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

Validation Manual for Lexica

Release 2.0

January 2004

Hanne Fersøe

Center for Sprogteknologi, Københavns Universitet
Njalsgade 80
2300 København S
Denmark

Acknowledgements:

This work builds on previous work for ELRA by Nancy Underwood and Costanza Navarretta from 1998 and on the ELRA manual for validation of existing SLR by SPEX, 2000.

TABLE OF CONTENTS

1 Introduction

This report constitutes deliverable D1.1A under the validation unit contract ELRA/0209/VAL-1. Its subject is the validation of Written Language Resources (WLR), specifically lexica, and it is a parallel report to deliverable D1.1B, which describes the validation of corpora. An updated version is planned for Dec. 2004.

The report builds on [1] with regard to specific validation criteria and processes, but where [1] presents general discussions of the issues of lexical validation, this report assumes a basic understanding of those issues and adopts a pragmatic approach to validation of lexica in the form of a manual. It also builds on [2] with regard to the general structure of the report. Appendix A contains the validation checklists in the form of templates for the expert validator to fill in. Filled in templates constitute a validation report. Appendices B and C present a general discussion of sampling and an example of an instruction for the validation of a specific resource.

The definitions, descriptions and report templates presented in this manual have been tested and adjusted based on the insights gained from testing. The test material was provided by ELDA from their catalogue of lexical resources.

Throughout the report 'validation' is understood as the quality evaluation of a database against one or more checklists of relevant criteria. For a specific criterion, the result of the validation may be an absolute value, yes or no, or it may be a relative value on a scale from 1 to 5 for certain types of content criteria.

Many different evaluation scenarios may be envisaged, but we have found that they can reasonably be summarized in the following 3 scenarios:

Scenario 1: The producer scenario
A lexicon producer develops the lexical data following a process involving specification, production (with internal validation), and external validation (see table below). In this scenario the lexicon producer defines the validation criteria that are considered relevant and necessary.
Scenario 2: The user scenario
A lexical resource already exists (it may or may not have been validated according to scenario 1), and a potential user wants to have it validated for a specific usage. In this scenario the potential user defines the validation criteria that are considered relevant and necessary.
Scenario 3: The context free scenario
A lexical resource is made available for distribution. In this scenario the distribution agency needs the resource to be validated against general, principled standardized criteria in order to be able to supply it with a declaration of contents.

This report intends to be useful for all 3 scenarios, but its checklists and methods should be seen as instruments for scenario 3, cf. below.

Lexica to be validated are understood as lexica which must be usable or useful within an NLP application, and this again means that they must meet certain minimal requirements in order to distinguish them from simple term or word lists. The outcome of a validation process is a report on how a particular lexicon fulfils all the different criteria defined for its validation. Appendix A to this document is a series of forms with checklists that may be used to record the results of the validation and as input for the validation report.

The overall lexicon validation methodology described here draws heavily on certain aspects of the framework for evaluating NLP systems, which was developed by the EAGLES Evaluation Working Group (EAGLES, 1996). The criteria and methodology are intended for the validation of existing WLRs (lexica), which may or may not have been developed according to a certain standard. Currently there is no such general standard, which can be directly applied in validating all types of lexica. However, a great amount of work aimed at standardising the features used in the production of lexica has been carried out by the EAGLES consortium [5], [6], and the ISLE consortium [7], and the reader is referred to [3] for an introduction.

A discussion of validation strategies is found in [2] and briefly presented below. Validation of language resources can be performed either during production, which means that the criteria to be employed (the standards or a relevant subset of them) are already taken into account, or it can be performed afterwards, as well. Furthermore, validation can be done in house (internal validation) or by another organisation (external validation). The two dimensions thus identified are shown in the following table (copied from [2]).

Validator	**Validation scheduling**	
	During production	After production
Internal	(1)	(2)
External	(3)	(4)

Table 1: Four types of validation strategies

The optimal strategy is to have all (1), (2), (3), (4) done, but for a limited validation approach the numbers in Table 1 above reflect the order of importance. The internal quality control during production is the most important quality safeguard. In contrast, to have only an external validation after the database is produced is the least preferable option.

Unfortunately, this last case may be typical for the validation of many of the lexica of the present ELRA catalogue, and so this deliverable addresses the formulation of the validation criteria for the lexica in the ELRA catalogue, but taking into account also forthcoming standards so that future lexica may also be validated using the criteria described here.

2 The Overall Model of Lexicon Validation

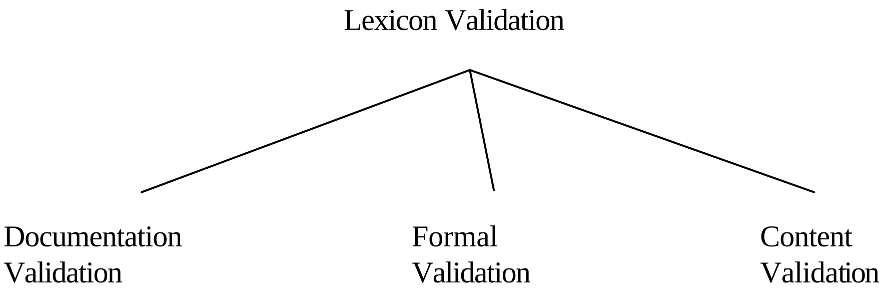
In validating a complex system such as a lexicon there are three major distinct categories of validation criteria, which must be taken into account:

1. *The lexicon's accompanying documentation* describing the lexicon with respect to its formal properties and its content. The documentation can bee seen as the written design specification containing the criteria against which the lexicon will subsequently be validated. Documentation is therefore extremely important for proper validation.

2. *The lexicon's formal properties* addressing on the one hand technical issues such as medium, delivery format, character set and on the other hand conformity with specifications addressing issues such as legality of attribute and value features. There are two possible types of specifications, which could be appropriate here: the suppliers' own specifications or some externally defined specifications or standard for a given type of lexicon.

3. *The lexicon's contents* addressing issues such as language(s), coverage and linguistic correctness of the coding.

The following diagram encapsulates the overall model of validation for lexica:



The rest of this manual is organised according to this overall model. Section 3 specifies the requirements for the contents of the accompanying lexicon documentation and describes the method for validation of the documentation. Section 4 describes the method for performing the formal validation, and section 5 deals with the validation of the contents of the lexicon.

3 Documentation Validation

In this section the requirements for the quality and contents of the accompanying documentation are described together with the method for validating documentation.

'Accompanying documentation' is understood as the descriptive, explanatory file(s) that accompany a lexicon, e.g. general documentation, specific documentation, 'read me' file(s), operating instructions, etc. The documentation accompanying a lexicon should clearly describe the standards to which the lexicon was created because in doing so it also defines the criteria by which it should be validated.

There is as yet no final, commonly accepted, set of standards that a lexicon resource must adhere to. However, the criteria detailed in this document were selected and described with a view to ongoing work in the area of standardisation [3], and they should be considered as basic, specifically the section on content validation.

The first source of information about a lexicon is given by the provider when he fills out the WRL-Lexica description form provided by ELRA. Other sources of information are the supplier's own specifications and documentation developed together with the lexicon, plus, of course, any extra information which may be considered useful.

