Issues in Designing a Flexible Validation Methodology for NLP Lexica

Nancy L. Underwood,
Center for Sprogteknologi
2300 Copenhagen, DENMARK
[nancy@est.ku.dk]

Abstract

This paper addresses a number of issues surrounding the validation of lexical resources. In particular it discusses the need for both flexibility and rigour in a general validation methodology and how this affects what it means to validate (as opposed to "evaluate") a resource. Such a methodology which takes account of these issues is described.

Introduction

This paper discusses several important issues which came to light during a project to develop a draft manual for validating a variety of electronic lexica prior to distribution (Underwood & Navaretta, 1997). At the outset, the plan was to define a set of criteria which the lexicon must fulfil and standards against which it should be evaluated in order to be judged as "good" (or "good enough") and eligible for distribution. However, considering the current situation with respect to the availability of lexical resources and the broad range of potential users, it became clear that a general methodology for validation must include a high degree of flexibility whilst maintaining enough rigour to have value. The next section of this paper describes the situation which demands such flexibility in the methodology and the effect this has had on our understanding of what validation is. The rest of the paper then, describes the methodology developed showing how these factors have informed all areas of the methodology.

The Need for Flexibility in Validation

The methodology and a draft validation manual described in this paper were developed under contract to ELRA for its distribution agency, ELDA (European Language resources Distribution Agency). As a distribution unit ELDA's remit is to identify, classify, collect, validate and distribute language resources, including lexica. Lexica to be distributed by ELDA must all be usable as potential components in an NLP application. However, apart from this requirement the range and diversity of acceptable lexica for distribution is rather wide. This must obviously be the case for an organisation which would like to distribute as many useful resources as possible. If we just consider for example, the lexica featured in the current ELRA catalogue available on the Web, we can see that the currently available resources already vary greatly along a number of different dimensions such as the languages treated, the size of a lexicon, its coverage and the type of information it provides. Currently 9 different languages are represented, either in monolingual or bi- and multi-lingual lexica, and the number of languages is set to rise. In terms of size they range from lexica with 100,000s of entries to specialised lexica with only 1-2,000 entries. The list includes large general language lexica as well as more specific ones covering particular subject domains (e.g. engineering, economics, sociology and data processing). As well as variation in subject domains, there is also variation in linguistic coverage so that whilst some claim to treat all parts of speech, other smaller lexica treat specific parts of speech or linguistic phenomena such as compounds, collocations or verb phrases. The types of entries they contain also varies (e.g. fully inflected forms vs. stems or affixes), as does the richness of information encoded. Some contain all of morphological, morphosyntactic, syntactic, phonological, semantic and usage information, whilst others contain only a subset of that information.

These are just some of the differences found in the lexica currently on offer, and it is clear that there are many other types of lexica out there. This wide diversity has a profound effect on the development of a validation methodology. Not only must the methodology be flexible enough to accommodate the diversity of the resources already available, but also take account of potential as yet unknown resources which will be offered to ELRA. As will be shown below, the uses to which a lexicon will be put are also somewhat ill-defined.

Validation vs. Evaluation?

Given the situation described above, it became clear that defining a general methodology for validation could require that the definition of validation itself may have to be refined for the current context. In determining what validation is, it was instructive to compare it with the related activity of evaluation in which it seems to be generally agreed (see e.g. Sparck Jones & Galliers (1996) Manzi et al (1996)) that the user and the context of use are central in evaluating an NLP resource. Thus a resource is evaluated with respect to how it can be successfully used to fulfil specific tasks within a specific setup. A vital part in setting up an evaluation then, is the discovery and formalisation of the user requirements for the resource being evaluated. However, in the current context, the distribution of lexical resources is not limited to a specific user-group, with well-defined requirements. Apart from the assumption that the lexica will be used in an NLP application, it is not possible to pre-determine the context or setup in which the lexicon will be used.

