Codage TEI des dictionnaires électroniques

Nancy IDE[a,b] et Jean VÉRONIS[b]

[a] *Department of Computer Science*
Vassar College
Poughkeepsie
New York, NY 12601, USA
`ide@cs.vassar.edu`

[b] *Laboratoire Parole et Langage*
Université de Provence et CNRS
29, avenue Robert Schuman
13621 Aix en Provence Cedex 1, France
`veronis@univ-aix.fr`

1. Introduction

La tâche du groupe de travail de la TEI sur les dictionnaires[1] était de fournir un ensemble de conventions au niveau des entrées de dictionnaires, la structuration de niveau supérieur (page de titre, matériau introductif, divisions en noms communs et en noms propres, en langues dans les dictionnaires bilingues, etc.) étant de même nature que dans bien d'autres types de textes. Le groupe de travail a par ailleurs limité son champ aux dictionnaires occidentaux modernes et a testé ses recommendations principalement sur des dictionnaires de taille moyenne, tels que le *Petit Larousse*, le *Petit Robert* ou le *Collins English Dictionary*. Les dictionnaires anciens et les dictionnaires « monumentaux » tels que l'*Oxford English Dictionary* ou le *Trésor de la Langue Française* ont été volontairement laissés de côté pour la première édition des *Guidelines* [3].

2. Composants de base

De nombreux types d'informations clairement identifiables figurent dans les entrées de dictionnaires : informations sur la forme du mot (orthographe, prononciation, césure, etc.), informations grammaticales (catégorie grammaticale, sous-catégorie, morphologie, etc.), définitions ou traductions, étymologie, renvois, sous-entrées, notes d'usage, exemples, etc.

1. Le groupe de travail sur les dictionnaires était composé de Robert Amsler, Susan Armstrong-Warwick, Nicoletta Calzolari, Carol Van Ess-Dykema, John Fought, Nancy Ide, W. Frank Tompa et Jean Véronis.

CRAWLER [krole] v.i. Nager le crawl.

FIGURE 1 – *Entrée d'un dictionnaire (d'après* Le Petit Larousse*)*

La première étape dans la réalisation d'une Définition du Type de Document (DTD) pour les dictionnaires est la spécification d'une typologie des éléments atomiques qui figurent dans les entrées, accompagnée d'une nomenclature adéquate pour ces éléments. Les éléments atomiques sont ceux qui constituent les champs de base spécifiques aux entrées de dictionnaire. Ces éléments ne contiennent aucun autre champ d'information : leur contenu est une séquence de caractères, éventuellement accompagnée d'éléments communs à tous les types de textes (dates, etc.). L'identification des champs fondamentaux d'information dans les dictionnaires avait reçu l'attention de nombreux chercheurs dans le passé et malgré des désaccords sur les détails, les champs d'information fondamentaux étaient relativement bien établis avant le travail de la TEI (voir par exemple [1, 4]).

Certains éléments de dictionnaires sont complexes, c'est-à-dire constitués de groupes d'éléments atomiques. Considérons, par exemple, la définition de la figure 1.

Cette entrée comporte trois parties distinctes : les informations relatives aux formes écrite et parlée de la vedette, les informations grammaticales et la définition. Dans de nombreux cas, il convient de rendre explicites ces associations ou regroupements ; à cette fin, nous avons défini un ensemble de balises groupantes permettant le marquage de relations logiques entre éléments. Ainsi, le codage de l'entrée ci-dessus serait :

```
<entry>
  <form>
    <orth>crawler</orth>
    <pron>krole</pron>
  </form>
  <gramGrp>
    <pos>v</pos>
    <subc>i</subc>
  </gramGrp>
  <def>Nager le crawl</def>
</entry>
```

La première information comporte deux sous-parties, marquées par les balises *<orth>* et *<pron>* ; la balise *<form>* assure leur association logique. De

roughcast ('rʌfˌcɑ:st) *n.* **1.** a coarse plaster used to cover the surface of an external wall. **2.** any rough or preliminary form, model, etc. *~adj.* **3.** covered with or denoting roughcast. *~vb.* **-casts, -casting, -cast.** **4.** to apply roughcast to (a wall, etc.). **5.** to prepare in rough. **6.** (*tr.*) another word for **rough-hew.** **- -'roughˌcaster** *n.*

FIGURE 2 – *Une entrée du* Collins English Dictionary

la même manière, le composant <*gramGrp*> comporte deux sous-composants, la catégorie grammaticale (<*pos*> pour *part-of-speech*) et les informations de sous-catégorisation (<*subc*>). La définition est un composant atomique, constitué du seul texte de définition, sans structure interne.