3.1 Basic information

Each lexicon resource must, ideally, come with documentation written in English. For lexical resources for other languages than English, documentation written in that language is recommended in addition to the documentation in English.

There must be a 'read me' file in the root directory of each medium describing all files (including the documentation files) contained in the database. The 'read me' file must specify the text editor or viewer that is necessary to access and read the documentation files.

The documentation in electronic form must preferably be in flat ASCII or in a generally accepted de facto standard format.

The documentation should contain suitable administrative, technical and content oriented information.

See Appendix A for validation steps and reporting form template.

3.2 Administrative information

Various kinds of administrative information is necessary in order to enable ELRA and eventually users of the resource to behave correctly and appropriately with regard to all aspects of the resource, such as requesting further information, checking the package for completeness, copyright issues and other rights.

This section of the documentation should therefore include information about

- contact person (name, address, affiliation, position or department, address, telephone, fax, e-mail)
- number and type of physical media (CDs , DAT tapes)
- the contents of each piece of physical medium
- copyright statement, and information on IPR

See Appendix A for validation steps and reporting form template.

3.3 Technical information

3.3.1 Directories and files

The documentation should specify

- the directory structure
- the files corresponding to the lexicon
- any other files forming part of the electronic material
- the procedure for unpacking, installing, viewing and accessing the lexicon

See Appendix A for validation steps and reporting form template.

3.3.2 Format and character set(s)

If the lexicon is in some SGML or XML format, this should be specified, and the parser which was used to check its consistency during production, should be specified, too, and the DTD as well. For other types of text files, the basic structure of the lexicon and its syntax should be given.

See Appendix A for validation steps and reporting form template.

3.3.3 Database system and/or platform

Regardless of the delivery format of the lexicon the information about the database system, which the lexicon was stored in, converted from, run on etc. (e.g. ORACLE version xx) and about the technical platform (LINUX, UNIX) may be useful, and if it is available it should be specified.

See Appendix A for validation steps and reporting form template.

3.3.4 Data structure of an entry

The documentation should specify

- the data structure for an entry with
    - the fields corresponding to the standard format of entries
    - the order in which the fields appear

    - whether a field must be obligatorily filled out.

See Appendix A for validation steps and reporting form template.

3.3.5 Lexicon size

The documentation should specify the size of the lexicon with respect to number of entries and space requirements.

See Appendix A for validation steps and reporting form template.

3.4 Content information

3.4.1 The natural language(s) of the lexicon

The documentation should specify the natural language(s) of the lexicon and whether the lexicon is monolingual, bilingual or multilingual.

See Appendix A for validation steps and reporting form template.

3.4.2 Entry Type

Lexica built for different purposes will comprise different types of entries (e.g. only verbs as in 'Verbmobil', or wordnets based on a concept hierarchy as in EuroWordNet, or full form entries etc.). The documentation should specify the types of entry in the lexicon, with a brief description of the sort of information included: e.g. "purely morphological information", "syntactic and semantic information plus textual definitions", "subcategorisation information only" etc.

Bi- and multilingual lexica may consist of mappings between translational equivalents of the different forms.

See Appendix A for validation steps and reporting form template.

3.4.3 Attributes and their values

The documentation should specify the legal attributes and their values. The list should be structured according to the type of information it documents, and it should preferably give examples as well.

I. Morphological, morphosyntactic and subcategorisation features expressed as attributes with their legal values.

II. Other linguistic features. This could include semantic information in an attribute/value format, as well as more discursive information aimed at informing the human user, for example:

- meaning definitions, collocations, synonym/antonym links, conceptual links in a semantic hierarchy, usage information, citations, literature references, explanatory notes, etc.

III. Administrative information.
Various types of administrative or "housekeeping" information might be included in a lexicon, for example:

- the name of the editor of an entry, the time an entry was created, the time an entry was last updated etc.

In addition to the above types of specification of the features, the dependencies between the different features should also be made explicit. If the format is SGML or XML this information will be encoded in the DTD, see section 3.3.2 above.

See Appendix A for validation steps and reporting form template.

3.4.4 Coverage of the lexicon

The documentation should specify two aspects of coverage: the linguistic domain or text type covered and the granularity or completeness of the coding within that domain.

Linguistic domain/text type

The documentation should give an indication of the domain the lexicon covers, e.g.

- general language
- a particular technical sublanguage (e.g. meteorology, computer science, linguistics, etc.)
- a corpus of a particular text type. Such a corpus might either correspond to a particular technical domain or be intended to reflect a particular style in general language, e.g. newspaper texts, novels, scientific journals, etc. or a particular level of expertise

and the degree of coverage within that domain, e.g.

- by number, by frequency or by percentage.

Granularity/completeness

The question of granularity concerns the depth of coding aimed at and the number (if any) of the reading distinctions for individual lexical items. By depth of coding, we mean the range of different types of information, which entries may carry (see 3.4.3), but also a more general indication of the depth of coding, e.g. expressed through the application(s) the lexicon was made for (3.4.5). The documentation should indicate on what basis reading distinctions are made and how fine-grained these distinctions are. For example, for a given (type of) lexical item, reading distinctions may be made on the basis of meaning differences, or syntactic behaviour. Thus the documentation should contain information to make the user able to find answers to questions of the following kind:

- What depth of coding has been aimed at?
- How are reading distinctions determined? :e.g.
  a) all readings found in the corpus
  b) all readings having a particular frequency
  c) readings selected
     randomly
     intuitively
     limited to a certain number etc.
- How fine-grained are they?

For each word class (syntactic category or POS = part of speech), the documentation should specify the degree of coverage in the lexicon:

- it should specify which word classes are considered closed classes, and whether all members of the relevant classes have been coded
- for open classes the documentation of coverage (completeness) is more complex, since this must be assessed with respect to the domain or text type covered. So for example, it may be that all words in a particular corpus are covered, or that only the most common nouns in a particular domain are coded, etc.

See Appendix A for validation steps and reporting form template.

3.4.5 Intended application of the lexicon

The documentation should preferably specify

- if the lexicon was developed either as part of a larger application (e.g. a machine translation system, a grammar checker , etc.),
- or with the intention of it being suitable for a particular applic ation(s),
- and in that case which application(s),
- whether the lexicon forms part of an existing system or has been developed for a potential system,
- which syntactic theory or formalism (e.g. LFG, CG, HPSG, etc.) the lexical coding was based on, if any.

The intended purpose or application of a lexicon is important because this will have affected the different types of information encoded in that lexicon and the granularity as well. If no applications are indicated in the documentation it will be assumed that the lexicon is intended to be a general purpose NLP lexicon with a wide range of potential applications.

See Appendix A for validation steps and reporting form template.

3.4.6 POS assignment

Whilst there is a general consensus across languages and among lexicographers in the assignment of word classes (syntactic categories or POS) such as noun and verb, the assignment of other word classes is not so clear cut. So, for example, words, which are considered determiners in one language, may have equivalents in other languages, which function as adjectives or pronouns. For specific application purposes, the lexicon developer may also have decided to 'collapse' certain word classes into one POS for practical reasons.

