EUROPEAN LANGUAGE RESOURCES ASSOCIATION
ELRA

ELRA - Distribution Agency (ELDA)
Dr. Khalid CHOUKRI
CEO

55-57, rue Brillat Savarin
F-75013, PARIS, FRANCE

Tel. +33 1 43 13 33 33
Fax. +33 1 43 13 33 30
Email: choukri@elda.fr

ELRA VCom

Methodology for a Quick Quality Check for WLR-lexica V2.0

Deliverable Identification:	**ELRA/0209/VAL-1 Deliverable D1.2**
Title:	**Methodology for a Quick Quality Check for WLR - lexica**

Release:	2.0		
Issued:	October 2005		
Origin:			
Author:	Hanne Fersøe, Sussi Olsen (Center for Sprogteknologi, CST, Denmark)		
Version:	X Internal draft	Circulated draft	Final
Status:	Public	Restricted	Confidential
Revision dates:			
Print date:			
Number of pages:			
Distribution:	VCom members only		

Document evolution

Version	**Date**	**Status**	**Notes**
0.1	January 2004	First draft	First presented in VCom meeting in Brussels, February 2004. Comments received during the period following this meeting.
1.0	January 2005	Second draft	Completely reworked taking into consideration the comments received and the experience gained from conducting full validation on a number of lexical resources in the ELDA catalogue. This methodology will be applied to those resources on the priority list, which have already been fully validated.
1.1	July 2005	Third draft	Partly reworked based on comments from Jan Odijk.
2.0	Oct. 2005	Fourth draft	Minor editorial up-dates
			.

TABLE OF CONTENTS

1 Introduction

1.1 The Need for a Quality Stamp

Quality is an important issue whenever an individual or an organisation decides to purchase something, which has a certain cost. For this reason it is argued that the language resources in ELRA's catalogue should be validated according to a set of commonly accepted standards and be given a standardised and uniform declaration of content, a quality stamp, just like many other products do.

Most of the Written Language Resources (WLR) offered for sale in the catalogue have not been validated yet, and the same holds for many new resources offered for distribution through ELRA.

1.2 Background for developing a QQC procedure

As a consequence of the above, ELRA has taken the lead in developing a methodology and a validation report template (a standard) for full validation of lexicon resources (WLR-Lexica) [1], and currently 9 lexical resources in the ELRA catalogue have been fully validated according to this methodology, see Table 1.

Full validation of a lexicon is, however, rather time consuming, and consequently costly, because a number of the steps in the validation procedure can only be made manually by experts. It is assumed that a less comprehensive, very basic quality assessment of a lexicon may be sufficient in order for a potential buyer to obtain an overall impression of the quality of the resource, and therefore it seems sensible to build on the experience gained from the full validations and establish a procedure - a methodology and a report template - for performing a Quick Quality Check (QQC). Such a brief assessment may also serve as a basis for deciding on a full validation.

This report presents a draft of a WLR-lexicon QQC methodology, which will be tested on a subset of the resources already fully validated:

ELRA Catalogue number		Resource Name
	Date of Full Validation	
M0037	SCIPER English - Spanish Bilingual Dict.	Nov. 2004
L0049	SCIPER French Monolingual full form Dict	Sept. 2004
L0052	SCIPER Spanish Monolingual full form Dict	March 2004
L0006	ILC Italian Morphological Lexicon	Sept. 2004
L0012	gilcUB-M-Dictionary	Oct. 2004

Table 1: subset of fully validated WLR-lexica in ELRA's catalogue

2 Types of information about a resource

Apart from cost, two other categories of information about a resource are basic for a potential customer to make the decision to purchase a lexicon:

Factual information about the contents of the resource

- What does the resource contain? (language, coverage, size, linguistic information, possible applications)
- What is the format? (flat files, mark-up, database, tools)
- Who made it? (producer's contact details)

The degree of conformance with the specifications

- Reliability of the information (e.g. is it French and only French, does it have the said linguistic information, etc.?)
- Usability (is it complete, is the documentation adequate, are the files accessible, etc.?)

