

LC-STAR: XML-coded Phonetic Lexica and Bilingual Corpora for Speech-to-Speech Translation

Folkert de Vriend¹, Núria Castell², Jesús Giménez², and Giulio Maltese³

¹ Speech Processing Expertise Centre (SPEX), Radboud University Nijmegen,
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

folkert@spex.nl

² TALP Research Center, Universitat Politècnica de Catalunya,
Campus Nord – {C5 / A0}, Jordi Girona, 1-3
08034 Barcelona, Spain

{castell/jgimenez}@talp.upc.es

³ IBM Italy Software Group - Rome Solutions Lab and EMEA Voice
Technology Development, Via Sciangai 53
00144 Rome, Italy

giulio.maltese@it.ibm.com

Abstract. The project “Lexica and Corpora for Speech-to-Speech Translation Components” (LC-STAR) aims to develop lexica for automatic speech recognition and text to speech synthesis for thirteen languages, and multilingual corpora for speech centered translation applications for nine languages. The project is led by a consortium comprising two universities and several industrial companies. All resources to be developed are encoded using the Extensible Markup Language (XML). This paper describes XML related issues in the LC-STAR project from three different perspectives; the XML encoding of the lexica, the XML encoding of the multilingual corpora and issues regarding the validation of XML encodings like that of the LC-STAR lexica.

1. Introduction

The European project “Lexica and Corpora for Speech-to-Speech Translation Components” (LC-STAR / IST-2001-32216) aims to develop both lexica for automatic speech recognition (ASR) and text to speech synthesis (TTS) as well as bilingual corpora for speech centred translation 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. A speech to speech translation demonstrator is under development to show the benefits of the created language resources for this kind of applications. The project started in February 2002 and ends in January 2005.

In the last decade an increasing interest in multilingual resources has arisen among the Natural Language Processing (NLP) community members for their exploitation in a number of tasks such as Machine Translation (MT) (Melamed, 1996) (Och & Ney,

2003) (Och et al., 2003), Information Retrieval (IR) and Information Extraction (IE) (Dumais, Landauer & Littman, 1996) (Baeza-Yates & Ribiero-Neto, 1999), Word Sense Disambiguation (WSD) (Schütze, 1997) (Diab & Resnik, 2002), enriching of Lexical Resources (LR) (Widdows, Dorow & Chan, 2002) etcetera, as well as for pedagogical applications (Maia & Sarmento, 2003) (Hoey, 2003). A very convenient manner to deal with the hierarchical information often present in these kind of multilingual resources is by means of the Extensible Markup Language (XML) (Bray, Paoli, Sperberg-McQueen, & Maler, ed., 2000).

The well-known advantages of XML such as standardisation, usability, portability, scalability and robustness counteract its main disadvantage: the huge disk/memory space consumption. Some examples of projects dealing with the creation of XML annotated lexical resources are the MATE (Multilevel Annotation, Tools Engineering), MULTEXT (Multilingual Text Tools and Corpora) and MULTEXT-EAST (Multilingual Texts and Corpora for Eastern and Central European Languages) projects. Also in the Speech Technology community, several approaches have been tried to the creation of XML multilingual speech corpora both for Speech Recognition and Synthesis (Isard et al., 1998) (Klabbers & Stoeber, 2001). The LC-STAR consortium has also adopted XML for encoding all the resources to be developed.

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. The project however is open for new partners who want to offer lexica for other languages. The lexica were created using electronically available text corpora and must contain common words, proper names and special words for voice-driven applications (Hartikainen et al., 2003). For the corpora the languages covered are Catalan, Finnish, German, Hebrew, Italian, Russian, Slovenian, Spanish and US-English. For the creation of the aligned bilingual corpora US-English transcribed speech data and transcriptions of conversations recorded in Catalan and Spanish were used. After human translation a trilingual corpus was obtained (Arranz, 2003). The bilingual corpora are still under development. All the developed resources in this project will be publicly available in 2005 through ELRA.