In addition, there are a number of word types, which do not readily fit into the generally accepted categories. Examples of these would be symbols (%, $), formulae (2/3=B), acronyms (ELRA, NATO), abbreviations (kHz, plc), dates (12/7/89), uniques, see [5], (e.g. the English infinitive marker '*to*'), punctuation (,.;:).

Finally transcategorisation must also be documented. Transcategorisation is the case where a certain form of a category also functions as another category, e.g. participles which function as adjectives (e.g. "the smiling child").

Thus the documentation must specify and preferably give examples of

- how POS is assigned to problematic classes such as possessives, demonstratives, quantifiers, numerals, articles, particles,
- whether more than one POS assignment is allowed,
- whether special word types are included in the lexicon, and which POS they are assigned
- whether the lexicon contains foreign words, and how they are treated,
- how transcategorisation phenomena are treated.

See Appendix A for validation steps and reporting form template.

4 Formal Validation

Formal validation is the checking of factual characteristics of the lexicon against the claims made in the documentation, and in that sense it should be seen as validation of the lexicon’s conformance with specifications. Formal validation is by nature language independent and amenable to (semi-)automatic checks.

4.1 Conformance with specifications, manual checks

4.1.1 Directories and files, functional verification and completeness check

Directories and files are delivered in a physical package, which should contain the complete set of electronic data and corresponding tools (description form, documentation files, lexicon files, DTD, viewer, etc.) on a physical medium or media. The validation consists in checking the completeness of the package against the documentation (e.g. number and types of CDs, directories, files, etc.) and in verifying the functionality described in the documentation by e.g. installing the CDs. The checklist for this step of the validation should basically be generated by the validator on the basis of the details specified in the documentation, but it should as a minimum have the following types of checks:

- does the package consist of the number and types of media specified
- can all media be decompressed, accessed, opened, installed, run, executed, printed, etc. as specified
- are all the files listed in the documentation present in the package
- do they conform to the directory structure specified in the documentation
- is the specified file naming convention, if any, adhered to
- do the files conform to the format (e.g. XML) and character sets (e.g. UNICODE) specified in the documentation
- are there any undocumented files present in the package
- are all the files readable

See Appendix A for validation steps and reporting form template.

4.1.2 Database system and/or platform

If the documentation specifies the database system the lexicon was stored in, converted from, run on etc. (e.g. ORACLE version xx) and the technical platform (LINUX, UNIX), and if these systems and platforms are available for validation, then it should be checked whether

- the lexicon can be converted to the specified data base format
- the lexicon can be uploaded to the specified data base system
- the lexicon can be run on the specified platform

Often it may not be possible to perform this check, but it may be of interest to potential users of the lexicon to have this information validated.

See Appendix A for validation steps and reporting form template.

4.2 Conformance with specifications, (semi-)automatic checks

4.2.1 Syntactic consistency of the lexicon

The input to all validation will be text files. The preferred format for lexica for validation and distribution by ELRA is an SGML or XML format, but other text files, which can be manipulated and read by humans on a computer, are also acceptable. The specific delivery format will have been specified in the documentation (see section 3.3.2) including the tools used for the lexicon development and for the checking of syntax during development.

For lexica in SGML/XML format the legal values and their legal attributes will be documented in the specification and they will be formally defined in the DTD, which thus contains the content model of the lexicon. The DTD, the lexicon files in e.g. XML-format, and the XML-parser used during lexicon development to check the consistency will be used again in validation. Thus this check can be performed automatically.

If the text files are not encoded in the SGML/XML format, the lexicon would nevertheless be structured in some way, for example by the use of parentheses or bracketing, or the lexicon might be very simple with little or no structure. In any case, the documentation would still have to contain a description of the syntax and the abstract data model of the lexicon, and it would still be expected that some automatic syntax checking had taken place during production, and that these automatic methods were applicable again in validation.

The object of this part of the formal validation is thus to check that all and only the declared legal attributes and values have been used in the lexicon, and the checks would be:

- the verification of the DTD (may have to be done manually)
- the syntactic consistency of files (verification of the coded lexicon entries)
    - are only legal attributes used
    - are only legal values used
    - are all obligatory fields filled

See Appendix A for validation steps and reporting form template.

4.2.2 Lexicon size

The producer of the lexicon will have made a number of decisions before and during the production process about the total number of entries and the number of entries belonging to each category (POS) or class or type of word included in the lexicon. These decisions are of great importance and interest to future users of the lexicon, and they will therefore have been described in the documentation (see section 3.4.4.). The validation consists in checking the lexicon against the documentation to verify whether the lexicon contains the specified number of entries in total, per POS etc. The checklist for validation of lexicon size must be generated on the basis of the information supplied in the documentation, but the size aspects that as a minimum would need checking would be:

- Number of entries, total
- Number of entries, per major relevant category (POS)
- Number of different types of entries (4.3.2)

Other relevant size aspects could be those related to space requirements, this may for instance be of interest to future users, who need to integrate the data into their own applications. Again, the producer will have documented this, and the validation consists in checking the conformance of the lexicon with the specifications.

See Appendix A for validation steps and reporting form template.

5 Content Validation

As opposed to formal validation, content validation is language dependent and thus language specific. The purpose of the validation is to check the *coverage* of the lexicon and the *correctness* of the linguistic coding, and this is a process, which requires knowledge of the language of the lexicon in order for the expert validator to be able to make the most adequate decisions in the creation of checklists and samples and in determining problem areas etc. It is therefore necessary that the validator be a native speaker or someone with a deep knowledge of the language of the lexicon.

Clearly it is not feasible to check all the entries and their features in a lexicon above a certain size. It is therefore necessary to select samples. The samples must reflect the two basic levels of checking, coverage and correctness. In addition it may be necessary to extract samples for language, application and lexicon specific checking, e.g. samples which are selected to check phenomena associated with lexica of specific languages and/or particular applications and/or specific kinds.

A general manual such as this cannot provide a definitive methodology for validating the content of all possible lexica and the features assigned within them, but the checks described here are sufficiently comprehensive on the one hand to serve as a validation template for many lexica, and on the other hand to serve as a pattern for developing other checks and checklists, when needed.

5.1 Validation of coverage

Coverage of a lexicon refers to the linguistic domain and the text type covered by the entries of the lexicon (3.4.4), and also to the completeness with which such a domain or text type is covered. Validation is made on the basis of checklists that are compared to the lexicon or it is performed on relevant samples taken from the lexicon.

When creating the checklists for validation of coverage, not only the statistical significance of the lists (see the discussion of sampling in Appendix B) but also the overall cost of the validation must be taken into account. Experience shows that cost considerations often prevail over the statistically ideal size considerations.

Two factors play a role in making checklists for validation of coverage:

- Methodology – which lists are relevant
- Size – how comprehensive are the lists (cost, statistical significance)

Lexica are usually created on the basis of a certain methodology regarding the selection of words from the language to be included in the lexicon. Creation of checklists for coverage validation should copy this method on a scale which yields the desired size, if possible.

To illustrate this in an operational way, three kinds of lexicon coverage are described below and with them the checklists and the checks relevant for their validation.

