

ELRA

Technical Centers

Call
 for creating a Network of Technical Centers for
 Written Language Resources Validation

Document Identification:	Validation centers Written/JO		
Title:	ELRA TECHNICAL CENTERS		
Release:	2		
Issued:	October 2001		
Origin:	ELRA		
Author/editor:	Jan Odijk		
Version:	<input type="checkbox"/> Internal draft	<input type="checkbox"/> Circulated draft	<input checked="" type="checkbox"/> Final
Status:	<input type="checkbox"/> Public	<input type="checkbox"/> Restricted	<input type="checkbox"/> Confidential
Revision dates:			
Print date:			
Number of pages:			
Distribution:	ELRA Board, ELRA VC		

TABLE OF CONTENT

1. PREAMBLE..... 3

2.WORK PACKAGES (WP) OF THE VC_WLR..... 3

 2.1 EXTENDING THE METHODOLOGY FOR DESCRIBING THE QUALITY AND CONTENT OF EXISTING WLR.....3

 2.2 IMPROVING THE QUALITY OF EXISTING WLR.....3

 2.3 QUALITY STANDARDS FOR WLR.....4

 2.4 VALIDATION OF NEW WLR.....4

 2.5 REPORTING4

3. ORGANIZATIONAL AND FINANCIAL ISSUES 4

 3.1 RELATION BETWEEN ELRA AND VC_WLR4

 3.2 RELATION BETWEEN ELDA AND THE VC_WLR5

4. FORMAT AND PROCEDURE FOR OFFER 5

1. Preamble

Describing, assuring and improving the quality of language resources are important tasks. The assurance of such quality is an important factor in ELRA's success. In the start up phase of ELRA it was foreseen that a Network of Technical Centers should be established to handle quality control.

To date a technical center for the validation of *spoken* language resources has been established. ELRA now intends to initiate the establishment of a network of technical centers for the validation of *written* language resources, the Validation Centers for Written Language Resources or VC_WLR. Written resources include lexicons as well as text corpora, possibly enriched with all kinds of annotations (POS-tags, syntactic structures, etc. etc.)

The procedure to establish the VC_WLR is identical to the one adopted in establishing the technical centers for spoken language resources, viz. they are to be established via an open call. Those European institutions willing to act as a VC_WLR for ELRA should send an offer to ELRA. The contents of this offer are described below. In particular, the offer must contain a proposal on how to address the problem of the detailed and thorough knowledge of a wide variety of languages required by the validation of multilingual resources.

ELRA's Board will decide which institutions will be selected. The selection of each candidate institution will be based on its ability to fulfill the tasks described in Section 2. The organizational and financial aspects are described in Section 3.

2. Work packages (WP) of the VC_WLR

2.1 Extending the Methodology for Describing the Quality and Content of Existing WLR

In the catalogue of ELRA many WLR are offered whose quality and content is not yet described in a satisfactory way. Some projects have resulted in linguistic resources distributed by ELRA that are comparable across languages in accordance with a commonly agreed content and format specification (e.g. PAROLE). However, almost no written data distributed by ELRA have been subject to validation by an external party and in accordance with a commonly agreed validation scheme (except for a limited number of PAROLE lexicons, and recently in the context of the ENABLER project). Though some research into the validation of linguistic resources has taken place and recommendations and guidelines have been formulated (e.g. Nancy Underwood et al., June 1998; Lou Burnard for text corpora), these have to be reviewed and where necessary adapted and extended to develop a concrete and workable methodology for the ELRA validation of written linguistic resources. The knowledge and expertise gained in the successful approach to validation taken in the SpeechDat family of spoken resources and by the existing ELRA validation center for spoken resources could be taken into consideration here, and its methods and approaches translated into an approach adapted for written language resources while maintaining the key elements that determined the success of the approach to speech.

The first task of the VC_WLR is to establish and/or extend the methodology for quality and content description so far developed.

The related document should focus on the quality and content of the WLR offered in the ELRA catalogue. A standard form should be developed for describing the content and quality of a WLR, starting from the form currently in use and taking into account the work carried out within TEI, OLAC, etc. The WLR in the ELRA catalog will have to be described according to this standard. This description will be used as a basis for providing any (potential) user with a quick overview in the ELRA catalogue relating to the quality and content of each WLR offered.