The LC-STAR consortium comprises two universities and a number of industrial companies with Siemens as the project coordinator. The universities participating are RWTH-Aachen (Rheinisch-Westfälische Technische Hochschule Aachen) and UPC (Universitat Politècnica de Catalunya). The University of Maribor is an external partner to the consortium. The other industrial partners are IBM, Nokia and NSC. The two official validation centres of the European Language Resources Association (ELRA) were contacted by the consortium for validation of the phonetic lexica. Of these two independent validation centres, SPEX (Speech Processing EXPertise Centre) is the one responsible for validation of the formal and phonemic part of the lexica, and CST (Center for Sprogteknologi) for the validation of the morphological and syntactic information.

This paper describes different XML related issues in LC-STAR. Section 2 gives background information on the reasoning behind the features of the XML encoding of the phonetic lexica. In section 3, information is provided on the XML encoding of the bilingual corpora that are being produced in the frame of LC-STAR. A flexible DTD, able to represent either bilingual or multilingual corpora, is described. Moreover, several levels of alignments for bilingual corpora could be represented by this DTD.

Finally, section 4 describes validation issues in LC-STAR that are related to the XML encoding of the phonetic lexica discussed in section 2.

2 Design of Phonetic Lexica

Annotated speech databases have been extensively developed in many languages and acoustic environments. However, there is a lack of linguistic oriented resources that specifically fulfil the needs of ASR/TTS components of Speech-to-Speech Translation (SST) applications. In LC-STAR resources are developed that are linguistic oriented. The lexica have to contain detailed grammatical, morphological (lemma), and phonetic information for each language.

The grammatical information needed per language within the scope of the lexica was specified at the beginning of the project. The available information in all languages was merged into a unique list of POS tags (part-of-speech tags). Most of the POS have an internal structure with attributes common to several languages (e.g. number or gender), but some POS relate only to a subset of languages (e.g. case), or are relevant only for specific individual languages (e.g. polarity for Turkish verbs). The advantage of such an approach is that a single description of grammatical features can cover the whole set of languages of the project. Additionally, this scheme is open and easily extendable for integrating further languages.

In addition to grammatical features, for each word lemma and phonetic transcription are specified. Phonetic transcriptions use SAMPA symbols, which are specified for each lexicon. Syllable boundaries and primary stress are also provided in the phonetic transcriptions. For each word, it is possible to specify more than one POS, and/or more than one lemma, and/or more than one phonetic transcription.

The formal representation of the lexica is implemented via an XML-based mark-up language. A formally specified grammar (DTD) containing all the described linguistic information allows for automatic validation of several important aspects of the XML-based lexica. A lexicon consists of a set of entry group elements. An entry group refers to a generic entry in a vocabulary. The spelling of an entry is the key to the entry group. For each entry group, it is therefore mandatory to specify an orthography (with allowed alternative spellings) plus one or more entry or compound entry elements. An entry refers to one specific grammatical / morphological category of a vocabulary entry, like *can* (verb) and *can* (noun). For each entry, one or more POS, one or more lemmas and one or more phonetic transcriptions must be specified. Special tags are used to mark the words that are included in a separate list of words relevant to special applications. For abbreviations, multiple expansions can be specified.

In some languages (e.g. Catalan, Hebrew, Italian, Spanish, Turkish) there are assimilation or agglutination phenomena that can be dealt with by using compound entries. Besides its spelling and its phonetic transcription, a compound entry is composed of a series of two or more entry elements (a subset of an entry) which are simply links to other entries. Each entry element must be characterized by an orthography and must contain one POS tag, together with all of its attributes. Figure 1 shows an example of the XML encoding of entry groups.

```

<ENTRYGROUP orthography="sveta" xml:lang="sl">
  <ENTRY>
    <NOM class="common" gender="masculine"
      number="singular" case="genitive"/>
    <LEMMA>svet</LEMMA>
    <PHONETIC>s v E - " t a:</PHONETIC>
  </ENTRY>
  <ENTRY>
    <ADJ gender="feminine" number="singular"
      case="nominative" degree="positive"/>
    <LEMMA>sveta</LEMMA>
    <PHONETIC>" s v e: - t a</PHONETIC>
  </ENTRY>
  <ENTRY>
    <ADJ gender="feminine" number="singular"
      case="accusative" degree="positive"/>
    <LEMMA>sveta</LEMMA>
    <PHONETIC>" s v e: - t a</PHONETIC>
  </ENTRY>
</ENTRYGROUP>

