Integrated system for encyclopaedia typesetting based on TEX

Marko Grobelnik, Dunja Mladenić,
Darko Zupanič, Borut Žnidar

Abstract

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

1 Introduction

Editing and typesetting of different kinds of encyclopaedic books (encyclopaedias, dictionaries, lexicons) is rather untypical job comparing to other kinds of texts. First of all, this kind of books are produced within relatively high budget projects within commercial companies. Many people are involved in such projects, typically having already quite established organisation of working procedures. Next, this kind of texts are never really finished. Book releases are always only better prepared stages of contents which is more or less dynamical (e.g. language in dictionaries). From technical point of view is important to mention huge amount text in such books and simultaneous appearance of several national languages along with all its peculiarities.

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

2 History

Our involvement in encyclopaedic books typesetting started rather accidentally few years ago. Slovenian publishing house *Cankarjeva založba* had a project to edit and to publish the book titled *The Encyclopaedia of the slovenian language* dealing with all possible aspects (grammatically, historical, linguistical, ...) 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 decided to try with standard interactive desktop publishing tools (VENTURA) to accomplish the job. The attempt failed again. After that, we decided to do it with TEX, which led to rather successful ending of the project. We made, of course, some mistakes, but we got many useful experiences for next books.

Because of the successful start, we got the job to build a system for editing and typesetting of the encyclopaedic type of books. Firstly, we adapted to the already established organisation of the work in the publishing house, changing it slightly in the direction of the automatisation of all possible phases of work. First projects, where we developed our system, were 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 project, the biggest being *The encyclopaedia of Slovenia* (12 books + index).

3 Solutions

Because of the local customs 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 to use 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 national character-set independent. The whole text corpus of a book is written in LEX formalism.

The following is an entry written in Lex format from *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 the three passes. First, the text in Lex is converted to the TeX format with a separate utility program which performs also 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 takes on IBM-PC i486/66 for 400kB text approximately 2 minutes.

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

4 Criticism

Advantages of our system are:

- The work with the system and Lex format is extremely simple. For the most of the books prepared with our system, only three people were involved: author(s), editor and typist.
- 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 text corrections and final-layout printout is in the range of minutes.
- Additional editorial support with text-manipulation utilities.

- The system is designed to support multilingual texts (dictionaries).
- Simple extendibility of the system.
- 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?).
- Two things must be done manually when preparing the text for the final printout: unresolved hyphens (narrow columns) and picture repositioning (to achieve artistic look).

5 Conclusions and Future prospects

We 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 the text input, we defined a language called LEX, which allows us having 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.