1 Copies of the draft manual are available from ELRA via elra@calvanet.calvacom.fr, comments and feedback are very welcome.
Indeed given the current situation with respect to the availability of lexical resources, it is not clear how stringent a user's requirements would be. For some languages, many different lexica exist whilst for others there are very few available, and one can imagine that a user may acquire a lexicon with more information than is required for a particular application and simply ignore or delete the superfluous information. Alternatively users may choose to further develop an existing lexicon in order to make it suitable for their own application, and may even accept a less-than-perfect lexicon rather than have no lexicon at all.

Thus the task of validation here cannot be to evaluate whether a particular resource meets the needs of a particular user(-group). Rather the goals of validation must be in some ways more limited, both confirming that the resource in question is what it claims to be and providing sufficiently detailed information about a lexicon to enable potential users to decide for themselves whether they wish to acquire it. This is not to say that the ultimate use to which a lexicon can be put is ignored. In fact this is taken very much into account during the design of an individual validation of a specific lexicon as will be seen below, but the general validation methodology cannot prescribe hard and fast user requirements which must be met.

Given the uncertainties surrounding both the properties of lexica to be validated and the requirements of potential users, the outcome of the validation process is a detailed report on a large number of characteristics to enable the user to make up their own minds. The validation methodology must be both rigorous enough to give confidence to potential users in choosing to acquire a lexicon and flexible enough to account for the wide range of lexica and potential user needs. In order to facilitate this, very detailed and stringent requirements are placed on the documentation accompanying the lexicon to be validated.

A Decompositional Model of Lexicon Validation

In order to achieve rigour in validating a lexicon. It must be decomposed into a number of characteristics to be checked and reported on (cf. quality characteristics in evaluation frameworks, e.g. TEMAA, 1996).

Each of these characteristics are further divided up into a number of sub-characteristics which will be checked and reported on in a comprehensive validation report.

It is envisaged that technical validation will first be carried out by, or under the supervision of, the distribution agency before sending it to specialised centres to validate its conformity with the relevant specifications and its linguistic content.

Minimal Requirements for a Lexicon

Although the validation methodology must be flexible enough to allow for the wide range of resources and user requirements mentioned above, this does not mean that all resources will be automatically validated. Some of the characteristics have been identified as minimal requirements which the lexicon must fulfil. These minimal requirements are checked at the beginning of the first (technical) phase of the validation.

One of the first basic questions to be asked is whether the resource in question can be considered a lexicon. As a minimal requirement a lexicon must provide more information than a simple word list. However exactly what this information must be is dependent on the type of lexicon being validated. For example, a lexicon for parsing should at least contain category and inflectional information. Inflectional information could either take the form of fully inflected words with all their associated features or a label indicating the inflectional paradigm to which the lexical items belong. On the other hand a more specialised resource such as a bilingual lexicon designed for machine translation purposes must minimally contain translation equivalents and parts of speech to enable disambiguation. Even more specialised lexica such as a valency lexicon for a specific part of speech would not have to contain category information, but instead information on the valency of each lexical item. Whether a putative lexicon provides more information than a simple word or term list should be clear from the documentation accompanying its delivery. If a resource does not meet this minimal requirement there is no point in validating it further.

This also means that all lexica must be delivered with detailed documentation whose contents are prescribed in the manual to enable a sufficiently detailed validation. Thus if the requisite documentation is not provided the validation cannot proceed. Similarly the lexicon must be complete and accessible in a format which can be used in the rest of the validation procedure (text files). A delivery which fails in any of the above criteria will be rejected for validation.

Technical Validation

Once it has been ascertained that the lexicon fulfils the minimal requirements, one of the most important aspects of technical validation involves checking the syntactic consistency of the lexicon files to ensure that the resource is in fact usable in an NLP system. The preferred format for delivery is SGML, whose consistency can be checked with respect to the DTD with an SGML parser. However, it must also be recognised that, at this time, not all lexica are produced in SGML, so the validator must also be able to parse other text formats. In either case the supplier of the lexicon is required to provide a description of the
syntax and the abstract data model of the lexicon, be it a DTD or not.
In addition a number of other technical characteristics concerning e.g. character sets, format and media are also checked and recorded.