5.1.1 Lexica that cover general language

The selection of words for inclusion in a general language lexicon is usually based on existing dictionaries, frequency lists, etc. and very often on corpus based selection criteria as well. In a general language lexicon the closed classes (e.g. pronouns, determiners, articles and prepositions) and series (e.g. auxiliary verbs, modal verbs, days of the week, months of the year) are expected to have 100% coverage. The open classes (nouns, verbs, adjectives etc.) are expected to be represented with a frequency reflecting their relative frequency in the language (e.g. most nouns) and to have been selected in such a way that frequent words in the language are included. This means that the expert validator must create resources (checklists) against which to check the lexicon. Such resources (checklists) are:

- Lists of closed classes relevant for the language of the lexicon
- Lists of relevant series
- Lists of frequent words, per word class
- Frequency of word classes, e.g. the percentage of nouns in a language or the frequency of nouns compared to other word classes

and it must be checked that the lexicon covers the sampled lists and complies reasonably with the percentages and frequencies.

5.1.2 Lexica that cover a particular sublanguage

The selection of words for inclusion in a domain specific lexicon is based on existing dictionaries, frequency lists, etc. of the sublanguage of the domain, usually combined with corpus based selection criteria as well. Lexica of one or more particular sublanguages may or may not also cover general language to a certain extent. If general language is also covered, the lists of closed classes and series described for general language lexica should be used for checking the coverage of the general language part of the lexicon. For the sublanguage part of the lexicon, i.e. the open classes, it may not be possible to create such checklists because they are likely to require access to the corpora used as the basis for lexicon creation, and this will rarely be possible for the validator. If it is possible and can be managed within the allowed cost frame the expert validator may generate checklists on the basis of e.g. introspection, otherwise the sublanguage coverage cannot be checked. Checklists for validation are:

- Lists of closed classes relevant for the language of the lexicon
- Lists of relevant series
- Relevant sublanguage lists (particular words from the sublanguage, frequency lists of the sublanguage etc.)

and the checks to perform are that the lexicon covers the sampled lists.

5.1.3 Lexica that cover a particular corpus

The selection of words for inclusion in the lexicon is based solely on a corpus, and external checklists are therefore not likely to be useful. Proper creation of relevant checklists requires access to the corpus and this will rarely be possible for the validator to have. It may therefore not be possible to check the coverage of corpus specific lexica.

See Appendix A for validation steps and reporting form template.

5.2 Validation of linguistic correctness

Clearly it is not feasible to check all the entries and their features in a lexicon above a certain size. It is therefore necessary to select representative samples and check these. A general discussion of sampling is presented in Appendix B. Here the conclusions of this discussion are adopted, and an operational method of sampling and validating the linguistic correctness at general and/or specific levels is described. As in validation of coverage, the methodology used for creating the samples and the size of the samples are significant factors.

Clearly it is obvious, also, that the desired level of correctness may well be relative to the intended application of the lexicon. The validator should therefore not be restricted to simply answering yes or no to correctness checks, but should rather give a relative score for the linguistic correctness (5.2.2). The score system has a granularity from 1 (lowest score) to 5 (highest score) where the validator must define the criteria for applying the scores in the validation of the linguistic correctness of a specific lexicon.

Finally it should be noted that the diversity of lexica yields a very large range from general level correctness (5.2.1) to specific level correctness (5.2.2) and therefore it may not be relevant to validate both levels for all lexica. This decision must be made by the expert validator w.r.t. a specific lexicon.

5.2.1 Linguistic correctness – general level

In brief, the conclusions of the discussion in Appendix B concerning method of sampling and size of samples at the general level are simple and straightforward.

Open classes must be sampled separately: If samples are made on a purely statistical basis there is no guarantee that all relevant problem areas, entry types, word classes etc. will be covered. Since they have different built in potential error types, all open classes must be represented in the complete sample.
Relevant distinctions within each open class must be sampled: e.g. nouns where de-verbal nouns and compounds are relevant distinctions.
Frequent words should be sampled: Frequent words tend to be linguistically more complex than less frequent words. The expert validator should make sure that the sample for each open class contains frequent words. These may either be identified through existing lists of frequent words (e.g. from the validation of coverage) or, for some languages, since frequent words are also often shorter, by sampling a reasonable number of words of 7 letters and less.
Sample size for closed classes: All members of closed classes should be checked.
Sample size for open classes: A minimum of a 1,000 entries in total for the open classes should be sampled for checking.

Once samples for checking have been selected or created, the expert validator's task is to check the correct coding of features within the entries. For the general level, this means checking the values assigned for all the features in an entry.

In practice, though, the allowed cost frame for the validation may prevent the checking of the entire sample or even of all the assigned feature values. In such cases the expert validator must make smaller samples but adhere to the same criteria, and it may even be necessary to select and check only a subset of the features.

See Appendix A for validation steps and reporting form template.

5.2.2 Linguistic correctness – resource specific areas

In brief, the conclusions of the discussion in Appendix B concerning method of sampling and the size of samples at the specific levels are that methods and criteria for sampling and sample size as well must be viewed in relation to the specificity of the resource. In the general discussion of sampling in Appendix B, 5 examples of specific phenomena (areas) and the considerations associated with them are described. In developing criteria for selecting specific samples it is indispensable that validators be experts, both w.r.t. language expertise and w.r.t. the specificity of the lexicon. Appendix C contains, as an example, a specific instruction in how validation, specifically sampling, may be performed for a specific lexicon (a EuroWordNet), and of how a validation template may be filled in..

The validation result should be expressed as a relative score, just as for the validation of linguistic correctness at a general level (5.2 and 5.2.2).

See Appendix A for validation steps and reporting form template.

5.2.3 Scores for linguistic correctness

The score system has a granularity from 1 (lowest score) to 5 (highest score) where the validator must define the criteria for applying the scores in the validation of the linguistic correctness of a specific lexicon.

6. References

[1] Underwood, N.L. & C. Navarretta, (1998). *A Draft Manual for the Validation of Lexica*. Final Report submitted to ELRA under the validation task contract.

[2] van den Heuvel, H, & L. Boves & E. Sanders (2002). *Validation of Content and Quality of SLR: Overview and Methodology.* Report submitted to ELRA under the ELRA/9901/VAL-1 contract.

[3] Monachini, M. & F. Bertagna & N. Calzolari & N. Underwood & C. Navarretta (2003) *Towards a standard for the creation of lexica.* Report submitted to ELRA under the validation task contract.

[4] EAGLES (1996). *EAGLES Evaluation of Natural Language Processing Systems. Final Report*. ISBN 87-90708-00-8. Copenhagen: CST.

[5] EAGLES (1996a). *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.

[6] EAGLES (1996b). *Subcategorization Standards. Report of the EAGLES Lexicon/Syntax Group.* SHARP Laboratories of Europe, Oxford Science Park, Oxford, UK.

[7] See ISLE homepage: http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

7. Appendix A Validation Report Template

The appendix contains a series of templates for a validation report for a lexicon. The values to be filled in by the validator are defined below.

Y = Yes
N = No
NA = Not applicable

Comment field:

- A comment is mandatory when the option NA is selected.
- The comment field should be used when descriptions are required
- At the conclusion of each set of validation steps (each form), the expert validator should comment briefly on the general result of that particular set of validation steps.

