Integrated system for encyclopaedia typesetting based on TEX

Marko Grobelnik, Dunja Mladenić, Darko Zupanič, Borut Žnidar
Artificial Intelligence Laboratory, J. Stefan Institute, Ljubljana, Slovenia.
name.secondname@ijs.si

Abstract

The paper presents a system used for several already published dictionaries, lexicons and encyclopaedias. The system is based on a TEX macro package accompanied by many special purpose utilities for editor (person dealing with contents and form) support.

Introduction

Editing and typesetting of different kinds of encyclopaedic books (encyclopaedias, dictionaries, lexicons) is a highly specialized endeavour compared with the typesetting and editing of ordinary texts. First of all, these kinds of books are produced as relatively high budget projects in commercial companies. Many people are involved in such projects, which typically have well established working procedures already. Next, these kinds of texts are never really finished. Book releases are always only better prepared stages of contents which continue to develop (e.g. language in dictionaries). From a technical point of view it is important to mention the huge amount of text in such books and simultaneous appearance of several national languages along with all their peculiarities.

To build a successful system one should consider all the aforementioned properties of such book making. In the following few sections we will briefly describe the history of our involvement in organising and executing encyclopaedic book projects, actual solutions, the pros and cons of our work and some prospects for the future.

History

Our involvement in encyclopaedic book typesetting started rather accidentally few years ago. The Slovenian publishing house *Cankarjeva založba* was in the process of editing and publishing the book titled *The Encyclopaedia of the Slovenian language* dealing with all possible aspects (grammatical, historical, linguistic, ...) of the Slovenian language. The attempt with the classical way to publish the book in a printing house appeared to be very inappropriate and expensive, because of the large quantity of very technical text, with many major revision changes. The publishing house tried to use a standard interactive desktop publishing package to accomplish the job. The attempt failed again. After that, we decided to do it with TEX, which led to successful completition of the project. We made, of course, some mistakes, but this proved a useful experience for further work.

Because of the successful start, we were asked to build a general system for editing and typesetting the encyclopaedic type of books. Firstly, we adapted ourselves to the already established organisation of work in the publishing house, changing it slightly in the direction of the automatisation of all possible phases of work. The first project on which we developed our system, was the composition of several smaller dictionaries. After that we got the job to technically organise the work for the biggest Slovenian general lexicon *Sova* (in English *Owl*), where we finally developed the technology and the system. Currently, we are involved in several minor and major projects, the biggest being *The Encyclopaedia of Slovenia* (12 books + index).

Solutions

Because of the existing practice in the Slovenian publishing houses, the system was prepared for IBM-PC, although all components are portable. The use of the system is text-editor independent, however, we suggest the use of open and flexible integrated environments (e.g. TEX-Shell, Emacs, Borland-IDE, ...).

For the purpose of the common text input, a language called LEX was defined. The language is primarily entry-oriented with special elements like pictures, capitals, phonetic support, entry qualifying and many commands for semantic structuring of text. Characters used for LEX constructs are independent of the natural character set. The whole text corpus of a book is written in LEX format.

The following is an entry written in LEX format from *The English-Slovene Modern Dictionary*:

```
<entry:lx>
<head:action> <ipa:"aeKŠN>
  dejanje; delovanje, proces; tožba;
<p:to be killed in ~> pasti v boju;
<p:to put into ~> sprožiti, pognati;
<p:to take ~> ukrepati;
<p:out of ~> pokvarjen, izločen;
<p:social ~> družbena akcija
<end>
```

For the printout, the text in LEX is processed in three passes. First, the text in LEX is converted to the TEX format with a separate utility program which also performs several consistency controls. Next, TEX needs two passes. In the first pass correct picture positions and column lengths are determined. The second TEX pass builds final layout. The whole 3-pass process for 400 kbytes of text takes approximately 2 minutes on an IBM-PC i486/66 computer.

Besides typesetting, additional software was developed for editorial support which works on text in LEX format. This includes dictionary inversion, multi-author support (e.g. one published book had 40 authors), sorting support, etc.

Pros and Cons

Advantages of our system are:

- Working with the system and LEX format is extremely simple. For most of the books prepared with our system, only three people were involved: author(s), editor and typist.
- There is no need to change text editor habits. The only demand for the text editor used, is ability to export ASCII files.
- The cycle time between the corrections in the text and the printout of the finallayout is in the range of minutes.
- Additional editorial support is provided with text-manipulation utilities.
- The system is designed to support multilingual texts (dictionaries).
- The system is easily extended.
- The system runs on any platform with TEX. Minimal platform is IBM-PC i386 with DOS.
- The system was tested on several real encyclopaedic books, some of them very extensive and complex.

Disadvantages are the following:

- The system is not WYSIWYG (is this really a disadvantage?).
- When preparing the text for the final printout two things must be done manually: unresolved hyphens (narrow columns) and picture repositioning (to achieve an artistic look).

Conclusions and Future prospects

We have presented a system for editing and typesetting of encyclopaedic type of books. The system is based on TEX with additional utilities for editorial support. All components are hidden within an integrated environment. For the purpose of text input, we have defined a language called LEX, which allows us full control across the text corpus for checking and other text manipulation operations.

Our plans for the near future are to make a complete commercial product for dealing with encyclopaedic type of books along with all necessary interactive typesetting features.