**Conformance Validation**

In validating conformance with specifications, the aim is to ensure that a lexicon contains all and only the features specified for it. In an ideal world such specifications would be an external standard such as that developed by EAGLES for the production of lexica (EAGLES, 1996a, 1996b). Such standards are to be encouraged because they help to facilitate the sharing and re-use of much needed resources. However, unless or until they are in widespread use in lexicon production, it would be counter-productive to insist on their use in this context. Such insistence in the current situation would seriously restrict the range of lexica available for distribution and re-use. Nevertheless, any lexicon must have a set of specifications (whether conforming to an externally defined standard or not) and in validating a lexicon, it is important to check its conformance with those specifications. Such specifications are required to be delivered as part of the documentation accompanying the lexicon. The checking of conformance is broken down into three criteria which can be checked relatively simply using UNIX-based tools comparing the text files containing the lexicon and its specifications.

- All attributes and values in the specifications should occur at least once in the lexicon file(s).
- All attributes and values in the lexicon file(s) should occur in the specifications.
- Attributes in the lexicon file(s) should have legal values (according to the specifications).

In the spirit of allowing potential users to make up their own minds about whether to acquire a lexicon, if it fails in any of the above criteria the validator is require to record which legal features are not in the lexicon or which features in the lexicon are not in the specification along with any cases of illegal value assignments. It is assumed that minor (or even not so minor) discrepancies will not necessarily render the lexicon useless for a potential user.

**Content Validation**

Finally the linguistic content is checked by experts in the lexicography of the language(s) of the lexicon. Since different languages typically pose specific problems in the construction of lexica, this is arguably the most flexible part of the procedure.

Because of the different properties exhibited by different languages, it is not possible to provide a definitive methodology for validating the linguistic coding of all possible lexica, which defines all the features which should be checked. To take just one example: in validating a German lexicon for use within an NLP application such as parsing or grammar checking one would expect the entries for prepositions to also contain information on the particular case(s) which they assign, whereas such information would not be of interest in other languages like English and Danish where only accusative case can be assigned by a preposition. Instead, there are guidelines for the validator in designing the content validation of a particular lexicon based on the language(s) in question and the specific application for which the lexicon is intended (if any). This part of the validation can be divided into two main activities: selecting the samples for entries to be checked and checking the coding in the entries.

**Selecting Samples**

In addition to allowing for the diversity of languages and applications of the lexica, it was also felt to be necessary to try and take into account practical factors such as time and budget which might be available for a validation. Clearly for most lexica, it is not feasible to check all the entries and their features. Therefore guidelines have been proposed to try and ensure that a representative (but cost-effective) sample can be taken.

The sampling proposed is divided into two types which can be termed "general samples", which allow the validator to gain an overall impression of the coding in the lexicon, and "language or application specific samples" which are designed to take into account properties and potential problems associated with lexica of specific languages and which are designed for particular applications.

**General Samples**

It is recommended that all entries for the closed class words, (e.g. pronouns, determiners, articles and prepositions) are checked. Although such classes are often considered to be rather simple, some of them (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 (or linguistic theory) of the lexicon being validated. In certain languages and theories some items in the class of pronouns, for example, may function as determiners and a separate class of determiners may not exist. It is a requirement that the accompanying documentation indicates how these minor closed classes are categorised. Turning to the open class words, the aim must be to ensure that a sample is sufficiently representative. A general lexicon, can be expected to contain a wide range of different word types with necessarily very different properties and it is necessary to ensure that all word types are checked taking into account the varying complexity of the different types of entries.

