Exploring XML-based Technologies and Procedures for Quality Evaluation from a Real-life Case Perspective

Folkert de Vriend\(^1\) & Giulio Maltese\(^2\)

\(^1\)Speech Processing Expertise Centre (SPEX)
Radboud University Nijmegen, The Netherlands
folkert@spex.nl

\(^2\)IBM Rome Solutions Lab and Voice Technology Development
Rome, Italy
giulio.maltese@it.ibm.com

Abstract

The use of Extensible Markup Language (XML) for the annotation of Spoken Language Resources (SLR) is becoming increasingly common these days. Therefore the Speech Processing EXPertise centre (SPEX), which is the SLR validation centre of the European Language Resources Association (ELRA), is also being confronted more with XML. The project \“Lexica and Corpora for Speech-to-Speech Translation Components\” (LC-STAR) is a project that uses XML for the encoding of its resources. For SPEX, XML-based annotations are still relatively new data formats, which is why at SPEX XML-based quality evaluation (validation) technologies and procedures are being explored. This is done using the XML encoded phonetic lexica developed in the LC-STAR project as a test bed.

1. Introduction

In projects in which large Spoken Language Resources (SLR) are being developed, the use of the Extensible Markup Language (XML) [1] to annotate the data is on the increase. When compared to annotations in NIST headers (“National Institute of Standards and Technology”), one of the main differences with using XML is that the annotations are separated from the speech file. They can even be on another server and be linked to the speech files by using the extended linking mechanisms XML and its related technologies (XLink, XPath and XPointer) provide. In comparison to annotations in SAM label files (“Speech Assessment Methods”), used in SpeechDat-like projects [6], XML offers a more convenient manner to deal with the hierarchical information often present in this kind of resources. Many tools are available to make annotations in an XML format directly, or alternatively convert annotations made with traditional annotation tools into XML automatically. This can be done as a background process. Especially for annotating multilingual SLR the support for Unicode in XML and XML-based tools and technologies is an advantage.

The Speech Processing EXPertise centre (SPEX) is the SLR validation centre of the European Language Resources Association (ELRA). For SPEX (www.spex.nl), SLR’s containing annotations in XML formats open a new area of validation potentials. Until recently only SLR’s in the SpeechDat format (SAM-oriented) were validated [2]. Examples of such projects SPEX is or was involved in for validation are SpeechDat (II) [3] and OrienTel [4]. Currently SPEX takes part in the projects \“Technology and Corpora for Speech to Speech Translation\” (www.tc-star.org) and \“Lexica and Corpora for Speech-to-Speech Translation Components\” (LC-STAR), which both use XML for annotation.

This paper explores XML-based technologies and procedures for quality evaluation of Spoken Language Resources. The LC-STAR project offers excellent opportunities for testing out these XML-based technologies and procedures in comparison with more traditional validation procedures. Section 2 gives some more background information on the LC-STAR project. Section 3 then discusses current issues in the validation of the phonetic lexica developed in LC-STAR. Section 4 elaborates upon the ways validation of XML-coded resources can be changed and ultimately enhanced by using the XML-based technologies that have emerged over the last couple of years. By doing so, the XML encoding of the resource will be exploited to its full potential.

2. LC-STAR

The project \“Lexica and Corpora for Speech-to-Speech Translation Components\” (LC-STAR) aims to develop both lexica for automatic speech recognition (ASR) and text-to-speech synthesis (TTS), as well as bilingual corpora for Speech-to-Speech Translation (SST) applications, especially those based on a statistical approach. These applications are targeted to be integrated into speech-driven interfaces embedded in mobile appliances and network servers, thus improving human-to-human and man-machine communication in multilingual environments. In the project, lexica for thirteen languages and bilingual corpora for speech-centred translation applications for nine languages are developed. The LC-STAR consortium has adopted XML for the encoding of all the resources to be developed.

Annotated SLR’s have been extensively developed in many languages and acoustic environments. However, there is a lack of linguistically oriented resources that specifically fulfil the needs of ASR/TTS components of SST applications. In LC-STAR, resources are developed that are linguistically oriented. The lexica have to contain detailed grammatical, morphological (lemma), and phonetic information for each language. A formally specified grammar (Document Type Definition) containing all the described linguistic information allows for automatic validation of several important aspects of the XML-based lexica. Figure 1 shows an example of a fragment of the phonetic lexica developed in the project.
3. Current validation issues in LC-STAR

The validation of the phonetic lexica is currently being done in two stages: a pre-validation and a full validation stage. The pre-validation phase is essential for signalling problems at an early stage of the production. Also, the other partners can still learn at this stage from the typical errors made by one producing partner. The pre-validation stage has already finished. For a broader perspective on the validation criteria and processes used for the phonetic lexica see [5].