```

Figure 1 : Example of XML encoding of lexica for Slovenian

3 Design of Multilingual Corpora

Usually, multilingual information is available in the form of raw text split in several files, sometimes one file per language, one file per document, or one file per language and document. At content level, we can distinguish between comparable corpora, parallel corpora or aligned corpora. When alignments are possible, a hierarchical structure would be best in order to preserve and process the data aligned at different levels. Because the granularity of the alignment may vary from non-aligned or corpus aligned to document aligned or section/ paragraph/ segment/ sentence aligned, and even, not very often though, word aligned. In general, the finer the granularity of the alignment the more these data are appreciated. For MT purposes at least a segment alignment is desirable, where a segment consists of as few sentences as possible.

3.1 Encoding issues for multilingual corpora

XML approaches to the creation of multilingual corpora have been already successfully developed and applied. Many of them are based either on the Text Encoding Initiative (TEI) guidelines for electronic text encoding and interchange (McQueen & Burnard eds., 2002) or on the Corpus Encoding Standard for XML (XCES) (Ide, Bonhomme & Romary, 2000) based on the Corpus Encoding Standard (CES) developed by the Expert Advisory Group on Language Engineering Standards (EAGLES). Other than these, there is the corpus and document interchange format (CDIF) used only in the British National Corpus (BNC). Both TEI and XCES are NLP oriented. TEI is too tough to work with, hard both to read and write for

researchers. Moreover, TEI presents problems to encode grammatical and prosodic information at the same time because segments may not embed. As to XCES, it is much simpler, and it seems to match the needs of the LC-STAR project resources well. Unfortunately, the currently available version of XCES is still a beta version.

Therefore, it was decided to define a new format by building a new DTD for multilingual corpora, bearing in mind that the goal of the LC-STAR project is the development of resources for speech-to-speech machine translation. The nature of the LC-STAR corpora, some coming from spontaneous speech recordings some not, required the representation schema to be flexible and rich enough to store the big amount of information involved while still being manageable. This new schema presents a multilingual corpus as a “document repository”, a collection of “documents” (e.g. news / dialogues). These are in their turn divided into “sections” (e.g. document paragraphs / dialogue turns). Each “section” may consist of several “segments”. The segment is intended to be the required alignment unit in the sense that all documents must be mandatory aligned at least at the segment level. Each segment consists of as many “lsegments” (language dependent segments) as languages are involved in the corpus (two in the case of a bilingual corpus). Examples of such an XML document may be seen in Figure 2. Figure 3 presents a second example, including linguistic features and using a more compact DTD.

```
<DOC_REPOSITORY DATE="10/2/2004">
  <DOC ndoc="000">
    <SEC nsec="000" S="1">
      <SGM nsgm="000">
        <LSGM language="CA"
          content="hola Mary , com està ?"/>
        </LSGM>
        <LSGM language="EN"
          content="hi Mary , how are you ?"/>
        </LSGM>
        <LSGM language="ES">
          content="hola Mary , ¿ cómo está ?"/>
        </LSGM>
      </SGM>
    </SEC>
  </DOC>
</DOC_REPOSITORY>
```

Figure 2 : LC-STAR trilingual Corpus XML sample aligned at segment level

An “lsegment” may either be simply a raw text string or consist of “words”, “compound-words”, “multi-words” and “acoustical-events”. A “word” is a linguistic unit or token that may carry linguistic features (lemma, part-of-speech, is_a_neologism?, is_a_letter_spelling?, wordnet, synset, is_an_acronym?, is_a_foreign_word?). Moreover, a “word” may also have acoustic features (is_interrupted?, is_badly_pronounced?). A “compound-word” is still a single word that is resulting from the union of separate words. A “multi-word” may consist of several words. They may form, for instance, named entities, dates, idioms, phrasal verbs, light verbs, compound nominals etcetera. Besides, multi-words allow the representation of syntactic and semantic parsing, i.e. a multi-word may be a syntactic phrase or play a semantic role. Finally, “acoustic events” comprise phenomena such

as filled pauses, noises, technical interruptions, unidentifiable parts of an utterance, repetitions or corrections and false starts.