From an informal survey of lexicon publishers, it transpired that the approach generally taken is to check all the closed class items and then take a rather large randomly selected proportion of say 15-20% of the total number of entries. However, there are two drawbacks to this type of approach in the current context. 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 in the sample becomes very large and potentially uneconomic. For a really large lexicon of say, 500,000 entries, (one of the largest lexica in the ELRA catalogue is over 500,000 entries) this would mean checking between 75,000 and 100,000 entries. Even with much smaller lexica of say 60,000 entries, the sample size would be between 9,000 and 12,000 entries. Therefore a less random sampling procedure has been
proposed in which the open word classes (nouns, verbs, adjectives, adverbs) should each be sampled separately. 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 reduces the risk that the samples will not be representative in terms of word classes. Also since different word classes have different types of features, it was considered potentially useful for a user to have separate information on the coding of the different classes. However, experience shows that even within one class, certain types of entry will carry much richer information and thus can be more prone to coding errors than others. So it is also recommended that the different word classes are also divided up according to a number of 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, membership, possession) and participates in a wide range of collocations, whereas re-nationalise has only one main meaning and one subcategorisation frame. So one simple rule of thumb is to concentrate on shorter words. This is clearly not the only criterion to be used and others potential criteria are briefly discussed in the draft manual. Such simple divisions of the open word classes are based on an inspection of the lexicon being validated and the validator’s knowledge of the language. Should a producer or the distributor request a more thorough-going validation, the validator could create ever more fine-grained distinctions. Once a number of such criteria are applied to the selection of samples to ensure their representativeness, the size of the sample can be reduce with some confidence. For the open classes in lexica up to around 60,000 - 80,000 entries, it is recommended that a proportion of around 2% of the total open class entries are sampled. 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, but it is assumed 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”) words than a smaller lexicon.

Language or Application Specific Samples In collecting language and application specific samples, validators use not only their knowledge and experience of the language(s) and any particular lexicographic problems which arise with that language, but also the information declared in the accompanying documentation. In deciding on specific elements to be checked, the validator should not only be looking for potential problems but also considering what elements are important and of interest to potential users, according to the type of lexicon being validated, and the sort of application it could be used in. 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, the draft manual does not give specific figures for the size of these samples and gives just a few indicative examples of the type of phenomenon which a validator may wish to check. Just to give one example, many Danish verbs also form phrasal verbs with a number of different prepositions or particles. E.g. the verb gå (to walk/go) participates in a number of phrasal verbs:  

\begin{verbatim}
gå ud  
(go out; be omitted; die)  
gå ud pâ  
(be to the effect that)  
gå ud fra  
(assume)  
\end{verbatim}

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.  

\begin{verbatim}
Jeg går i biografen.  
(I’m going to the cinema.)  
Jeg går ud pâ gaden.  
(I’m going out into the street.)  
Jeg går ud af værelset  
(I leave the room.)  
\end{verbatim}

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, suitable for many different natural language processing tasks, the treatment of phrasal verbs and the distinction between them and verbs with valency bound prepositions should probably be checked. Very similar phenomena also exist in other languages.

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.

Checking Coding Once samples for testing have been selected the validator’s task is to check the correct coding of features within the entries. For the general samples, this means checking the values assigned for all the features in an entry. For the more specific samples the validator may decide to concentrate only on specific features in entries. Validators must first familiarise themselves with the list of attribute/value pairs used in the lexicon and defined in the documentation, and then make a checklist of features to be checked. Such checklists help guide the validation and record its findings.
However, care should be taken in assessing and recording the coding in a lexicon. In those cases where the validator is in agreement with the original lexicographer's coding, there is no problem with just checking off the feature as being correctly coded. However, the validator should bear in mind that whilst certain features have values which are more or less fixed (and can easily be judged as either correctly or incorrectly coded), other features may be more open to interpretation. For example, gender would seem to be an attribute whose value will be either right or wrong. On the other hand, more syntactically oriented features such as countability (if it is used at all) might be more open to interpretation since it seems that a number of non-count nouns can, in the right context also be countable, e.g.: "This shop sells a number of different cheeses/wines/spring waters...". Similarly the assignment of subcategorisation frames is also very much dependent on the original lexicographer's linguistic approach and interpretation. It may not always be an error not to assign a particular feature. Again in the case of subcategorisation frames, it may be that certain possible frames may not be assigned to a particular word. This could be determined by the depth of coding which has been aimed at by the original lexicographer.