Score (criteria to be defined by validator):
1: lowest score
2:
3:
4:
5: highest score

A.1 Identification of validated resource

Lexicon name or reference number:	
Supplier's name:	
Date received by ELDA:	
Technical validator's name and contact details:	
Expert validation site's name and contact details:	
Date on which validation completed:	

A.2 Validation of Documentation

A.2.1 Basic Information – Validation steps

A.2.1	Criterion	Y/N/NA	Comments
1.	Is there any accompanying documentation?		
2.	Is the documentation written in English?		
3.	If the language of the resource is not English, is there any documentation written in the source language?		
4.	Is the description form filled in?		
5.	Is there a read me file in the root directory of each medium?		
6.	Is the read me file ASCII format?		
7.	Does the read me file specify the editor(s) needed to be able to read the documentation? if yes which?		
8.	Does the documentation contain administrative information?		
9.	Does the documentation contain technical information?		
10.	Does the documentation contain information about the content of the		

	lexicon?		
11.	Validator's summary comment to validation of basic information:		

A.2.2 Administrative Information – Validation steps

A.2.2	Criterion	Y/N/NA	Comments
1.	Are all the contact person details documented?		
2.	Is the number and type of physical media documented?		
3.	Is the content of each piece of physical media documented?		
4.	Are copyright issues documented?		
5.	Does the documentation contain information about IPR issues?		
6.	Validator's summary comment to validation of administrative information:		

A.2.3 Technical Information Validation steps

A.2.3	Criterion	Y/N/NA	Comments
	Directories and Files		
1.	Does the documentation contain a specification of the directory structure(s)?		
2.	Does the documentation list the files of the lexicon?		
3.	Does the documentation list other files forming part of the electronic material?		
4.	Does the documentation specify the procedure for unpacking, installing, viewing and accessing the lexicon?		
	Format and character sets		
5.	Is the SGML or XML parser used for consistency checking during production specified?		
6.	Is the DTD for the SGML/XML coded material specified?		
7.	For non-SGML/XML formats, is the basic structure and its syntax specified?		
	Database system and/or platform		
8.	Is a database system specified?		
9.	Is the platform specified?		
	Data structure of an entry		
10.	Does the documentation specify the data structure of an entry?		
11.	If yes, does the documentation specify the different fields of an entry?		
12.	If yes, does the documentation specify the order in which the fields must appear?		
13.	If yes, does the documentation specify whether a field must be obligatorily filled out?		
	Lexicon size		
14.	Is the number of entries specified in the documentation?		

A.2.3	Criterion	Y/N/NA	Comments
15.	Does the documentation specify space requirement?		
	Summary		
16.	Validator's summary comment to validation of technical information:		

A.2.4 Content Information – Validation steps

A.2.4	Criterion	Y/N/NA	Comments
	The natural languages of a lexicon		
1.	Is the lexicon monolingual?		
	If yes, which natural language does it cover?		
2.	Is the lexicon bilingual?		
	If yes, which two natural languages does it cover?		
3.	Is the lexicon multilingual?		
	If yes, which natural languages does it cover?		
	Entry Type		
4.	Does the documentation specify the types of entry in the lexicon?		
5.	Does the documentation include a description of the sort of information included with an entry?		
6.	Does the documentation specify the languages of translational equivalents?		
	Attributes and their values		
7.	Does the documentation specify all the legal attributes and their values?		
8.	Does the documentation specify the dependencies between the different attributes?		
	Coverage of the lexicon		
9.	Does the documentation specify the domain (text type) of the lexicon?		
10.	Does the documentation specify the degree of coverage?		
11.	Is the basis for and the depth of reading distinctions documented?		
12.	Does the documentation specify the degree of coverage for each syntactic category (POS)		
13.	Does the documentation specify the principles for coverage for the open classes?		
14.	Does the documentation specify for all classes whether they are treated as open or closed classes?		
	Intended application of the lexicon		
15.	Does the documentation specify which application(s) the lexicon is suitable for?		
16.	Does the documentation specify whether the lexicon forms part of a particular system?		
17.	Does the documentation specify the syntactic theory/formalism the coding is based on?		
	POS assignment		

A.2.4	Criterion	Y/N/NA	Comments
18.	Does the documentation specify the principles for POS assignment?		
19.	Does the documentation specify whether the lexicon contains special word types, foreign words, etc.?		
	Summary		
20.	Validator's summary comment to validation of content information:		

A.3 Formal Validation

A.3.1 Conformance with specifications, manual checks – Validation steps

A.3.1	Criterion	Y/N/NA	Comments
	Directories and files, functional verification and completeness check		
1.	Does the package consist of the specified number and types of media?		
2.	Can all media be handled as specified?, including		
	Accessed?		
	Opened?		
	Installed?		
	Run?		
	Printed?		
	Other?		
3.	Is the set of files complete?		
4.	Are the files organised according to the directory structure specified in the documentation?		
5.	Do the files conform to the format and character sets specified in the documentation?		
6.	Are all the files readable?		
7.	Are there any undocumented files present in the package?		
	Data base system and/or platform		
8.	Can the lexicon be converted to the specified database format?		
9.	Can the lexicon be uploaded to the specified database?		
10.	Can the lexicon be run on the specified platform?		
	Summary		
11.	Validator's summary comment to formal validation, manual checks:		

A.3.2 Conformance with specifications, (semi-)automatic checks – Validation steps

A.3.2	Criterion	Y/N/NA	Comments
	Syntactic consistency of the lexicon		
1.	Is the DTD or data model of the lexicon complete as specified in the documentation?		

A.3.2	Criterion	Y/N/NA	Comments
2.	Does the lexicon contain only legal attributes?		
	Are all legal attributes used?		
	Does the lexicon contain only legal values?		
	Are all legal values used?		
	Are all obligatory fields filled?		
Lexicon size			
3.	Does the lexicon contain the total number of entries specified?		
4.	Does the lexicon contain the number of entries specified for each major relevant grammatical category?		
5.	Does the lexicon contain the number of different types of entries specified?		
6.	Do the file sizes correspond to the sizes stated in the documentation?		
Summary			
7.	Validator's summary comment to formal validation, (semi)-automatic checks:		

A.4 Content Validation

A.4.1 Validation of coverage – Validation steps

A.4.1	Criterion	Y/N/NA	Comments
Closed classes and/or series			
1.	List below the checklists used to validate the closed classes and/or series and describe them (relevance and size) in the comments field. <insert the list of relevant checklists here>		Give a brief description of each checklist
2.	Are all members of the checklists present for:		
	Determiners?		
	Pronouns?		
	Adpositions (prepositions and postpositions)?		
	Conjunctions?		
	Auxiliary verbs?		
	Modal verbs?		
	Uniques (e.g. infinitive markers)?		
	Other relevant closed classes/series? (please copy rows if necessary)		
Open classes			
3.	List below the checklists used to validate the open classes and describe them (relevance and size) in the comments field. <insert the list of relevant checklists here>		Give a brief description of each checklist
4.	Are all members of the checklist present for: (please copy rows if the lexicon covers multiple domains/sublanguages that should be checked separately)		
	Nouns		
	Verbs		

A.4.1	Criterion	Y/N/NA	Comments
	Adjectives		
	Adverbs		
	Other word classes (please copy rows if necessary)		
5.	Describe the checklist generated to check the relative coverage of word classes?		
6.	Does the relative coverage of word classes in the lexicon correspond to that of the checklist?		
7.	Describe particular checklists developed to check a particular sublanguage, list them below with the relevant question they check, and answer the question.		
	Summary		
8.	Validator's summary comment to coverage:		

A.4.2 Validation of linguistic correctness, general level – Validation steps

When the score is less than 5, please indicate the number of errors in the **Comments** column (absolute, relative or in percent).