However, it is still under discussion the way in which word (and/or phrase) alignments in bilingual corpora are to be implemented. Several strategies are being studied. While some imply redundancy, some are not functional enough, and some others are not really human readable. Different linking standards are being considered, apart from naming conventions.

```

<LSGM language="CA">
  <W L="hola" P="INT">hola</W>
  <W L="Mary" P="NOM">Mary</W>
  <W L="," P="PUN">,</W>
  <W L="com" P="CON">com</W>
  <W L="estar" P="VER">està</W>
  <W L="?" P="PUN">?</W>
</LSGM>
<LSGM LAN="EN">
  <W L="hi" P="INT">hi</W>
  <W L="Mary" P="NOM">Mary</W>
  <W L="," P="PUN">,</W>
  <W L="how" P="CON">how</W>
  <W L="be" P="VER">are</W>
  <W L="you" P="PRO">you</W>
  <W L="?" P="PUN">?</W>
</LSGM>
<LSGM LAN="SP">
  <W L="hola" P="INT">hola</W>
  <W L="Mary" P="NOM">Mary</W>
  <W L="," P="PUN">,</W>
  <W L="¿" P="PUN">¿</W>
  <W L="cómo" P="CON">cómo</W>
  <W L="estar" P="VER">está</W>
  <W L="?" P="PUN">?</W>
</LSGM>

```

Figure 3 : Trilingual “segment” incorporating some linguistic features (lemma “L” and part-of-speech “P”)

3.2 LC-STAR bilingual corpora

Bilingual resources are very valuable for a wide range of NLP applications such as word sense disambiguation, learning foreign languages environments or Machine Translation, especially for the statistical approach. When these corpora are aligned at word level, they are really useful, but of course this kind of resources requires a big amount of human work. Two different kinds of bilingual resources are under development in the LC-STAR consortium. The first kind are eight bilingual phrasal lexica that should be created. From the tourist domain, 10K segments in US-English have been selected. For each target language, the US-English segment together with its translation will be represented according the described DTD. These bilingual phrasal lexica should contain 10K segments where each segment should have two Lsegments: one with the US-English segment as it is, and one with the target

decomposed in words which are each followed by the lemma and the basic POS tag (using the same tagset as for the lexica discussed in the previous section). Optionally, more Lsegments for alternative translations could be added. Target languages for these phrasal lexica are: Catalan, Finnish, German, Hebrew, Italian, Russian, Slovenian, and Spanish.

The second kind of bilingual resources under development are three parallel bilingual corpora that should be created from a trilingual corpus in Catalan, Spanish, and US-English. Each corpus should have about 500K words per language. In this case, segments will be always decomposed in both languages of the bilingual corpora. The final goal for these corpora is to enrich them with alignments in deeper levels than the segment. A compact version of the DTD will be used for these corpora.

4 Validation of Phonetic Lexica

The formal and phonemic validation of the LC-STAR phonetic lexica that were discussed in section 2 is carried out by SPEX. Validation is 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 production and is typical for larger projects with longer development time. The pre-validation stage for the LC-STAR phonetic lexica has already finished (see also Hartikainen et al. 2004).

4.1 Current validation issues

Until recently only Spoken Language Resources in the SpeechDat format were validated at SPEX (Hoegge et al., 1999). These have annotations that are SAM (“Speech Assessment Methods”) oriented. XML encoded resources are relatively new for SPEX and pose new questions for validation.

Producers of XML-encoded resources can do some validation themselves by defining and using a DTD. As pointed out in section 2, for LC-STAR a special DTD was designed for project-internal use. This was done at the beginning of the project and thus the DTD described one basic frame work for all lexica to be developed. For validation purposes however there is a need for a tighter DTD so later on in the project the generic DTD was supplemented by a language specific set of DTD rules for each language. This set of rules is only used for validation and is more strict on grammatical attributes. For instance, it ensures that a noun in Turkish can not be given an attribute “gender” while a noun in Italian can, and in fact should have that attribute for each noun.

Validation based on a DTD is very useful but many of the formal checks that are usually performed by the validation centres cannot be done with a DTD (no matter how language specific it is). This is mainly because the DTD cannot be used for checking element content. For LC-STAR formal checks were performed testing on: valid phonetic symbols, numbers of entries per domain, and empty fields for lemma and phonemic representation. These checks were performed with special validation software written in Perl.