Thus in devising the checklist and looking at the coding, the validator must decide on the relative importance of the different features. Since a whole range of different types of lexicon can be validated including some which concentrate only on a particular level or phenomenon (e.g. synonymy, collocation or subcategorisation frames), the importance of specific features is closely tied to the type of lexicon and its stated purpose or application. A report summarising the criteria used in selecting samples and the results obtained in content validation will be produced. This will be synthesised with the results from the formal validation to produce a report which is as informative as possible to allow potential users to decide whether the lexicon is useful for them. The draft manual contains a template for the overall report to guide the validation and ensure consistency between validations.

Factors Affecting Feasibility

The methodology proposed is still in draft form and it is expected that it will be subject to a number of revisions and refinements once it has been thoroughly tested in the field. Whilst we have tried to take into account the variability of a number of different dimensions the overall feasibility of this approach can only be seen once it has been tested.

One highly variable factor affecting feasibility is the cost of carrying out these recommendations in different countries. Whilst the distribution agency is a centralised body, it is expected that the expert validators will be located in a number of different countries depending on the languages involved, thus giving rise to very different costs. 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 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 the distributor could choose in particular instances, to have a more thorough-going validation.

Another factor which may at first sight appear to make the methodology less than feasible is the amount of effort required from the expert validator in designing a content evaluation from scratch for a specific lexicon. However, once the sampling techniques and characteristics have been developed for specific lexica, they should be at least partly re-usable for other similar lexica. Especially for the more closely related languages and lexicon types, particular phenomena to be checked will be pertinent to a range of individual validation tasks. Expert validators are requested to make the language and lexicon specific phenomena they have identified available as feedback which can be incorporated as appendices to later versions of the manual. Over time, it is expected that a library of content characteristics can be built up, thus cutting down on the work needed for each new validation.

Conclusions

The assumption that as many useful lexical resources as possible should be available for re-use within the NLP community has a profound effect on the notion of what it means to validate a lexicon for distribution. Flexibility and rigour have become keywords in this enterprise.

The methodology is flexible in the range of different lexicon types which can be considered and in the sort of linguistic features which an expert validator may define as the most important to check, as well as leaving open the possibility for a more thorough-going validation to be commissioned for a particular lexicon. Making the output of validation a detailed report, allows for lexica which may not be perfect in all characteristics to still be made available to potential users.

Allowing many different types of lexicon to be accepted for distribution has the effect that there are not many of absolute requirements in terms of e.g. size, richness of information or coverage. However this is not to say that the validation process lacks rigour. As we have seen there are a small number of minimal requirements which must be met and, perhaps more importantly, the validation process involves the decomposition of the lexicon into many detailed characteristics, all of which must be checked and reported on.

This combination of flexibility and rigour is made possible largely due to the detailed and stringent requirements on the documentation which must accompany a lexicon delivery, and which are described in detail in the draft manual. As mentioned earlier, the methodology is still in draft form and will be subject to testing and possible revision. Although the context in which the methodology has been developed has dictated the need for flexibility and a certain open-endedness, there is nothing in principle to preclude the setting of absolute requirements on more of the characteristics being checked. Thus the methodology might also be regarded as a framework which can support the design of validations for specific lexica where the user requirements and context of use are known and translated into absolute requirements which the lexicon must fulfil.
Acknowledgements

The methodology and a draft validation manual described in this paper were developed under contract to ELRA with the help of stimulating and fruitful discussions with and suggestions from Costanza Navarretta, Bente Maegaard, Anna Braasch and Henning Spang-Hanssen. In addition I would like to thank the experts from ELRA Panel for Validation of Written Language Resources (EPV-WLR) who gave valuable feedback on earlier versions of the manual.

References