3. Structure hiérarchique et portée

D'une façon quasi-systématique, les entrées de dictionnaires sont structurées de façon hiérarchique : une entrée comporte souvent deux ou plusieurs sous-parties, chacune correspondant à des homographes grammaticaux, qui peuvent se subdiviser à nouveau en sens et sous-sens (figure 2). Dans certains cas, un ou plusieurs niveaux peuvent être absents (par exemple, le niveau des homographes grammaticaux).

L'organisation hiérarchique des dictionnaires permet la factorisation des informations sur certains niveaux de la hiérarchie. Les informations ont donc une portée, comme les variables d'un langage informatique structuré en blocs tel que Pascal : les informations précisées à un niveau donné de l'hiérarchie s'appliquent à tous les niveaux emboîtés. Dans les dictionnaires, les informations relatives à la prononciation, à la forme orthographique, à la catégorie grammaticale, etc. sont généralement mises en facteur à la tête de l'entrée car elles s'appliquent aux différents sens. Par exemple, dans l'entrée « roughcast » de la figure 2, l'orthographe et la prononciation s'appliquent à l'entrée entière, « nom » s'applique aux trois premiers sens, etc. Le codage SGML reflète cette structure :

```
<entry>
  <form>
    <orth>roughcast</orth>
    <pron>'r^f,ca:st</pron>
  </form>
  <hom>
```

```
    <gramGrp>
      <pos>n</pos>
    </gramGrp>
    <sense n='1'>
      <def>a coarse plaster used to cover...</def>
    </sense>
    <sense n='2'>
      <def>any rough or preliminary form, model, etc.</def>
    </sense>
  </hom>
  <hom>
    <gramGrp>
      <pos>adj</pos>
    </gramGrp>
    <sense n='3'>
      <def>covered with or denoting roughcast.</def>
    </sense>
  </hom>
  ....

</entry>
```

4. Problèmes et difficultés

Les dictionnaires figurent parmi les types de textes les plus complexes traités par la TEI. Chaque entrée d'un dictionnaire est un objet fortement structuré, dans lequel de nombreux mécanismes d'abréviation et d'organisation typographique permettent une présentation condensée des informations. De plus, la structure des entrées de dictionnaires varie considérablement d'un dictionnaire à l'autre et dans un même dictionnaire : il semble presque que l'on puisse trouver n'importe quel type d'information à n'importe quelle position d'une entrée dans un dictionnaire ou un autre. Toutefois, malgré ces variations, les lecteurs humains sont capables d'interpréter relativement aisément les entrées de dictionnaire et ce, le plus souvent sans consulter les explications introductives. Il est donc clair qu'il existe un certain nombre de principes et de régularités sous-jacentes qu'une norme de codage se doit de saisir. La première difficulté à laquelle a été confronté le groupe de travail sur les dictionnaires a donc été la définition d'un schéma de codage suffisamment général pour couvrir la plupart des dictionnaires, tout en permettant de décrire les particularités de chacun. Ce conflit entre généralité et pouvoir descriptif existe pour de nombreux types de textes, mais il semble atteindre son point culminant dans le cas des dictionnaires.

Un deuxième type de problème de codage provient du fait que les dictionnaires, contrairement à la plupart des autres types de textes, sont à la fois des textes et des bases de données[2]. Les dictionnaires ont bien évidemment l'apparence de textes et possèdent de nombreuses caractéristiques communes à tous les types de textes. Néanmoins, les utilisateurs ne lisent pas en principe les dictionnaires de manière linéaire de A à Z comme ils le font pour la plupart des textes, mais accèdent à des entrées à partir d'une clé (la vedette) dans le but de récupérer divers champs d'information associés à cette clé (prononciation, information grammaticale, étymologie, définitions, etc.). Cet accès non linéaire est typique de l'accès aux bases de données. Il est encore plus clair avec les dictionnaires électroniques qui offrent d'autres modes d'accès : l'utilisateur peut accéder à tous les mots dont la définition contient un mot donné, à tous les mots remplissant un certain nombre de critères (par exemple, tous les verbes relevant du domaine nautique, apparaissant avant 1900), etc. En outre, si l'affichage sur l'écran ressemble toujours plus ou moins à du texte, la représentation interne est rarement celle d'un texte linéaire.