A.4.2	Criterion	Score	Comments
	Closed classes and/or series		
1.	Describe each sample, including size and features, for validation of closed classes and/or series and list them below: <insert the list of samples here>		
2.	For each sample give a score for the correctness of the linguistic coding where 5 is the highest score: (copy this section for each sample)		
	Morphological information?		
	Syntactic information?		
	Semantic information?		
	Multilingual information?		
	Other levels of information (list them)?		
3.	Please give a brief summary of the correctness of closed classes		
	Open classes		
4.	Describe each sample, including size and features, for validation of open classes and list them below: <insert the list of samples here>.		
5.	For each sample give a score for the correctness of the linguistic coding where 5 is the highest score: (copy this section for each sample)		
	Morphological information?		
	Syntactic information?		
	Semantic information?		
	Multilingual information?		
	Other levels of information (list them)?		

A.4. 2	**Criterion**	**Score**	**Comments**
6.	Please give a brief summary of the correctness of closed classes		
Summary			
7.	Validator's summary comment to general linguistic correctness:		

A.4.3 Validation of linguistic correctness, resource specific areas – Validation steps

When the score is less than 5, please indicate the number of errors in the **Comments** column (absolute, relative or in percent).

A.4. 3	**Criterion**	**Score**	**Comments**
Resource specific areas selected for validation			
1.	List the areas selected for validation, and for each area describe the sample(s), including size and features, used to check each area <insert the list of areas and samples here>		
2.	For each sample give a score for the correctness of the linguistic coding where 5 is the highest score: (copy this section for each sample)		
	Area 1		
	Sample a		
	Sample b		
	Area 2		
	Sample c		
	Sample d		
Summary			
3.	Validator's summary comment to resource specific linguistic correctness		

8. Appendix B Sampling for validation - General Discussion

General samples

This section discusses how to choose general samples, and how large these samples should be. When deciding on the size and type of samples to take, not only the statistical significance of the sample but also the overall cost of the validation must be taken into consideration. Therefore the discussion takes into account both the experience and practice of a number of producers involved in validating lexica and also the overall cost of the validation.

Firstly for the closed class words, e.g. pronouns, determiners, articles and prepositions, we recommend checking all items in these classes. The reason for this is that although such classes are often considered to be rather simple, certain classes, e.g. pronouns, often have complex sets of morphosyntactic features associated with them. How these words are in fact divided up into different classes is often dependent upon the particular language (and the linguistic theory traditionally linked with it) of the lexicon being validated. In certain languages some items in the class of pronouns, for example, may function as determiners and a separate class of determiners may not exist. Thus suppliers of lexica are required to specify how these minor closed classes are categorised.

Sampling entries in a lexicon cannot be approached in the same way as sampling a product line from, for example, a sweet factory. In the latter case it is to be expected that the products from a particular line should be more or less the same. A lexicon, on the other hand, generally contains a wide range of different word types with necessarily very different properties. In validating lexica, then, it is necessary to ensure both that all word types are checked and also that the varying complexity of the different types of entries is taken into account, focusing checking on the more complex types of entry.

The approach generally taken by publishers of lexica is to take a rather large randomly selected proportion of say 15-20% of the total number of entries. However, there are two major drawbacks to this type of approach. Firstly, one cannot be certain that even such a large sample will representatively include all the different entry types. Secondly, as soon as a reasonably large lexicon is to be validated, the number of entries to be checked becomes unmanageable. For example, in a lexicon with 60,000 entries one would be committed to checking a sample of 9,000 to 12,000 entries. With a really large lexicon of, say, 120,000 entries, the sample size would be between 18,000 and 24,000 entries. Clearly it would be impractical to check such large samples in the context of validation. Therefore we recommend a more principled approach to selecting the general samples, which will then allow smaller samples to be taken, with more confidence in their validity.

If expert validators are able to take into account how the lexicon is built up and which types of words tend to carry more or less information in their entries this can increase their confidence of the validity of the sample taken. Thus the open word classes (nouns, verbs, adjectives, adverbs) should each be sampled separately. There are two reasons for this. Firstly different word classes have different types of features and so it would be useful for potential users of the lexicon to have separate information on these different classes. Secondly, since word classes may differ widely in size (e.g. in a general language lexicon it would be expected that there are many more nouns than adverbs) taking a sample from each word class avoids the risk that the samples will not be representative. However, experience shows that even within these classes, certain types of entry will carry much richer information and thus be more prone to potential coding errors than others.

Therefore, it is recommended that the different word classes are also divided up according to other criteria. For example in many cases the shortest words are also the most frequently used words, and the most frequently used words tend to have a larger number of different syntactic and semantic possibilities than the less frequent words. Compare the English verb *have* with the verb *re-nationalise*. As well as functioning as both a main verb and an auxiliary, *have* has several different meanings (e.g. ownership, family relationship, possession) and participates in a wide range of collocations, whereas *re-nationalise* has only one main meaning and one subcategorisation frame. Similarly with nouns, compounds tend to have fewer semantic possibilities than simple nouns. Compare the compound *bookkeeper* (which describes a person with a particular job) with the simple noun *book* (which can refer to either a concrete object, the contents of such an object, a list of bets and odds etc.). Thus one rule of thumb in trying to capture representative samples would be to concentrate more on the shorter (more frequent words). However, in the case of nouns, another distinction would be that between deverbal and non-deverbal nouns, where the former will have inherited subcategorisation frames from the verbs they are related to whilst the latter may not have, compare *education* with *house*. In these cases the longer word will have more structured information, including the subcategorisation patterns, associated with it.

So, the class of nouns could be divided up into deverbal nouns and non-deverbal nouns, with non-deverbal nouns being divided into those under 7 letters long, and those over 7 letters long. Adjectives could be similarly divided into such subclasses, whilst verbs could also be classed according to their length. Such divisions of the open word classes are the simplest and clearly, based on an inspection of the lexicon being validated and the expert validator's knowledge of the language, the validator could create even more fine-grained distinctions. We return to this point below.

