Tagging early dictionaries with SGML by adapting the Text Encoding Initiative Guidelines: Basnage de Bauval's *Dictionnaire Universel*

Agnès TUTIN et Chantal WIONET, Lille et Grenoble, France

Abstract

In this paper[1], we present an ongoing project of tagging early French dictionaries, *Le dictionnaire Universel* de Basnage de Bauval (1701) in order to build a lexical database which would enable fine-grained queries for linguistic and historical studies.

The tagging model is SGML and as a tagset basis we chose to adopt the Text Encoding Initiative Guidelines for print dictionaries. We show that in spite of numerous inconsistencies, computerizing *Le dictionnaire Universel* is a feasible task and exhibits many regularities in the text which can be partially automated with finite state automata. Applying a systematic grammar on the text renews the metalexicographic studies of early dictionaries.

1 Computerizing early dictionaries by using the Text Encoding Initiative Guidelines

In the past few years, much research has been carried out into computerizing 17^{th} and 18^{th} century French dictionaries : some dictionaries like Jean Nicot's *Thresor de la langue françoyse* (1606) by T.R. Wooldridge, from Toronto University and some versions of the *Dictionnaire de l'Académie française* (1694-1935) (T.R. Wooldridge, Toronto University; Isabelle Leroy-Turcan, Université de Lyon III) already are available on Internet. Computational versions of these kinds of text are of high interest both for language historians and for dictionary researchers.

Making computerized versions of early dictionaries would allow several uses for the scholars in this field : hypertextual browsers, lexical databases, statistical results... But such processing tools require that information classes are previously identified in the dictionary entries, which is far from being straightforward.

For our application, Basnage de Bauval's *Dictionnaire Universel* (1701), we considered it essential that the tagging model should meet the following requirements:

a) It should be flexible and should make it possible to encode simultaneously an editorial perspective (the original typography is retained) and a lexical perspective (where information classes are identified) [Ide & Véronis 95].

b) It should be widely used by the academic community so as to be implemented in several computing tools and to enable interchange with other scholars.

These requirements led us to adopt the Standard Generalized Markup Language and the Text Encoding Initiative Guidelines [Sperberg-McQueen & Burnard, 1994]. As a first tagset for our dictionary, we chose to adopt the DTD recommended in the TEI guidelines for contemporary dictionaries. This really represented a challenge insofar as early dictionaries appear to have a fuzzy structure compared to modern dictionaries.

The DTD for print dictionaries was designed for modern standard monolingual and bilingual dictionaries and not for early unstructured ones. Given that the dictionary structures are extremely variable, the TEI DTD is designed to be flexible in order to capture this variety : (i) no specific order is required between the elements, (ii) the description of many elements remains quite general, and (iii), a large set of attributes can be used to refine the description if necessary.

The TEI guidelines have been successfully adapted to simple contemporary dictionaries such as the *Petit Larousse Illustré* [Tutin & Véronis 98], even if one must not underestimate the extent of the adapting task which is far from being straightforward. The tagset proposed for the print dictionaries could be quite easily adapted to our tagging project even if some modifications appeared as essential (see section 4).

2 Principles for computerizing early dictionaries

2.1 The DTD provides a framework for metalexicographic analysis

Tagging early dictionaries is not a straightforward task. The microstructure of the dictionary is not easily brought to light as the typographic conventions and the metalinguistic descriptions are not used in a consistent way and one has to refrain from modelling the analysis from a contemporary point of view. The analysis should be carried out primarily from a content analysis and requires a deep understanding of the entry fields.

The tagging process is not neutral and obviously constrains the analysis performed. This appears inevitable and even desirable in order to allow queries to be made easily on the text and also to reveal the regularities and irregularities of the microstructure. For example, we think that the part of speech field should be systematically tagged even if in some rare entries, this information does not appear. The lack of information (an empty element in SGML markup) may be in itself an interesting piece of information for the dictionary researcher, and should be annotated according to us.

2.2 The typographic information should be dissociated from the content information

The tagging process we advocate tries not to misinterpret the textual content. The original text will in any case be retained with all its typographic properties. Given that the typography cannot be systematically relied on (it is not used consistently in the different kinds of fields), the formal aspect and the informational content are dissociated. The typographical properties are marked with attributes, while the informational content is mostly tagged with elements. For example, the mentioned word in examples which generally appears in italics receives the attribute REND and the value "it" while the attribute of the <author> receives the value "smc" (small capitals).

D'ABORD-QUE [...] *D'abord qu'*il le vit, il lui voulut donner un coup de bâton. ABLAN. [...]

...

<PCIT> <Q><OREF **REND="it"**> D 'abord-qu'</OREF> il le vit, il lui voulut donner un coup de bâton.</Q> <BIBL> <AUTHOR **REND="smc"**> Ablan. </AUTHOR> </BIBL> </PCIT>

...

