ELRA Validation Methodology and Standard Promotion for Linguistic Resources

Hanne Fersøe[1], Monica Monachini[2]

[1]Center for Sprogteknologi (CST) – Københavns Universitet
Njalsgade 80, Copenhagen, Denmark
hanne@cst.dk
[2] Istituto di Linguistica Computazionale (ILC) – Consiglio Nazionale delle Ricerche
Via Moruzzi 1, Pisa, Italy
monica.monachini@ilc.cnr.it

Abstract

This paper describes the results of work made for ELRA during 2003-2004. It describes the methodology for validation of written language resources (WLRs), specifically lexica, which has been developed for ELRA and tested on a few resources in the ELRA catalogue. It discusses the importance of key issues in lexicon creation and validation such as the adoption of standards for the coding of linguistic content and the importance of documentation. It reports on the experience gained from applying the methodology to lexical resources in the ELRA catalogue arguing that the checks must be reasonable, informative, on a suitable level of detail, and generic. It proposes a set of basic elements to be included in future discussions on establishing standards for lexicon resources. In conclusion it sketches the work to be undertaken in 2004 to promote validation and the adoption of standards.

1. Introduction

In 2000 ELRA's Board set up a Validation Committee with the aim to 'maximize the "ease of use" and "suitability" of the language resources (LR) which may be needed for LE-systems' (1). The term 'validation' is understood as the activity involved in quality evaluation of a database against one or more checklists of relevant criteria. The purpose of validation is to improve the quality of language resources and support their adherence to standards.

1.1. ELRA Validation Centres

The committee was set up to support one of the primary activities of ELRA according to its statutes: to give advice, coordinate, and carry out LR validation at European level. Following an open call two ELRA Validation Centres were established: Speech Processing EXpertise Center (SPEX) as the coordinator of the network for validation of spoken language resources (SLR), see (2), and Center for Sprogteknologi (CST) as the coordinator of the network for the validation of written language resources (WLR), see (3). The progress of the activities was reported in (Heuvel, H. van den et al, 2003).

1.2. Validation and Standards

Quality assessments of LRs based on a validation methodology, which builds on current standards and best practises, serves the goals of "ease of use" and "suitability", particularly when the quality criteria and best practices promoted through the methodology are also applied in resource production projects. In this context the relation between standards for validation and standards for the coding of content of resources is clear. A standard framework for the coding of linguistic content is a key issue in the construction of quality LRs and management of their internal accuracy. So well in line with the "ease of use" and "suitability" goals of the committee, activities aiming at providing definitions and descriptions of a basic set of standard linguistic elements for language resources are also promoted.

1.3. Structure of the Paper

This paper describes the results of the work done for ELRA concerning the development of the validation methodology and the promotion of standards and best practices specifically for lexicons.

Section 2 is the description of the methodology, and section 3 presents the lessons learned from applying it to resources in the ELRA catalogue. Section 4 describes the proposed standard for the creation of lexica and discusses its relation to previous work and current activities in the area of standardization and lexicon production. Section 5 presents the future work currently initiated by ELRA through the WLR Validation Center.

2. Presentation of the validation methodology

The Validation Manual for Lexica (Fersøe, 2004) presents an overall model for lexicon validation, which distinguishes between three major categories of validation criteria: those related to the documentation of the lexicon, those related to the formal properties of the lexicon, and those related to the lexicon content. The manual assumes a general understanding of the basic issues of validation and proposes an operational methodology including detailed checklists for each major category of validation criteria and a template for a validation report.

2.1. Documentation Validation

Validation of a lexicon's documentation is the act of checking that certain very basic information is present in the documentation. This involves a human reading the documentation and checking it against the criteria. It is also the first step in the full validation process, meaning that it is through the documentation that the validator first learns what the lexicon producers were intending and how they specified that. By lexicon documentation we mean the explanatory files that accompany the lexicon files themselves. These are files such as general and specific documentation, 'read me' files, operating instructions etc. The documentation is understood, generally, as the written presentation of the lexicon, and, specifically, as the design specifications containing criteria against which the lexicon will subsequently be validated. These criteria can be divided into three types of information, which must be present in the documentation in order for it to meet a good quality standard.

Firstly, the documentation should be written in English (also for lexical resources for other languages than English), and it should clearly present core administrative information: contact data for the resource (e.g. name, address, e-mail, URL), the number and types of physical media involved (e.g. CDs), the precise contents of each piece of physical medium, and copyright statement and IPR information, if relevant.