Sample Size

As for the actual size of the samples taken, as already mentioned, we recommend that all elements of the closed classes be checked. For the open classes in lexica up to around 60,000 - 80,000 entries, we recommend a proportion of around 2% of the total open class entries. However the total sample size for open class words should not be less than 1,000 entries, so that in the case of a lexicon of only, say 7,000 entries, the proportion would in fact be higher than 2%. For larger lexica and certainly those with considerably more than 100,000 entries, the proportion can be reduced, for example, to 1%. This may seem somewhat counter-intuitive. However, we assume that the smaller lexica will contain a higher proportion of frequent words (with their associated greater richness of information) than a very large lexicon. In other words, very large lexica are assumed to contain a larger number of less common (and therefore less problematic in coding terms) words than a smaller lexicon. Thus the recommended size for general samples can be summarised in the following table:

word classes	proportion	minimum number in sample
closed word classes	100%	
open word classes:		
lexica with up to 80,000 entries	c. 2%	1,000 entries
lexica with over 80,000 entries	c. 2% - 1%	1,000 entries

In the above we have simply referred to the number of "entries" in a sample. However, different types of lexica may be structured in such a way that they contain different types or levels of entries, which are linked. For the present purposes of deciding on the size of a sample, we assume that the term "entry" refers to the most specific type of entry. For example if a lexicon had morphological entries which were also linked to syntactic entries, it is clear that linked to one morphological entry, one could expect a number of different syntactic entries. In such cases all syntactic entries related to a particular morphological entry should be included in the sample.

It was mentioned above that validators could, in fact, create more fine-grained distinctions within the open word classes, based on a more thorough understanding of the contents and structure of the lexicon being validated. In addition, on the basis of such an understanding they could also choose larger samples than have been recommended above. Such an approach of course, would increase the validity of the sample and so the quality of the validation. However it would also increase the time taken and therefore the cost of the overall validation. Nevertheless we feel that within this validation framework, there should be scope for flexibility where validators can offer such a validation package (for a higher price) so that the supplier or ELDA could choose in particular instances, to have a more thorough-going validation.

Language/application/lexicon specific samples

In developing criteria for selecting language (and application) specific samples, validators will draw firstly on their knowledge and experience of the language(s) and any particular lexicographic problems which arise with that language, but also on the information declared in the accompanying documentation. In deciding on specific elements to be checked, the validator should thus not only be looking for potential problems but also considering what elements are important and of interest to potential users in a lexicon for the language in question either for “all-purpose” NLP applications or more specific ones in cases where the supplier only claims a limited number of potential applications for the lexicon. So the first task for the expert validator is to compile a list of language specific phenomena to be checked. Since the number and type of different language or application specific entries or features to be checked is highly dependent on the particular lexicon being validated, we will not give specific figures for the size of these samples.

Given the highly specific nature of this task, in the following we just give a few indicative examples of the type of phenomenon, which a validator may wish to check. It is however foreseen that individual lists of language/lexicon specific phenomena could be provided as feedback to this manual and be incorporated in an appendix. It is also expected that, especially for the more closely related languages and lexicon types, particular phenomena to be checked will be pertinent to a number of different languages and lexica.

Example 1 Phrasal verbs in e.g. Danish and English

Many Danish verbs also form phrasal verbs with a number of different prepositions or particles. E.g. *stige* (to rise) participates in a number of phrasal verbs (here we just give 2 examples): *stige af* (to get off/ alight), *stige på* (to get on/board). Or a verb like *gå* (to walk/go) participates in many more phrasal verbs:

gå ud	go out; be omitted; die
gå ud på	be to the effect that
gå ud fra	assume

at the same time both *gå* and *gå ud* can also occur with many directional or locative PPs without resulting in a phrasal verb, e.g.

Jeg går i biografen.	I’m going to the cinema.
Jeg går ud på gaden.	I’m going out into the street.
Jeg går ud af værelset	I leave the room.

Whether a verb is functioning as a phrasal verb or a simple verb with a prepositional argument affects its syntactic analysis, its pronunciation and its translation. So, in a comprehensive Danish lexicon, designed for many different processing tasks, the treatment of phrasal verbs and the distinction between them and verbs with valency bound prepositions should be checked. Similar phenomena also exist in other languages (e.g. English).

Example 2 Prepositions and case assignment in e.g. German

In German, prepositions assign case to their arguments. Certain prepositions only assign one case, whilst others assign different cases depending on the context in which they occur. For example, a preposition like *über* (above) assigns either accusative or dative case depending on whether a direction or a location is indicated as its object, e.g.

> *Bitte häng den Mantel über den Tisch* (direction)
> Please hang the coat above the table. (accusative)
>
> *Der Mantel hängt über dem Tisch* (location)
> The coat is hanging above the table. (dative)

So in a German lexicon aimed at syntactic processing, the coding of prepositions with respect to their different possibilities for case assignment should be checked.

Similarly German verbs exhibit different case marking properties, for example, the verb *folgen* (to follow (somebody)) assigns dative case to its object, whilst the closely related verb *verfolgen* (to pursue (somebody)) assigns accusative case. So the coding of the case assigning properties of verbs should also be checked.

Example 3 Transcategorisation in e.g. English

In English, certain word forms can function as more than one category. So, for example, the majority of participles also function as adjectives as well as verbs. E.g.

> *The winning competitor suddenly lost speed.*
> *The Irish competitor is winning the race.*

However there are also lexicalised forms of adjectives which resemble participles, e.g.

> *The incoming mail is filed over there.*

but they have no corresponding verb (there is no such verb as *to income*, although the lexicalised adjective is clearly derived from the phrasal verb *come in).* Thus, in an English lexicon designed for syntactic processing, both the treatment of transcategorisation phenomena and how they are distinguished from other merely apparent cases of the phenomenon should be checked. Transcategorisation is in fact common in many languages.

Example 4 Bilingual Lexica

In the case of bilingual lexica the criteria for selecting a sample to be checked will also include translation related criteria. If the validator is aware of certain specific words or semantic/syntactic classes of words whose translation may be problematic, these would need to be checked.

In determining the phenomena to be checked, the validator must always bear in mind the type of lexicon being validated. Thus, whilst the above examples generally apply to very rich, general purpose lexica, other more specialised lexica may also be validated. Such lexica may, for example, deal with a highly specific technical sublanguage or collocations or subcategorisation frames etc., deliberately ignoring other types of information which one would find in a general lexicon.

The validator may approach the selection of entries to be checked in two ways. Either particular words which should have the property to be checked are known beforehand, and so the validator extracts those specific words as a sample, or the sample can be selected on the basis of the relevant attribute/value pairs to be checked. It is expected that, in practice, a combination of the two methods would be applied.

Example 5 WordNets

WordNets are a specific type of lexica, based on a concept hierarchy, and the approach to sampling and validation differs from that of more traditional dictionary-like lexica. Samples will typically be

(a) based on language internal relations implemented
(b) based on equivalence relations
(c) based on instances of top ontology concepts (coverage of hierarchical nodes)
(d) based on frequency of a lemma

and they must have a reasonable size related to the statistics provided for their coverage

(a) for categories having less than 5 instances (synsets): all should be checked
(b) for categories having 5 to 10 instances: 50% should be checked
(c) for categories having more than 10 instances 2% should be checked

9. Appendix C Specific instruction: Validation of a EuroWordNet

This instruction was developed by Senior Researcher Anna Braasch, Center for Sprogteknologi, as a direct result of her work with the validation of the German EuroWordNet, which was carried out as a test of the methodology of this manual. The cost allowed for the validation was very modest, and this of course determines the scope of the validation.