Typical for validation of XML-encoded SLR’s is the fact that producers can do some of the validation themselves by using a Document Type Definition (DTD). As pointed out in the previous section in the LC-STAR project a special DTD was designed for project-internal use. Still there are many tasks left for the validation centres to perform.

First of all, validation of SLR’s that will be distributed to end-users via ELRA, like the LC-STAR lexica, has to be carried out by independent validation centres. Currently these are SPEX for SLR’s and the Center for Sprøgteknotologi (www.cst.dk) for written Language Resources.

Second, the consortium partners in LC-STAR have to be sure that not only their own material is valid according to the latest version of the generic DTD, but also that the lexicon of another partner is valid according to this DTD. One of the functions of the validation centres is assuring that each partner provides SLR of equal quality. Only this equal quality in the final SLR allows a fair exchange between partners at the end of the project. So although partners can check their XML against the DTD themselves, there still is the need for an independent centre to check all the XML lexica against the correct DTD.

Third, although DTD-based validation is very useful, it does not nearly cover all checks usually performed by the validation centres. This is obvious when taking into account the non-formal checks on phonetic transcriptions for instance, but also for the formal checks there is still a lot that cannot be checked by simple DTD validation. For instance for LC-STAR, automatic checks were performed testing whether:

- Correct numbers of entries per domain (names / words) are present according to the specifications.
- Only valid phonetic symbols are used according to the documentation provided.
- Fields for lemma and phonemic representation are never empty.

In these cases the checks were performed with special software written in Perl. An example of one of the shortcomings of DTD-based validation is the fact that the DTD cannot be used for checking element content. It cannot even be used for checking on empty elements. One can state in the DTD that the element LEMMA has PCDATA (parsable character data). But because the content model definitions in a DTD are only used for controlling the structure of the XML document, the occurrence of an empty element LEMMA will be valid according to any XML parser, regardless of the PCDATA statement. All in all a big drawback of DTD-based validation is that the datatype system of DTD’s is very weak and only applies to attributes.

4. Exploring XML-based validation technologies and procedures

In the prevalidation phase for LC-STAR, SPEX used Perl software without XML-based technologies for most of the processing of the lexica. This was a logical decision because at SPEX most validation experience comes from working with SpeechDat or SpeechDat-like SLR’s that have meta files in a flat SAM or tab-separated format [6]. However, by applying SpeechDat-like programming procedures to the LC-STAR lexica none of the available XML tree structure information is being used. The XML file is just treated as a “very decorated” text stream and no full advantage is taken of the rich encoding. Because SPEX is increasingly being confronted with SLR’s that use an XML encoding for their annotations, SPEX investigates the possibilities for using XML-based technologies and procedures for carrying out validation checks that go further than DTD-based checks, thereby taking full advantage of the available XML encoding.

The LC-STAR project offers opportunities for testing out these technologies and procedures next to the “traditional” way of text stream processing.

4.1. Substitutes for current validation procedures

Two of the interesting technologies built on the XML specifications with which some of the standard validation tasks carried out at SPEX can be performed are W3C XML Schema’s [7] and XSL Transformations (XSLT) [8]. For the LC-STAR phonetic lexica SPEX had to perform checks on the following kinds of element content:

- Spaces and underscores: spaces cannot occur in certain types of elements (proper names), but underscores can.
✓ Digit characters: digits are not allowed in certain types of elements.
✓ Stress and boundary markers: all LEMMA elements should contain at least one stress marker.  

Now where the DTD as discussed cannot be used for checking these kinds of element content, W3C XML Schema’s (WXS) can. With WXS one can define restrictions on string datatypes by defining a pattern which can be a regular expression. This way different restrictions can be specified for different elements. For the LC-STAR phonetic lexica a schema especially designed for validation purposes could be made. A basic schema could be generated from the current lexicon DTD. For this there are special software tools. Then one can start defining extra restrictions in the schema on the element content. For instance in a schema definition for the check on spaces and underscores one could simply use the “Name” datatype for which spaces are illegal but underscores are not.

Some checks performed during validation have to be done manually, for instance checks on phonetic transcriptions and POS tagging. Because of the labour-intensive nature of these checks they are usually based upon a sample of the complete SLR. These samples are derived from the SLR by the use of special selection software written in Perl. However, taking samples out of an XML file comes down to taking out certain nodes of the XML tree structure which is something that can be done with XSLT. Especially the latest version XSLT 2.0 [9] is interesting because it offers support for regular expression matching.

Another advantage of the use of XSLT for selecting samples for manual validation is the easy way in which the information stored in the XML can be transformed into a human readable format. Validators checking the POS tags of a lexicon file, for instance, could be directed to a website where they can open an HTML page that was directly generated from the XML file under validation. This way selection and reformattting of the sample is being done in one XSLT instructions file.

4.2. New validation procedures

Some XML-based techniques enable validation procedures that were previously not (easily) possible. Two examples are the checking of character sets and on-line WXS validation.