Secondly, the documentation should describe the formal properties of the lexicon. These are constituted by the basic technical information needed in order to access and use the data: character set(s) used, data format (e.g. mark-up language), system(s) needed to view and/or access the data, and the number, names and organisation of files belonging to the lexicon, plus the procedure for accessing them.

Thirdly, the documentation should contain the content information necessary to serve as a specification of the linguistic content. This covers the items lexicon size, lexicon coverage, intended application(s), natural language(s), data structure of an entry, entry types, attributes and their values, POS assignment and other relevant linguistic specifications.

2.2. Formal Validation

Validation of formal properties is the act of checking that the lexicon complies with the corresponding specifications of the documentation. Formal validation is the second step in the full validation process and it may involve both manual and semi- or fully automatic checks. Manual checks are e.g. the functional verification checks, where the validator checks the functionality described in the documentation by e.g. decompressing, installing, opening, running etc. the files according to the instructions, and the completeness checks, which verify the completeness of the package against the documentation (number of CDs, files, etc.). Semi- or fully automatic checks may be carried out to check the syntactic consistency of the lexicon by e.g. applying an appropriate XML-parser to the lexicon DTD and XML-files.

2.3. Content Validation

Validation of content is the act of checking that the content of the lexicon complies with the specifications. It is a manual process, which requires general linguistic and language specific expertise; it is the most complex of the validation tasks. Content validation is basically checking of coverage and of linguistic correctness. In both, some of the checks are general and could be applied to all lexica, while other checks must be designed for the specific lexicon and language(s) in question. Coverage refers to linguistic domain and text type covered by the lexicon and also to the completeness with which a domain or a text type is covered. The validator must create checklists, which copy the methodology of the lexicon creation on a scale which yields a size of a reasonable statistical significance; often the cost of the effort must also be considered. Correctness refers to linguistic coding. Here relevant samples must be selected at the general level to check the coding of open classes, closed classes and frequent words, while at the specific level samples must be selected which reflect the particularity of the resource.

The value of standards having been applied during lexicon development is thus central for the design of lexicon specific validation criteria. Accepted standards for resources regarding coverage and linguistic content may help the producer before production by offering stable criteria for designing and specifying the resource and for the internal validation during production. Also, the standard may serve as a checklist for the external validator's design of the specific parts of the validation.

3. Lessons learned from applying the methodology

The methodology and its associated checklists were first applied to three lexica with the purpose of testing the manual before it was finalized in its current version. One lexicon was then validated after a number of revisions caused by the first three validations were made. The validated lexica were selected from a prioritized list of lexica from ELRA's catalogue, two of them are semantically organized WordNets, and two of them are NLP-lexicons with feature based morpho-syntactic descriptions, they cover different languages, and they are all monolingual.

3.1. Validation Results

The validation reports show that the documentation of the resources is not particularly complete, well structured nor informative, and that for all the resources there are examples of essential information missing. The reports also show that, in spite of the missing information, it was possible to access and inspect all lexica, but because of the missing documentation it was not possible to check a number of formal aspects, e.g. completeness of the package. It was also possible to check many aspects of the content, but relevant aspects may have been overlooked because of missing documentation.

The questions that arise from applying the methodology concern: Appropriateness - do the proposed checks impose a reasonable quality requirement? Information level - will the result of the checking be informative? Level of detail - Is the level of checking too detailed or too superficial? Generic versus specific checking - Are the checks sufficiently generic to cover all types of lexica?

3.1.1. Documentation

The proposed criteria for documentation validation seem difficult to fulfil, and it might be concluded that the checks are too detailed and not reasonable. On the other hand, although the level of checking proposed is rather detailed, the information requested, if present, will prove to be very informative for a user of the resource. Based on our experience from compiling an edited collection of available validation methodologies in the context of the ENABLER project (Fersøe 2003), it is our impression that the low score on documentation quality may well be due to the fact that in many cases the documentation is created for internal use and not modified to present the resource to external users. The proposed validation methodology for documentation will be tested on more lexica, and subsequently be adjusted. For now the conclusion is that the checks proposed assume a reasonable, balanced and informative documentation standard at a suitable level of detail for a resource offered for distribution. We recommend that resource producers adhere to it, because it would add value to their resource.