2.3 Attributes are used to standardize the lexical information

The content tagging is as far as possible standardized with the use of attributes and fixed values but the textual content is not modified. For example, the metalinguistic markers concerning the part of speech and the gender are standardized with an attribute TYPE. This will make the consultation of the text easier (several metalinguistic marks can be used for the same information).

DEBARASSÉ, ÉE. part.pass. & adj. [. . .]

<GRAMGRP>

<POS TYPE = "PPSE"> part. pass. </POS>

</GRAMGRP> &

<GRAMGRP>

<POS TYPE = "ADJ"> adj. </POS>

</GRAMGRP>

When the grammatical field is lacking in the text, it is thus provided by the means of attributes. For example, in the following entry DÉCISIF, the part of speech has been omitted. It is supplied by the attribute TYPE and the value "adj".

DECISIF, IVE. Qui decide; qui resoud. [. . .]

[...]

<FORM TYPE="LEMMA">

<ORTH REND="caps"> DECISIF</ORTH>

<INFLEX REND="smc" GEN="fem" NUMBER="sing"> , ive.</INFLEX>

</FORM>

<GRAMGRP>

<POS TYPE="ADJ"> </POS>

</GRAMGRP>

[...]

2.4 The tagging process is as far as possible automated

The tagging process has been automated as far as possible with the use of finite state automata. Relying on formal markers enables one to develop rigorous and reproducible methods. Several elements (typography, the order of the fields within the entry, the metalinguistic markers) seem regular enough to enable several fields to be semi-automatically tagged. This kind of process allows us (i) to formalize very precisely the patterns encountered in the entries, (ii) to avoid inconsistencies in the tagging process, (iii) and to shorten the treatment. The transducers are built with the help of the *Intex* system [Silberztein 93].

For example, a transducer associates specific tags for metalinguistic markers. Thus, e.g. the string ". s. f." will automatically be tagged <GRAMGRP> <POS type="s"> s.</POS> <GEN type="f"> f.</GEN> </GRAMGRP> . Such automata have been systematically built for the key-words, the related entry, the grammatical part, the usage marks and the cross references.

3 Adaptations of the TEI structure

The entry structure in early encyclopaedic dictionaries is slightly different from contemporary ones. In the following example, we can see that homonymic and derivative entries (DECOUPLÉ) are not numbered but are treated in specific subentries, while collocations are treated in specific paragraphs.

DECOUPLER v. act. Detacher des chiens couplez, particulièrement pour les lâcher après le gibier. *Decouplez* vite vos chiens.

Decoupler, se dit figurément, des gens qu'on employe dans la poursuite de quelques affaires. [...]

DECOUPLÉ, ÉE. part.pass. & adj.

On dit aussi d'un vert galant, d'un jeune homme bien taillé & bien vigoureux, Qu'il est bien *découpl'e.*

The entire entry is tagged with <entry> and subentries are annotated with the help of <re> (related entries). For collocation paragraphs and etymological unstructured comments, specific elements have been added (<CollGrp> and <CEtym>[2]). The entry structure is represented by Fig. 1[3].

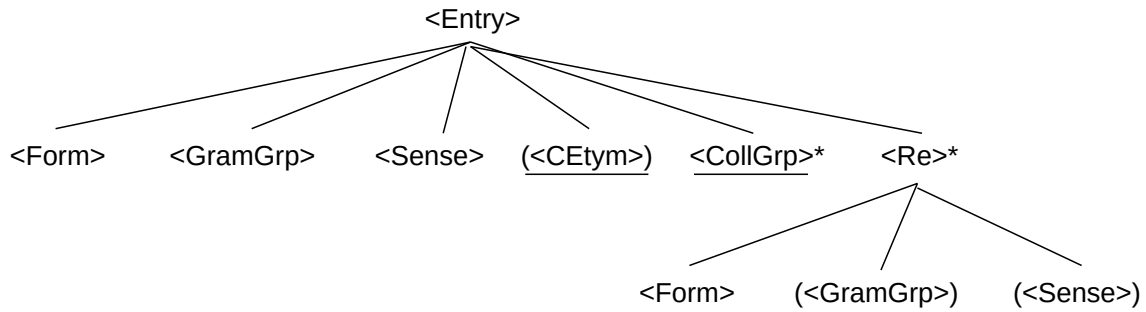

Figure 1: Structure of entries in Basnage's *Dictionnaire Universel*

Most of the elements presented in Fig. 1 are split up into elements which cannot be described here for lack of space. Most TEI elements were retained, while we had to add some specific elements to deal with the fuzziness of some fields. This was the case with the usage field which is quite unstructured in early dictionaries. For example, in the following entry, the usage field is a sentence and is not standardized as in modern dictionaries. The usage field is annotated by a specific tag <Cusg> while the introductory words are tagged with <Lbl> and the usage information is tagged by <Usg> .