1. Required documentation

For the validation of a EWN, the following documentation is required:

- EuroWordNet General Documentation (a thorough introduction to the general concept and principles)
- EWN Viewer user documentation (description of the tool functionalities providing the presentation of EWN data in various forms)
- Description of the language specific data set.

1. Requirements to the content of the language specific documentation

- Which approach (=model) is followed in the development of the **L**WN?
  (L = German, English, Spanish…)
  Merge, i.e. based on L language resource(s) or Expand (based on WN1.5)
  (The approach is relevant for the content quality check of the LWN)
- Selection of local Base Concepts and LS vocabulary? Building of the local core wordnet? (cf. EWN-Doc., p.57) Additions to the core wordnet?
- Statistics (necessary pre-requisites for validation):
  - table overview of representation of language internal relation types by number of entry words, subdivided according to PoS
  - equivalence relation types from the LWG to the ILI list subdivided according to PoS
  - coverage of TopOntology concepts for each PoS
- A textual part explaining the figures (e.g. concerning non-exploited internal relations, implemented levels of description, etc.) is required

2. Content validation of the EWN data set –some relevant points

Which info must be present on the entry word (minimum requirement)?
The semantic wordnet has to contain the following main information
(A) From the monolingual point of view

- Synonyms of the entry word (synset members)
- Hyperonym/hyponym relations (chains)
- Other language internal relations implemented (as stated in the documentation of the data set)

(B) From the multilingual point of view

- Equivalence relations to the ILI (mappings)

A further interesting question from the conceptual point of view is the coverage of the Top Ontology concepts and the distribution through the concept nodes.

The validation is to be performed along two lines, using the criteria of consistency and correctness:

(a) monolingual content (internal relations stated)
(b) relations to ILI (eq-links) and TopOntology

The criterion of exhaustiveness cannot be applied, as it is not the goal of the resource. Thus, the selection of concepts and entry words is not a subject of this validation.

3. Selection and size of samples

Selection criteria (tentative list)

(e) based on language internal relations implemented (cf. above)
(f) based on equivalence relations (mappings from a synset to the ILI)
(g) based on instances of top ontology concepts (coverage of hierarchical nodes)
(h) based on frequency of a lemma

Method of selection

(a) select word class (PoS)
(b) prepare a list for each PoS of appropriate candidates of semantically complex (‘machen’), medium (‘Feuer’) and simple (‘Vater’) concepts
(c) select an appropriate subset according to selection criteria and size (fixed no., randomly selected from the set resulting from the concept-based search, etc.)

Content validation

(a) check members of the synset (record lacking members, if any)
(b) check eq-links to ILI
(c) check hyperonyms/hyponyms, meronyms/holonyms of the synset members
(d) near-synonymes, -antonymes, etc.
(e) check language and ILI glosses
(f) other features, if any (variant)

Steps

(a) check the topmost level (synset)
(b) expand downwards node by node as regards relations (hyponyms/hyperonyms and meronyms/holonyms; (near)-antonyms, etc.

Reasonable size (tentatively proposed for all open word classes)
Check the statistics provided in the Documentation for instances

(d) for categories having less than 5 instances (synsets): all should be checked
(e) for categories having 5 to 10 instances: 50% should be checked
(f) for categories having more than 10 instances 2% should be checked

	RESOURCE SPECIFIC VALIDATION			
	Please validate a minimum of 5 (if possible) areas that the documentation explicitly states that the resource covers or that is clearly relevant to the *resource??* language. Extract a suitable sample and verify the correctness of the entries.	*HER: indsæt kort beskrivelse af res. (nøgleord e.l.?)*		**German WordNet:** Semantic network of relations between word meanings organized around the notion of a synset. Synset is a set of words referring to the same concept, etc. (=set of synonyms)
	Area 1			**Synset list**
Q1	Short description of area	Comment		Check: Are all words listed true members of the synset?
Q2	Do the entries in the sample comply with the documentation, *and/*or with your knowledge of the source language?	1...5	5	50 entry words, randomly selected (35 n, 15 v), the synset of each meaning is checked
Q3	*Please comment on observations, if the grading is lower than 5*	Comment		There are detected some inconsistency problems wrt. sexus and lexicalized concepts (1)
	Area 2			**Language internal relations (A) hyperonymy -hyponymy**
Q4	Short description of area	Comment		Hierarchical relationship comprising several levels. Check: hyperonym - hyponym chains, all levels
Q5	Do the entries in the sample comply with the documentation *and/* or with your knowledge of the source language?	1...5	4	Occurrences of contra-intuitive and incorrectly stated relations
Q6	*Please comment on observations, if the grading is lower than 5*	Comment		Following problem types detected: Unbalanced set of hyponyms (1) Wrong relationship (2)
	Area 3			**Language internal relations (B) near – antonymy, caused_by, has_derived**
Q7	Short description of area	Comment		Check: are the relations properly stated?
Q8	Do the entries in the sample comply with the documentation *and/*or with your knowledge of the source language?	*1..5*	5	Checked: 50 entry words, randomly selected (40 n, 10 v)
Q9	*Please comment on observations, if the grading is lower than 5*	Comment		these relations are only in a few cases implemented (1), but all are correct
	Area 4			**Language internal relations (C) meronymy - holonymy**
Q10	Short description of area	Comment		Check: are the relations properly stated?
Q11	Do the entries in the sample comply with the documentation *and/* or with your knowledge of the source language?	*1..5*	4	Checked: 50 nouns selected with focus on supposed mero/holonymy relations

Q12	*Please comment on observations, if the grading is lower than 5*	Comment		A few incorrect relations detected (1)
	Area 5			**Representation of selected TO nodes**
Q13	Short description of area	Comment		Check of selected instances of a node wrt. description consistency
Q14	Do the entries in the sample comply with the documentation *and/* or with your knowledge of the source language?	*1..5*	5	10 TO concepts and several subordinates were randomly checked
Q15	*Please comment on observations, if the grading is lower than 5*	Comment		
Q16	**Area 6**			**Equivalence relations to ILI**
Q17	Short description of area	Comment		Check of eq-links between SL word/meaning and ILI
Q18	Do the entries in the sample comply with the documentation *and/* or with your knowledge of the source language?	*1..5*	5	The mappings are in accordance with description and proviso given in the EWN documentation
Q19	*Please comment on observations, if the grading is lower than 5*	Comment		For this check the EWN General Document, Ch. 2.3 is an inevitable prerequisite
Q20	Other problems detected			
				The documentation provides no explanation of the large number of not implemented language internal relations (cf. Table 7) and lacking definitions, etc.
	SUMMARY			
R1	Please summarize on the overall *impression wrt* correctness of the resource	Comment		At the higher levels the look-ups show a well worked-out resource The problems detected (and specified in the notes below) appear all at deeper, embedded levels of the records, probably caused by automatically established relations and links and from the ‘merge approach (cf. GW Documentation, Ch.2).

Scores: 1 to 5:
1: A large number of major errors detected
2: Several error types with a large number occurrences pro type
3: Several error types with a few occurrences pro type
4: A few minor errors detected
5: No errors detected