3.1.2. Formal Properties

The proposed criteria for formal validation concern checking of conformance with specifications. If the documentation does not state e.g. how many and which files there should be, then completeness cannot be checked, and if it does not state the legal attributes and values, then consistency cannot be checked. It is reasonable and of informative value that such basic formal properties as completeness and consistency may be checked. On the other hand the reports seem to indicate that the level of detail in the checks regarding directories, files, platforms are not entirely appropriate, and these checks will need to be reworked, taking into account more validation results.

3.1.3. Linguistic content

The most demanding and difficult part of the validation is the content validation. Using the same check lists for all lexica will certainly not result in reasonable and informative quality information for all lexica. The current validation strategy of dividing the checks into general checks that should be applied to all lexica and resource specific checks to be designed by the validator seems to work. The difficulty consists in designing the resource specific validation in such a way that the checks most relevant for the resource in question are made and so that the validation can be made at a reasonable cost. Both of these presuppose the existence of a comprehensive documentation.

4. Standard promotion

The lack of broadly accepted standards or the presence of overlapping ones, prevents LRs from being shared and interchanged. In this light, ELRA has made the promotion of standards one of its main missions. Standards are intended to offer measures to surmount barriers: operational by guaranteeing technical and methodological support; conceptual by offering the best possible framework for the activities; infrastructural by solving the fragmentation of this kind of activities in Europe. ELRA seeks to provide a framework for lexicon development and use that takes into account the needs of a multilingual community and, especially, addresses the requirements of real end-users/industrial users. In this effort ELRA intends to capitalize on and reuse results from previous EC and national projects and standardizations activities.

4.1. The Outset

The driving criteria of Monachini et al (2003) in the design of an ELRA-standard for the creation and description of lexicon content at any level of linguistic representation are based on the outcomes of the previous ten-years standardization efforts that from EAGLES (Monachini and Calzolari, 1996; Sanfilippo et al, 1996 and 1999) through the PAROLE-SIMPLE experience (Lenci et al, 2000) arrives to ISLE (Calzolari et al, 2003), based on the fact that these recommendations have been worked out by a team of experts from both academia and industry, thus guaranteeing maximum coverage and compatibility, as well as a valuable degree of acceptance and/or diffusion in the community. In the light of ELRA's aim to aid the process of validation, the focus of the defined EAGLES-based standard is moved from lexicon developers to validators. Three main leading motives have been considered particularly fertile and suited to a standard for lexicon development and validation: the EAGLES methodology, aiming at a maximal decomposition and high granularity across different languages; the ISLE approach to the identification of the so-called *basic notions*; and, finally, the flexibility and modularity of the MILE model.

4.2. Basic notions

The basic notions are the lexical dimensions that play a role at any level of linguistic description and concur to define e.g. a morpho-syntactic unit, a syntactic structure, or a word meaning from a monolingual point of view. Moving to the multilingual level, the notions are the set of conditions and operations to be imposed on monolingual descriptions.

4.2.1. Decomposition

By virtue of maximal decomposition, the notions are resolved into their minimal constitutive sub-elements, thus allowing easier reusability or mappability into different theoretical approaches: small units can be assembled in different frameworks, according to different theory/application-dependent purposes. The pursuit of granularity does not restrict the standard to the very basic dimensions of a lexical entry: whenever consensus can be found on a more complex linguistic dimension, such shareable commonly agreed lexical objects are provided[1].

4.2.2. Modularity

Horizontally, modularity guarantees the independence of monolingual descriptions: the basic notions are distributed over independent but linked modules, corresponding to the different layers of linguistic encoding. Vertically, within the various layers, flexibility allows for different degrees of granularity in encoding, i.e. for both shallow and deep representations. Modularity and flexibility are crucial requisites of a standard.

4.2.3. Linguistic Representation

The specifications are presented along the lines of the different levels of linguistic representation.

[1] For example, at the semantic level, 'semantic relation' is an example of basic notion, whereas 'Qualia relation' is a more complex dimension associated to the Generative Lexicon.

Morpho-syntactic recommendations are arranged according to obligatoriness and appear in the form of attribute-value tables, followed by a glossary and application criteria. They constitute, throughout the community, a *de-facto* standard, being applied to a large number of European and non-European languages[2] and in different frameworks.

