

DOCUMENT RESUME

ED 413 754

FL 024 775

AUTHOR Jakopin, Primoz
TITLE EVA - A Textual Data Processing Tool.
PUB DATE 1995-00-00
NOTE 4p.; In: Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995); see FL 024 759.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Alphabets; *Computational Linguistics; *Computer Software; Dictionaries; *Discourse Analysis; Foreign Countries; *Language Research; Linguistic Theory; Literacy; Slovenian; Uncommonly Taught Languages

ABSTRACT

EVA, a text processing tool designed to be self-contained and useful for a variety of languages, is described briefly, and its extensive coded character set is illustrated. Features, specifications, and database functions are noted. Its application in development of a Slovenian literary dictionary is also described. (MSE)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

At the end of 1982, EVA developed out of an experiment, to test if a humble personal computer of the day and place, Sinclair ZX Spectrum, could be put to reasonable use for text and data processing instead of a mainframe computer. The positive answer led to a full-blown text editor with integrated data base and graphics facilities in 1985, which in 1986 has been ported to the ATARI ST computer family (named STEVE). The DOS version has been in use since 1991; the Windows NT/Windows 95 version is under development. EVA has served as a software tool for processing of a sizeable amount of textual material and for preparation of various dictionaries in the Slovenian academic environment, where UNIX-based work stations have found very little response in the humanistic domain so far.

From the start, the program has been designed to be as flexible as possible, in order to allow the user to accommodate his own needs and situations with little external support. It is more or less self contained, with its own keyboard table, screen characters, DTP mode, graphics editor and OCR facility. To answer the needs for wide character sets such as UNICODE, a capability to process 8-bit and 16-bit characters in the same file has been introduced in 1993. If a line of text contains only characters with codes below 256, it is stored as a string of 8-bit characters in RAM as well as on disk. If, on the other hand, it contains one or more characters with codes above 255, it is stored as a 16-bit entity. In figure 1 the list of the first 1209 characters from the EVA character set is given.

Of special interest are codes 13 to 28, which modify the look of printed characters and sets 279 to 310, 327 to 354 and 365 to 369. The latter three groups represent the upper-, lower-, and through- diacritical characters which can all be combined with any other character to create characters, which are not part of the EVA character set.

A large set of data base functions includes general purpose routines such as sequential or indexed sorting and searching, as well as more specialised functions such as splitting of text into sentences, word or fieldwise translation and markup, or the computation of entropy. The majority of interfaces for import and export to other software is dual: ASCII file format (with user-definable filtering), RTF format (again with an open filter, for interface with Microsoft Word), WordStar file format, STEVE file type, open data record interface and the PCX picture format. For input there are also WordPerfect and TIFF interfaces.

Currently EVA is also used in production of a lemmatization dictionary of Slovenian, based on the Dictionary of the Slovenian Literary Language (93 151 headwords). So far, entries for nouns (54 522 lemmas to make 468 281 word forms) and adjectives (22 961 lemmas and 277 831 word forms) have been completed.



REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: TELRI - Proceedings of the First European Seminar: "Language Resources for Language Technology", Tihany, Hungary, Sept. 15 and 16, 1995	
Author(s): Heike Rettig (Ed.)	
Corporate Source:	Publication Date: 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

<input checked="" type="checkbox"/> Check here For Level 1 Release: Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) and paper copy.	The sample sticker shown below will be affixed to all Level 1 documents <div style="border: 1px solid black; padding: 5px; text-align: center;"> PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY _____ <i>Sample</i> _____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) </div> <p align="center">Level 1</p>	The sample sticker shown below will be affixed to all Level 2 documents <div style="border: 1px solid black; padding: 5px; text-align: center;"> PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY _____ <i>Sample</i> _____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) </div> <p align="center">Level 2</p>	<input type="checkbox"/> Check here For Level 2 Release: Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but <i>not</i> in paper copy.
--	--	---	--

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here → please

Signature: 	Printed Name/Position/Title: Norbert Volz, M.A. TELRI Project Manager	
Organization/Address: Institut für deutsche Sprache R 5, 6-13 - 68161 Mannheim Postfach 101621 - 68016 Mannheim	Telephone: +49 621 1581-437 E-Mail Address: volz(at)ids-mannheim.de	FAX: +49 621 1581-4158 Date: 28/11/97