XML fully supports Unicode, which is ideal for a project like LC-STAR in which bilingual corpora and phonetic lexica are being produced for many different languages. The 13 languages covered by the lexica in LC-STAR are Arabic, Catalan, Finnish, German, Greek, Hebrew, Italian, Mandarin, Russian, Slovenian, Spanish, Turkish and US-English. Unicode supports all of these languages, which would make it possible to construct for instance a trilingual corpus that has aligned Chinese, Russian and English sentences in one and the same XML file. The downside of this freedom is that there is more need to control what character subsets of Unicode are being used in the different parts of the XML file. Chinese characters could appear in XML elements that should only contain English characters. Also not all applications – especially older ones - can fully handle Unicode and as Eric Wilde states: “For a very long time to come, it is not realistic to expect IT infrastructures to support full Unicode” [10]. So it is important to check on character sets. For the LC-STAR phonetic lexica the following checks were performed on character sets:

✓ All and only SAMPA phoneme symbols are used as contents of the element PHONETIC.
✓ Only ISO 8859-X is used as contents of the element LEMMA for European languages, Arabic and Hebrew. For Chinese only GB2312 is used.

For checking character sets in XML files, a specialized schema language can be used instead of WXS, namely CRVX (Character Repertoire Validation for XML). CRVX is a schema language for specifying character repertoire constraints. Restrictions are supported for character sets which are subsets of Unicode [11]. In LC-STAR we check for instance that in the Italian lexica the LEMMA elements contain characters from the ISO-8859-1 character set. At present this is only done on a manual sample basis because programming checks for all used Unicode character subsets is too laborious. However, in a CRVX Schema the check on the ISO-8859-1 set could be elegantly formulated as shown in Figure 2.

**Figure 2**: Example of CRVX code

```xml
<crvx structures="namespaceXML">
  <context path="ENTRY/LEMMA">
    <restrict charrep="\p{IsBasicLatin} \p{IsLatin-1Supplement}">
    </context>
  </crvx>
</crvx>
```

SPLEX distinguishes two different kinds of settings in which validation criteria are developed. In the first setting SLR are completed in a framework where validation by a non-producing partner is an integral part of the SLR production process. In the other setting, validation is not an integral part of the SLR production and is done post-hoc [12]. LC-STAR is an example of the first setting. In this setting, the validation centre is involved from the beginning of the specifications of the SLR and validation criteria are closely linked to the specifications.

Since XML is designed to be able to handle distributed data over the internet, it enables us to create a framework in which the development of XML-encoded SLR is directly linked to the formal criteria while keeping it clear who is at the creation side and who is at the validation side of the SLR development process. In this framework, the validation centre can create a special WXS for validation purposes - as discussed in the previous section - and put this on a server where it can be externally referenced. This way the producing consortium partners always have access to the latest set of...
validation criteria as laid down in the WXS². Producers can validate the XML they produce against the WXS from the start of the project by referencing the on-line available WXS from their production and editing environment. For instance many XML editors will check the XML file against the referenced WXS automatically when you save the file you have been working on, forcing you to adhere to the WXS.

This way all parties stick to their main task in the framework: the producers produce XML and the validation centre defines rules the XML should adhere to. They do this in their own production environment and only the data set that is the smallest (the set of rules in the WXS file) is the one that is being transported over the network causing the least network load.

In most projects that SPEX is involved in validation is carried out in stages [12]. At several points in time a version of the SLR is checked. At this point in time errors are found and reported back to the producer. If these errors are severe the producer of an SLR sometimes needs to do a significant number of repair operations. Now with the XML-encoded data the producer can validate its resources. This makes it possible to detect errors earlier and to be more time and cost effective in producing the resources.

5. Conclusions

XML-based techniques and tools are considered by SPEX a very promising - and on principle preferable - alternative for validation of XML with “traditional” text stream processing procedures that are not truly XML-aware. The LC-STAR project offers an excellent opportunity for comparing both techniques. At least the use of W3C XML Schema and XSLT for standard validation purposes seems to offer many interesting possibilities, especially when making use of the continuous validation functionality. With the CRVX Schema language, checks on Unicode character repertoires can be performed in an easy and elegant way. Finally, on-line access to validation rules in a W3C XML Schema (WXS) file provides producers with a self-monitoring tool that is always up to date. Ultimately, the continuous checking possibilities during production will result in errors getting detected earlier, which reduces the costs of fixing errors found at an intermediate validation stage.

Right now SPEX is testing if some of the XML-based techniques and tools discussed in this paper can already be implemented for the full validation stage of the LC-STAR project at the moment seems somewhat vague, as there is not much documentation publicly available and several parts of the standard seem to be still very much under development.

6. Acknowledgements

The LC-STAR project was funded under EU Contract IST-2001-32216.

7. References


² By the nature of this validation setting the criteria always change over the duration of the project.