Syntactic specifications use the notion of syntactic frame, which corresponds to a set of possible syntactic structures, the head and its syntactic arguments, with their phrasal realizations, associated with an entry. Semantic specifications focus on the notion of semantic frame that specifies the predicative argument structure of a lexical unit. The syntactic and semantic frames are closely connected and rules are provided to map them to each other. The specification of these two layers derive directly from the PAROLE-SIMPLE framework, where they have been applied to lexicons of twelve European languages, becoming a sort of *de-iure* standard in the community[3].

Syntactic and semantic frames have strong discriminating power in sense disambiguation, thus constituting the hearth where the multilingual *constrain* and *add* operations are issued.

4.3. Relation to BLARK

Standards and validation are crucial in the development of the components of a BLARK (Basic LAnguage Resource Kit) for a language (Binnenpoorte et al 2002). A BLARK defines for a language, what is the “minimal set of LRs” needed for that language in order that it be possible for language and speech technologies to be produced for the benefit of the speakers of that language.

5. Future work

Future work of the validation centre will focus on providing many more validation reports and on revising and adjusting the validation methodology based on these to make it as generally applicable as possible. Extensive validation according to the manual is time-consuming and costly, so another task is to develop a method for a quick validation of a lexicon resource. The idea is that it should take less than a day for a person to make it, and that it should check a subset of the elements checked in a full validation, and that there will be no checks on content, i.e. the correctness of e.g. POS assignment.

In promotion of standards and best practices the work will concentrate on refining and tuning the current standard proposal. The instrument will be the mapping of linguistic specifications of ongoing lexicon production projects with the currently proposed standards in order to identify potential gaps and differences and to understand how such recent experiences can best be incorporated into the framework.

6. References

(1) http://www.elra.info/services/valcom.php

(2) http://www.spex.nl/validationcentre/

(3) http://www.cst.dk/validation/index.html

Binnenpoorte, D., De Vriend, F., Sturm, J., Daelemans, X., Strik, H. & Cucciarini, C. (2002). A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch. In Proceedings of LREC 2002, pages 1862-1866.

Calzolari, N., Bertagna, F., Lenci, A. & Monachini, M. (2003). Standards and Best Practice for Multilingual Computational Lexicons. MILE (the Multilingual ISLE Lexical Entries), ISLE Deliverable2.2-2.3, CLWG,Pisa.

Erjavec, T. & Monachini, M. (1997). Common Specifications and Notation for Lexicon Encoding of Eastern Languages. Deliverable 1.1. Multext-East Project, COP-106.

Fersøe, H. (2003). Edited Collection of Validation Methodologies. ENABLER-Deliverable-D4.3. See www.enabler-network.org/reports.htm

Fersøe, H. (2004). Validation Manual for Lexica. Report submitted to ELRA under the ELRA/0209/VAL-1 contract. See (1).

Heuvel, H. van den (2002). Validation of Content and Quality of SLR: Overview and Methodology. Report submitted to ELRA under the ELRA/9901/VAL-1 contract.

Heuvel, H. van den, Choukri, K., Höge, H., Maegaard, B., Odijk, J. & Mapelli, V. (2003). Quality Control of Language Resources at ELRA. Proceedings Eurospeech 2003, Geneva, pp. 1541-1544.

Lenci, A., Bel, N., Busa., F., Calzolari, N., Gola, E. & Monachini, M. et al. (2000). SIMPLE: A General Framework for the Development of Multilingual Lexicons. International Journal of Lexicography, 13(4), 249-263.

Mapelli, V. & K. Choukri (2003). Report on a (minimal) set of LRs to be made available for as many languages as possible, and map of the actual gaps. ENABLER-Deliverable-D5.1.

Monachini, M. & Calzolari, N. (1996). Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages. EAGLES Document EAG-LSG/IR-T4.6/CSG-T3.2, Pisa, Italy.

Monachini M. et al (2003) Towards a standard for the creation of lexica. Report submitted to ELRA under the ELRA/0209/VAL-1 contract.

Sanfilippo, A. et al. (1996) Subcategorization Standards. Report of the EAGLES Lexicon/Syntax Group. SHARP Laboratories of Europe, Oxford.

Sanfilippo, A. et al. (1999). Preliminary Recommendations on Lexical Semantics Encoding. Final Report, SHARP Laboratories of Europe, Oxford.

[2] Erjavec and Monachini (1997) extended the morpho-syntactic specifications to cover Eastern European languages and, in ISLE, they have been tested on the Asian languages.

[3] WordNet has become known, and its building block, the *synset*, should be taken into consideration.