DAGUER. V. act. Frapper avec une dague. **Ce mot est vieux.**

[...]

<ENTRY> . . .

<CUSG> <LBL> Ce mot est </LBL> <USG TYPE="TIME"> vieux</USG> .

</CUSG>

. . . </ENTRY>

4 Results and conclusion

Our project aimed at tagging a small sample of the Basnage's dictionary (around 360 entries) for evaluating the feasability of the tagging process. The plain text has been keyed in manually, since OCR is out of the question for this kind of early printed text.

The DTD has been developed according to the TEI guidelines, and has been designed to be re-usable for other early encyclopaedic dictionaries (Furetière and Trévoux). Several adaptations and additions proved to be necessary since some fields like usage and etymology are far less structured than in contemporary dictionaries, but surprisingly, most TEI elements were restained.

The tagging process has been partially automated with the help of transducers applied to the plain text. The tags have been checked and completed manually with the help of an SGML editor (Softquad's *Author/Editor*).

Computerizing a sample of the Basnage's *Dictionnaire Universel* proved to be not only feasible but fruitful from the linguistic viewpoint. Contrary to expectations, the text exhibited enough regularities for building a grammar of the fields, the DTD. Using transducers forced us to inventorize the metalinguistic formulations for the different kinds of fields. Here again, we found more consistency than we expected.

Notes

[1] This research is supported by a grant from the "Délégation Générale à la Langue Française".

[2]Elements specific to our application have been underlined.

[3]Abbreviations used:
<Form>: Form of the entry (lemma, inflection),
<GramGrp>: Grammatical information,

<CEtym>: Etymological comment,
<CollGrp>: Comment on collocations,
<Re>: Related entry.

References

[Ide/Véronis 1995] Ide, N., Véronis, J. (1995). Encoding dictionaries. *Computers and the Humanities*, 29:2, 167-179. [reprinted in Ide, N., Véronis, J. (Eds.) (1995). The Text Encoding Initiative: Background and Context. *Kluwer Academic Publishers*, Dordrecht, 342p.].

[Leroy-Turcan 1995] Leroy-Turcan I. (1995), L'informatisation du *Dictionnaire Etymologique ou Origines de la Langue Françoise* de Gilles Ménage (1694), in *Informatique et Dictionnaires Anciens*, Paris, Didier-Erudition, pp. 131-142.

[Silberztein 1993] Silberztein M., (1993), *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*, Paris, Masson.

[Sperberg-McQueen/Burnard 1994] Sperberg-McQueen, C.M., Burnard, L. (1994), *Guidelines for Electronic Text Encoding and Interchange*, Text Encoding Initiative, Chicago and Oxford.

[Tutin/Véronis 1998] Tutin A., Véronis J., 1998, Electronic Dictionary Encoding: Customising the TEI Guidelines, *Actes d'EURALEX 98*, Liège, 3-8 august 1998.

[Wionet/Tutin 1995] Wionet C. & Tutin A. (A paraparaître), *L'informatisation du Dictionnaire Universel de Furetière revu par Basnage (1702): premier bilan*, Champion.

[Wooldridge 1995] Wooldridge T.R. (1995), La base lexicographique du *Dictionnaire de l'Académie française* (1694-1992): quelques mesures, in *Informatique et Dictionnaires Anciens*, Paris, Didier-Erudition, pp. 157-164.

Appendix: a simple entry

DAGUET . Terme de Venerie. Jeune cerf, qui est à sa premiere tête; qui pousse son premier bois.

Daguet. adv. Sourdement; en cachette. Il s'en est allé, il a tiré sur ses chausses *daguet*. Cela est bas & populaire.

```
<ENTRY>
    <FORM TYPE="LEMMA"> <ORTH REND="CAPS"> DAGUET</ORTH> .</FORM>
    <GRAMGRP> <POS TYPE="S"> </POS> <GEN TYPE="MASC"> </GEN> </GRAMGRP>
    <SENSE>
        <CDOMAIN> <LBL> Terme de </LBL> <DOMAIN> Venerie</DOMAIN> . </CDOMAIN>
        <DEF> Jeune cerf, qui est i est à sa premiere tête; qui pousse son premier bois. </DEF>
    </SENSE>
    <RE>
        <FORM TYPE="HOMOGRAPH"> <ORTHRE REND="SMC"> Daguet</ORTHRE> . </FORM>
        <GRAMGRP> <POS TYPE="ADV"> adv.</POS> </GRAMGRP>
```

```
        <SENSE>
            <DEF> Sourdement; en cachette. </DEF>
            <EG> <Q> Il s‘en est allé, il a tiré ses chausses <OREF REND="IT"> daguet</OREF> . </Q>
            </EG>
            <CUSG> <LBL> Cela est </LBL> <USG> bas et populaire </USG> .</CUSG>
        </SENSE>
    </RE>
</ENTRY>
```