4.2 XML-based validation strategies

Since the emergence of XML many different technologies have been developed that provide ways of accessing and manipulating the XML. For these technologies also many pieces of off-the shelf software have become available, some of them are Perl modules. Now although the standard Perl software approach followed for LC-STAR, compared to DTD based validation, is capable of doing many more of the usual checks performed in validation, none of these technologies that have been built on top of the XML standard are being used. This way no full advantage is taken of the available XML tree structure information. For this reason and the fact that SPEX is increasingly being confronted with validation of resources that have their annotations encoded in XML, SPEX investigates the possibilities for using these new XML based technologies and software tools for carrying out validation checks.

Interesting validation functionality is offered by schema languages. For the LC-STAR phonetic lexica checks were performed on the following kinds of element content:

- Spaces and underscores: spaces cannot occur in certain types of elements (proper names), but underscores can.
- Stress and boundary markers: all elements LEMMA should contain at least one stress marker. Boundary markers may occur.
- Only ISO 8859-X is used as contents of the element “lemma” for European languages, Arabic and Hebrew. For Chinese only GB2312 is used.

As previously noted the DTD cannot be used for checking these kinds of element content. Schema languages however can. With W3C XML Schema’s (Thompson et al., 2001; Biron & Malhotra, 2001) one can define restrictions on string datatypes. These restrictions can be defined by a pattern which can be a regular expression. This way one could implement the check on LEMMA elements that must have at least one stress marker.

For checks on character sets like the check on ISO 8859-X and GB2312, a special kind of schema language can be used instead of W3C XML Schema, namely CRVX (Character Repertoire Validation for XML). CRVX is a schema language for specifying character repertoire constraints for XML documents. Restrictions are supported for character sets which are subsets of Unicode (Wilde, 2003). Figure 4 shows an example of CRVX code specifying that a LEMMA element must contain characters from the ISO-8859-1 character set:

```
<crvx structures="namespaceXML" version="1.0"
  xmlns="http://dret.net/xmlns/crvx10">
  <context path="ENTRY/LEMMA">
    <restrict charrep="\p{IsBasicLatin}\p{IsLatin-1Supplement}"/>
  </context>
</crvx>
```

Figure 4 : Example of CRVX code restricting LEMMA content

Other applications of XML-based technologies that are of interest to SPEX are remote validation via on-line accessible schema rules, and XSL Transformations (XSLT) (Clark, 1999) for taking samples out of resources and transforming them for the purpose of manual validation checks.

5 Conclusions

By going into detail about the reasoning behind the encodings of both the phonetic lexica and multilingual corpora under development in LC-STAR, this paper has shown that XML offers a convenient way of encoding the hierarchical information often present in these kinds of richly encoded linguistic resources. The paper described a DTD for phonetic lexica that is able to code, in a uniform way, a wide range of linguistic phenomena for a set of very different languages. Validation issues related to these lexica were also discussed. Moreover, the paper described a flexible DTD for multilingual corpora. As the bilingual corpora are still under development, validation criteria have not been defined for them yet. They will be based on the experiences with the validation of the phonetic lexica.

XML based techniques and tools are considered by SPEX a promising alternative for validation of XML with “traditional” not truly XML-aware text stream processing procedures. At least the use of W3C XML Schema for checking element content in general and CRVX for specialised checks on character sets seems to be very convenient for validation purposes. The LC-STAR project offers an excellent opportunity for testing out these validation techniques.