Les dictionnaires présentent donc une forte dualité entre leur structure de surface (le texte) et leur structure profonde (le contenu informationnel). Une grande partie des informations de la structure profonde n'est pas explicite dans la structure de surface et nécessite la connaissance des conventions d'abréviation et de présentation des dictionnaires. Par exemple, dans l'entrée «roughcast» ci-dessus, la structure de surface – c'est-à-dire la position linéaire des divers éléments – ne dit pas explicitement que «nom» (n.) ne s'applique qu'aux trois premiers sens, etc.

La dualité structurelle des dictionnaires est source de difficultés de codage par le conflit qu'elle entraîne entre deux vues différentes du dictionnaire. Un utilisateur donné peut préférer le codage d'un point de vue textuel qui conserve la structure de surface (afin, par exemple, de rester fidèle à une version imprimée pré-existante). Cependant, le type d'inférence nécessaire à la récupération de la structure informationnelle profonde à partir de la structure de surface peut être difficile, voire impossible, pour un ordinateur. Si un utilisateur s'intéresse à la vue «base de données» (par exemple afin de visualiser et manipuler le dictionnaire à l'aide d'outils informatiques), il aura besoin d'un codage explicite des informations qui ne sont qu'implicites dans la structure de surface. Dans certains cas, les utilisateurs souhaiteraient même avoir accès aux deux vues simultanément. Étant donné que les deux vues du dictionnaire sont souvent en conflit, leurs codages peuvent être très différents. Un deuxième défi important pour le groupe de travail de la TEI sur les dictionnaires était de permettre le codage des deux vues, soit indépendamment, soit simultanément.

2. Il est à noter que, malgré le fait qu'une base de données puisse être générée à partir des informations de n'importe quel texte, un dictionnaire est une base de données par destination.

Le lecteur pourra trouver une discussion plus approfondie de ces difficultés dans [2].

5. Conclusion

Les propositions de la TEI ont été testées par le groupe de travail sur de nombreuses entrées de dictionnaires dans différentes langues. Plusieurs équipes dans le monde sont à l'heure actuelle en train de les appliquer à la création ou à la rétro-conversion des dictionnaires les plus variés et il est probable que cette utilisation en grandeur réelle aboutira à des propositions de révision et peut-être de simplification ou d'harmonisation. De même, l'extension aux dictionnaires anciens, ou aux gros dictionnaires comme l'*Oxford English Dictionary* ou le *Trésor de la Langue Française*, ne manquera pas de faire apparaître de nouveaux problèmes et difficultés. Les principes de base de la TEI semblent suffisamment robustes pour supporter une telle extension, mais il est concevable que de nouvelles balises ou de nouveaux attributs doivent être développés.

Dans de nombreux cas, il semble que les limites du langage SGML aient été atteintes : si puissant et utile qu'il soit, il a été conçu pour la représentation de documents simples, tels que manuels techniques ou correspondance commerciale ; la complexité de textes tels que les dictionnaires (ou les textes littéraires en général : manuscrits anciens, éditions critiques, etc.) semble indiquer la nécessité d'un langage de représentation de données de nouvelle génération, doté d'une plus grande flexibilité et d'une plus grande capacité expressive. Ne serait-il pas paradoxal que des préoccupations lexicographiques et littéraires contribuent à l'émergence de nouveaux langages informatiques?

Bibliographie

[1] Robert AMSLER, Frank W. TOMPA, « An SGML-Based Standard for English Monolingual Dictionaries », in *Information in Text: Fourth Annual Conference of the UW Center for the New Oxford English Dictionary*, University of Waterloo Center for the New Oxford English Dictionary, Waterloo, Ontario, 1988, 61–79.

[2] Nancy IDE, Jean VÉRONIS, « Encoding Dictionaries », in IDE and VÉRONIS (Eds.), *The Text Encoding Initiative : Background and Context*, Kluwer Academic Publishers, Dordrecht, 1995, 167–179.

[3] C.M. SPERBERG-MCQUEEN, Lou BURNARD, *Guidelines For Electronic Text Encoding and Interchange (TEI P3)*, ACH-ACL-ALLC Text Encoding Initiative, 1994.

[4] The DANLEX Group, *Descriptive tools for electronic processing of dictionary data*, Niemeyer, Tubingen , Lexicographica, Series Maior, 1987.