Output of WP2.1:

- Document describing methodology concerning quality and content
- Content and quality description of all ELRA WLR

2.2 Improving the Quality of Existing WLR

Existing WLR may have errors that could be removed with reasonable effort. The task of the VC_WLR is to establish a procedure to remove these errors. Especially a procedure has to be established which handles the errors reported by users of WLR (bug reporting procedure). Further, the existing WLR can be improved by better documentation, by reformatting according to established standards and by content changes. A similar procedure for spoken language resources has been proposed and is currently being implemented and experimented with,

hence it is sensible to investigate to what extent the procedure proposed for SLR can be adopted for the improvement of WLR and what modifications and or extensions are necessary or desirable.

The quality of the existing WLR should be gradually improved in accordance with a priority scheme that has to be worked out in close cooperation with ELRA's validation committee. The scheme has to be approved by the ELRA board.

Output of WP 2.2:

- Report describing the procedure to be used to improve existing WLR
- Improve existing WLR according to a priority scheme

2.3 Quality Standards for WLR

The VC_WLR have to play a leading role in establishing quality standards for WLR. For this task the VC_WLR have to cooperate with organizations involved in the production of WLR such as the consortia of the PAROLE and SIMPLE projects, and with ELRA's distribution agency (currently ELDA). Additionally, the extent to which existing recommendations, guidelines and proposed standards from groups such as the EAGLES and ISLE projects can be incorporated should be considered throughout.

Output of WP 2.3:

- Report describing the procedure for building up relationships with significant WLR producers and standards groups
- Following on from the report, the establishment of those relationships

2.4 Validation of New WLR

Owners of WLR regularly offer their WLR to ELRA for distribution. ELRA has the distribution carried out by its distribution agency (currently ELDA). Each time a WLR is offered for distribution, the task of the VC_WLR is to establish in cooperation with the owner of the WLR a manual containing

- The specification of the content of the WLR,
- The validation criteria for checking the quality of the WLR,
- The procedure to validate the WLR.

Based on this manual the VC_WLR have to validate any new WLR offered for distribution.

Output of WP 2.4:

- Report on the validation procedure as specified in a specific contract between ELDA and the center(s)

2.5 Reporting

Twice a year the VC_WLR must report work undertaken to date to the board of ELRA via the head of the validation committee.

Output of WP 2.5:

- Status reports

3. Organizational and Financial Issues

3.1 Relation between ELRA and VC_WLR

Concerning the tasks 2.1, 2.2, 2.3, 2.5 as described above the relation between ELRA and the institution(s) that are appointed as VC_WLR will be regulated by a contract between ELRA and those institutions. The contract has to be renewed after every fiscal year of ELRA by the Board of ELRA. Three months before the end of each fiscal year of ELRA the Board of ELRA will decide on the financial support to be given to the VC_WLR for the next fiscal year to perform the tasks 2.1, 2.2, 2.3, 2.5. Annually, a letter of intent will describe a budget for the year for the VC_WLR.

The initial amount made available will be approximately 15K EUR.

The ELRA validation committee will act as a steering committee for all activities related to validation of written resources. All actions proposed by the validation committee and agreed upon between the validation committee and the appointed VC_WLR will have to be approved by the ELRA Board.

3.2 Relation between ELDA and the VC_WLR

Separate contracts will be made with ELDA concerning task 2.4 on a case-by-case basis.

4. Format and Procedure for Offer

To apply to be a VC_WLR, send your offer by e-mail (as ASCII or RTF files, approx. 2000 words) to the CEO of ELRA (Khalid Choukri, choukri@elda.fr) and to the head of the ELRA validation committee (Harald Hoega, harald.hoega@mchp.siemens.de). The e-mail should contain:

1. Name of the proposing institute
2. The name of the person at the institute who will be the head of the VC_WLR.
3. A statement outlining the suitability of the institute to act as a VC_WLR.
4. A proposal on how the institute plans to provide for the required detailed and thorough knowledge of a wide variety of languages.
5. A list of personnel who will work on the tasks to be undertaken by the VC_WLR.
6. A possible start date
7. Sketch of the work for the work packages described that can be carried out within the fiscal year 2002 (1.1.02 – 31.12.02) for a budget of inferior or equal to 15KEUR. For each work package a rough estimate for the costs should be given.

Proposals are due by Friday March 1, 2002.