References

- Arranz, V., Castell, N. & Giménez, J. (2003). Development of Language Resources for Speech-to-Speech Translation. International Conference RANLP - 2003 (Recent Advances in Natural Language Processing). Borovets, Bulgaria.
- Baeza-Yates, R. & Ribiero-Neto, B. (1999). Modern Information Retrieval. Addison Wesley/A CM press.
- Biron, P. V. & Malhotra, A. (2001). XML Schema Part 2: Datatypes. World Wide Web Consortium, Recommendation REC-xmlschema-2-20010502, May 2001.
- Bray, T., Paoli, J., Sperberg-McQueen, C. M. & Maler, E., ed. (2000). Extensible Markup Language (XML) 1.0 (Second Edition), W3C.
- Clark, J. (1999). XSL Transformations (XSLT) version 1.0. Technical report. World Wide Web Consortium, Recommendation. REC-xslt-19991116, 16 November 1999.
- Diab, M. & Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 255-262), Philadelphia, PA.
- Dumais, S., Landauer, T. & Littman, M. (1996). Automatic cross-linguistic information retrieval using latent semantic indexing. In SIGIR 96 Workshop on Cross-Linguistic Information Retrieval (pp. 16-23).
- Hartikainen, E., Maltese, G., Moreno, A., Shammass, S. & Ziegenhain, U. (2003). Large Lexica for Speech-to-Speech Translation: From Specification to Creation. In Proceedings Eurospeech 2003, Geneva.
- Hartikainen, E., Maltese, G., Moreno, A., Shammass, S., Ziegenhain, U., Fersø, H. & Van den Heuvel, H. (2004). Creation and Validation of Large Lexica for Speech-to-Speech Translation Purposes. In Proceedings LREC 2004, Lisbon.

- Hoeghe, H., Draxler, C., Van den Heuvel, H., Johansen, F.T., Sanders, E. & Trof, H.S. (1999). Speechdat multilingual speech databases for teleservices: across the finish line. In Proceedings EUROSPEECH'99 (pp. 2699-2702), Budapest, Hungary.
- Hoey, M. (2003). What can the corpus tell us about linguistic creativity?. Plenary lecture at the CL 2003 - Corpus Linguistics - Conference at the University of Lancaster.
- Ide, N., Bonhomme, P. & Romary, L. (2000). XCES: An XML-based Encoding Standard for Linguistic Corpora. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC), Athens, Greece.
- Isard, A., McKelvie, D. & Thompson, H.S. (1998). Towards a Minimal Standard for Dialogue Transcripts: A New Sgml Architecture for the HCRC Map Task Corpus. In Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP98), Sydney.
- Kay, M. (2003). XSL Transformations (XSLT) version 2.0. Technical report. World Wide Web Consortium, Working Draft. WD-xslt20-20031112, 12 November 2003.
- Klabbers, E. & Stoeber, K. (2001). Creation of Speech Corpora for the Multilingual Bonn Open Synthesis System. 4th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW4), Perthshire, Scotland.
- Maia, B & Sarmento, L. (2003). Corpora and the general public. In Proceedings of PALC 2003 (Practical Applications of Language Corpora), Conference at the University of Lodz, Poland.
- Melamed, I. D. (1996). Automatic construction of clean broad-coverage translation lexicons. In 2nd Conference of the Association for Machine Translation in the Americas, Montreal, Canada.
- Och, F. J. Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z. & Radev, D. (2003). Syntax for Statistical Machine Translation. Final Report of Johns Hopkins 2003 Summer Workshop on Syntax for Statistical Machine Translation.
- Och, F. J. & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics archive, Volume 29, Issue 1 (March 2003) (pp. 19-51). MIT Press Cambridge, MA, USA.
- Schütze, H. (1997). Ambiguity resolution in language learning. Stanford CA: CSLI Publications.
- Sperberg-McQueen, C. M. & Burnard, L. (eds) (2002). Guidelines for Text Encoding and Interchange. Published for the TEI Consortium by the Humanities Computing Unit, University of Oxford.
- Thompson, H. S., Beech, D., Maloney, M., & Mendelsohn, N. (2001). XML Schema Part 1: Structures. World Wide Web Consortium, Recommendation REC-xmlschema-1-20010502, May 2001.
- Van den Heuvel, H., Iskra, D., Sanders, E. & De Vriend, F. (2004). SLR Validation: Current Trends and Developments. In Proceedings LREC 2004, Lisbon.
- Widdows, D., Dorow, B. & Chan, C.K. (2002). Using Parallel Corpora to enrich Multilingual Lexical Resources. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC) (pp. 240-245), Las Palmas, Spain.
- Wilde, E. (2003). Character Repertoire Validation for XML Documents. Twelfth International World Wide Web Conference (WWW2003), Budapest, Hungary.