dictionaries is quite recent (Wooldridge 1985). Pruvost (1995: 17) notes the landmark significance of the 1993 Toronto Colloquium on Early Dictionary Databases (Lancashire & Wooldridge 1994). Lancashire (1992) is preparing an English Renaissance Knowledge Base with similar aims to those of the Académie project. The philological care taken in representing faithfully the original texts allied to the technological sophistication that is now the norm in Humanities computing make it possible to create research resources that give scholars full access to early texts without having to depend entirely, as in the past, on repeated partial linear readings or on the filtered and diachronically mar-

ked interpretations of historical dictionaries (such as the *OED* or the *TLF*).

6.

The paper will comprise an illustrated description of the structure and data of the combined ASD dictionary-text database.

References

- I. Lancashire (1992). "Bilingual Dictionaries in an English Renaissance Knowledge Base", *CCH Working Papers*, 2: 69-88.
- I. Lancashire & T.R. Wooldridge (1994). *Early Dictionary Databases*. *CCH Working Papers*, 4. [The papers of this volume are being reissued in electronic form in the *CH Working Papers*. See, for example, I. Leroy-Turcan on the computerization of Ménage's etymological dictionary, *CHWP*, B.10, at <http://www.chass.utoronto.ca/epc/chwp/>]
- I. Leroy-Turcan (1996a). "Conflits de générations et usages littéraires concurrents dans la première édition du *Dictionnaire de l'Académie française* (Paris, Coignard, 1694)", lecture given at University of Toronto, February.
- I. Leroy-Turcan (1996b). "Modalités de mise en oeuvre de l'informatisation de la première édition du *Dictionnaire de l'Académie française* (1694)"; paper given at Les Journées "Dictionnaires électroniques du français des XVIe et XVIIe siècles", Clermont-Ferrand, June.
- J. Pruvost (1995). "Un demi-siècle d'or pour les dictionnaires de langue français", *Actes du Colloque 1994 La Journée des dictionnaires*. Université de Cergy-Pontoise, Centre de Recherche Texte/Histoire: 5-22.
- T.R. Wooldridge (1985). *Concordance du Trésor de la langue françoise de Jean Nicot (1606): matériaux lexicaux, lexicographiques et méthodologiques*. Toronto, Éditions Paratexte.
- T.R. Wooldridge (1994). "Projet d'informatisation du *Dictionnaire de l'Académie* (1694-1935)", to appear in the Proceedings of the *Colloque du tricentenaire du Dictionnaire de l'Académie française* (Institut de France, Paris, November 1994).
- T.R. Wooldridge (1995). "Bases dictionnaires, bases philologiques, bases de connaissances culturelles", to appear in the Proceedings of the *Colloque "Autour de l'informatisation du Trésor de la langue française"* (Nancy, May 1995).
- T.R. Wooldridge & I. Leroy-Turcan (1995). "Metalinguistic Keywords as a Structural Retrieval Tool for Early Dictionaries", *JADT 1995*, Rome.