Examples of factual information, which is currently available at first sight in the catalogue, are presented below in section 2.1.

2.1 Examples of Factual Information

Currently, a customer's immediate source of factual information about a resource is ELDA's 'Catalogue of LRs', which can be consulted electronically at http://www.elda.fr/rubrique6.html. Below the presentation of 4 lexicons (WLR) in the catalogue are shown as examples:

Ref.	**Name**	**Type &**	**No of entries**	**Language**	**ELRA M**	**Non-M**	**Date**
L0001	DICO-MORPH_lemme. MEMODATA	Morpho-syntactic information	400,000 entries	French	R 12090 C 15112	R 15112 C 18890	23/01/97
L0002	DICO-MORPH_Collocation. MEMODATA	Collocation lexicon	35,000 entries	French	R 6992 C 8740	R 8740 C 10925	23/01/97
L0003	DICO-SYNT. MEMODATA		90,000 inflexional forms	French	R 8861 C 11077	R 11077 C 13846	23/01/97
L0004	Dutch Lexicon. (LanTmark)	General vocabulary	64,000 entries	Dutch	R 7680 C 61440	R 12800 C 102400	23/01/97

The link in the first column for the first resource (L0001) gives access to additional information:

L0001: DICO-MORPH_lemme. (MEMODATA)

Entries: more than 400 000
Language: French
Format: ASCII with separators
Medium: CD-ROM
French reusable lexicon for morphological works which produces the canonical form from the inflexional form. This lexicon is divided into the following lexical categories: nouns (55,000), verbs (8,000), adjectives (16,850), adverbs (2,000), other words (30,000).

This information is useful, but it is insufficient as a quality assessment, often it is not available, and it is not standardised for all the resources.

3 The QQC Methodology

The QQC methodology is based on two principles:

A. The QQC itself consists of checks performed against a few very basic requirements expressed as minimal criteria within three areas, in parallel with the procedure for full validation:
    - documentation checks, i.e. the suitability of the documentation and the content of the resource (factual information)
    - formal checks, i.e. the usability of the resource
    - correctness checks, i.e. the reliability of the content and usability information.

B. Generally a QQC should take about 5-6 hours of work (for one person at CST).

This procedure is somewhat different from the one adopted for SLR. In SLR correctness of e.g. transcriptions is not checked during QQC. This makes good sense, since this type of content checks are very time consuming and less relevant than e.g. checking the formal qualities of a speech database, since these are key factors that determine its usefulness.

For WLR-lexica the situation is different. Experience from full validations performed on a number of resources shows that the technical, formal issues such as media, number of files, file structure, file names etc. are less relevant QQC issues, since most WLR-lexica consist of one or two files only, they are usually not delivered in a database structure, and they usually fit on just one CD-ROM disc. The factual information stated in the documentation and matched against correctness checks of the content of the lexicon entries is a much more significant quality indicator (reliability), and from this it follows that the documentation of a lexicon is of extreme importance. Experience shows that documentation may consist of anything from no documentation at all, over 1.5 pages for a 500,000 entries dictionary, to hundreds of pages for a 20,000 entries dictionary. Clearly both the first and the last are prohibitive for performing the QQC. So the general suitability of the documentation must also be taken into consideration.

The results of the QQC are reported in a standard report employing a star notation in line with the notation developed for QQCs for Spoken Language resources:

Meaning of the quality stars:

1. * The minimal criteria for this aspect of the lexicon are not fulfilled or poorly fulfilled.
2. ** The minimal criteria for this aspect of the lexicon are reasonably well fulfilled.
3. *** The minimal criteria for this aspect of the lexicon are all fulfilled or well fulfilled.

The documentation checks, both suitability and content, are done manually whereas the formal checks include manual as well as automatic procedures. The correctness checks of the content have to be done manually.

The results of the validation checks described below in sections 3.1-3.3 will be summarised in Table 2, the Quality Assessment Table. The filled in table will appear in the cover page of the validation report together with a few concluding remarks.

QQC part	Quality value			General remarks
	*	**	***	
1. Documentation - suitability				
2. Documentation - content				
3. Formal aspects - usability				
4. Content - reliability				

Table 2: Quality Assessment Table for a QQC on WLRs in ELRA's catalogue

3.1 Documentation checks (suitability), manual

Documentation must be suitable, i.e. clear, sufficient and to the point.
Documentation must be in English, or alternatively, in the source language of the lexicon.

3.2 Documentation checks (content), manual

Documentation must clearly describe and specify the following topics:

- Copyright issues and contact persons
- Format and character set of the lexicon file(s), including naming conventions and how to open and handle the file(s), if relevant
- Lexicon size
- The languages of the lexicon and whether it is mono-, bi- or multilingual
- Type and structure of entries, including all legal attributes and values and their mutual dependencies
- Coverage of the lexicon, domain type
- Principles for coverage of each syntactic category (POS)
- Principles for coverage of the open and closed word classes
- Intended applications

3.3 Formal checks (usability), manual/automatic

The conformance of the following topics with the documentation is checked:

- Can the media and files be opened and handled as specified?
- Virus check
- Format and character set of lexicon files
- Lexicon size both the total number and the number of different types of entries
- The structure of entries
- Only legal values and attributes
- Are all legal values used?
- Are obligatory fields filled in?

3.4 Content checks (reliability), manual

The content check will be performed on a few (about 30) randomly selected words. For languages that CST is not able to handle in-house, it is not possible to perform content checks. Coverage for the different word classes is not included in a quick quality check

The correctness (reliability) of the following linguistic information will be checked depending on the type of lexicon:

- PoS tags (for all kinds of lexica)
- Morphological information, if relevant
- Syntactic information, if relevant
- Semantic information, if relevant
- Translational equivalents, if relevant

4. Layout of the QQC Report

The QQC report consists of a cover sheet with basic QQC data such as name of resource, ELRA catalogue number and date of QQC. On the same page the filled in Table 2 is inserted; it will be called *Quality Assessment Table*, and a concluding comment is added by CST.

CST will use the checks from sections 3.1 - 3.3 as a basis for the more detailed reporting. This detailed reporting that will both serve as input to the Quality Assessment Table and it will form part of the report.

The checks must be answered with very short comments such as ‘OK’, ‘missing’, or the like, or if necessary, with more comprehensive comments up to 1-2 lines of text. There is not enough time to make longer comments, and long comments are not the intention of a QQC.

For ease of reading, different fonts are used for the checks and for the answers.

The report template is attached after Section 5. References.

5. References

[1] Fersøe, H (2004). *Validation Manual for Lexica, V2.0*. Report submitted to ELRA under the validation unit contract ELRA/0209/VAL-1.

Documentation - suitability

Documentation must be suitable, i.e. clear, sufficient and to the point.

Documentation must be in English, alternatively in the source language of the lexicon.

Documentation - content

Copyright issues and contact persons

Format and character set of the lexicon file(s), including naming conventions and how to open and handle the file(s), if relevant

Lexicon size

The languages of the lexicon and whether it is mono-, bi- or multilingual

Type and structure of entries, including all legal attributes and values and their mutual dependencies

Coverage of the lexicon, domain type

Principles for coverage of each syntactic category (POS)

Principles for coverage of the open and closed word classes

Intended applications

Formal issues

Can the media and files be opened and handled as specified?

Virus check, specifying software

Format and character set of lexicon files

Lexicon size both the total number and the number of different types of entries

The structure of entries

Only legal values and attributes

Are all legal values used?

Are obligatory fields filled in?

Content

PoS tags (for all kinds of lexica)

Morphological information, if relevant

Syntactic information, if relevant

Semantic information, if relevant

Translational equivalents, if relevant