Securing Interpretability: The Case of Ega Language Documentation

Dafydd Gibbon[1], Catherine Bow[2], Steven Bird[2,3], Baden Hughes[2]

[1]Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Germany
gibbon@spectrum.uni-bielefeld.de
[2]Department of Computer Science and Software Engineering, University of Melbourne, Australia
{cbow, sb, badenh}@cs.mu.oz.au
[3]Linguistic Data Consortium, University of Pennsylvania, USA
sb@ldc.upenn.edu

Abstract

The prime consideration in designing sustainable language resources is to ensure that they remain interpretable for coming generations of users. In this paper we adopt a new perspective on resource creation - securing the interpretability of data, using a case study of Ega, an endangered African language for which a small amount of legacy data is available. Basic steps to securing interpretability are to transfer files to durable media, and where possible, to convert all legacy data into XML files with Unicode character encodings. In the absence of agreed 'best practice' standards, we propose a methodology of 'better practice' to assist in the transition process towards this goal. We discuss a number of issues involved in securing interpretability of the lexicon, character encodings, interlinear glossed text, annotated recordings and nomenclature in linguistic descriptions, and describe our solutions.

1. Introduction

In recent years there has been a heightened state of awareness of the issues involving endangered languages, and the urgency of the task of collecting and preserving linguistic data from these languages while they still exist. A range of initiatives such as EMELD, HRELP and DOBES[1] aim to promote best practice in the collection, documentation and archiving of endangered language data, offering information, training, tools, and opportunities for discussion for interested parties. Not only are the languages under threat of extinction, but with the rapid development and reduced effective life span of technologies, even data already collected risks becoming uninterpretable by future generations of users. Unless special care is taken, many resources become unusable within a decade of their creation. This issue is particularly acute in the case of endangered languages, since if the resources have become uninterpretable in the permanent absence of native speakers, there is no prospect of recreating the resource or indeed using it to revitalise the language in future. The term 'uninterpretable' can refer both to understanding the content of archives and to decoding archive data; we refer here to the latter as a prerequisite for the former.

While there are guidelines available for the collection of endangered language data for preservation (Bird and Simons, 2003), there is little reporting on the methodology for securing interpretability of available linguistic data by preservation of both forms and functions in appropriate formats. Quite often a researcher 'inherits' a data set in any number of different formats, yet there are few guidelines available for securing the interpretability of these resources according to 'best practice' methodologies. Even 'best practice' is still under discussion, and so the specific goals for documentation and archiving remain underspecified.

In this paper we propose a systematic methodology to preserve interpretability, and demonstrate its application to the specific case of the Ega language[2] in a practical scenario. While we do not claim that our methodology meets all the criteria of 'best practice', our goal is to assist in the transition towards such a goal. In the following sections we discuss a practical scenario and deal with five kinds of legacy data: the lexicon, character encodings, interlinear glossed text, annotated recordings, and linguistic descriptions.

[1]See www.emeld.org, www.hrelp.org, www.mpi.nl/DOBES.

2. Towards 'better practice'

2.1. Scenario

Suppose we have one person month in which to process legacy language documentation materials which have been presented for archiving. What are the most important steps which must be undertaken in order to preserve interpretability? The case of legacy resources available for the Ega language represents a fairly common scenario, in presenting a lexicon, interlinear texts, annotated digital audio and video recordings, linguistic descriptions collected by various researchers in various formats — and limited processing time. The problems which arise from these resources include

- legacy fonts (font mixtures in the same document; unavailable fonts),
- lexicon structure (interpretation of lexical data categories and lemmatisation decisions),
- annotation conventions (glossing; phonetic, prosodic, visual annotation),
- terminology (in the present case, from specialized English and French linguistic traditions).

[2]Ega (new Ethnologue code: EGA) is an endangered language spoken in South Central Ivory Coast, tentatively classified as Western Kwa, for which relatively small quantities of legacy data resources are available, mainly in various electronic formats.

We provide a high level summary of process; detailed solutions are to be found in the current Ega repositories.[3]

2.2. Procedure

Our general methodology is shown in Figure 1. The very basic first step in securing interpretability is to secure the original file by transferring it to standard digital storage media, with associated font files in the case of text data, and with any available metadata. A second step ensures basic interpretability by providing human-readable versions of documents or lexical databases, such as PDF renderings, since these will contain important clues to character encodings as well as text and record structures, should any of the more detailed documentation be lost. In a third step, paper documents are scanned for digital storage.

Beyond this, current wisdom requires, ideally, that future interpretability be attained by converting all legacy data into XML files conforming to community-agreed conventions (Yergeau et al., 2003), with Unicode character encodings, prior to permanent archiving. We could call this ideal method 'best practice'. There are practical constraints, however. This approach presupposes the availability of standard XML encodings of linguistic data types, suitable conversion tools, suitably trained linguists or archivists to perform the conversion, and adequate financial and human resources. This is impractical in many cases, so instead we propose 'better practice' (or maybe 'not quite worst practice') to be 'the minimal documentation of a resource which significantly helps to secure interpretability over the longer term'.

We distinguish between *language interpretability* and *archive interpretability*; the latter is a prerequisite for the former. Language interpretability minimally requires interlinear documentation of core semiotic features of a language, specifically

1. surface forms (ideally including matching representations of pronunciation and (where relevant) orthography as annotations for audio and/or video recordings),
2. structures (basic phonological and morphosyntactic segmentation and classification),
3. meanings (literal glosses and free translations).

We concentrate here on archive interpretability, which is concerned with sustainable file formats, character encodings, and linguistic markup conventions, all of which are potentially highly volatile. The tools we developed were implemented as prototypes in appropriate scripting languages.

3. Securing interpretability of the lexicon

3.1. Problem statement

In common with many language documentation efforts, the Ega project uses 'Shoebox',[4] a hybrid text markup editor combined with a database management system, with powerful search, display and output functions. Shoebox is popular for its flexibility: users can insert new fields on the fly and re-order existing fields at will. It is also popular for its support of ad hoc character encodings: users can represent information from different languages in different fields. Despite the flexibility, a lack of documentation regarding the use of these features renders the original data uninterpretable. There are three significant interpretability problems with Shoebox data:

1. If character encodings are not documented and the encoding linguist is not available, it may not be possible to recover the intended *character* from its encoding, rendering certain fields useless.
2. Since Shoebox does not require standard labels in markup, if the interpretations of field names (e.g. 'lx' for lexeme) and abbreviated content (e.g. 'N' for noun) are not documented, it may not be possible to recover the intended *interpretation* of the content.
3. If the microstructure of an entry (e.g. the permissible sequence of fields) is not documented, it may not be possible to recover the intended *relationships* between fields (e.g. whether a field consisting of a comment, translation or cross-reference applies to the previous field, a subset of fields, or to the entry as a whole).

3.2. Approach

Lexicon interpretability is a complex issue to which we cannot really do justice here. We addressed these specific problems by adding automatically extracted corpus information, exporting lexicon data to a well-defined XML format, and designing and implementing a method for providing metadata of the following kinds about the lexicon microstructure:

1. Content documentation, which gives information about the interpretation of each field and its contents.
2. Structure documentation, which gives information about the possible optionality and repetition of a field, and the field on which it depends.

4. Securing interpretability of character encodings

4.1. Problem statement

Legacy character encodings cannot be interpreted accurately without access to the original font definition along with suitable software and expertise. This problem is not specific to languages which use exotic characters, in fact, it is relevant to languages which use just a single font. The case of Ega language documentation is quite typical: a variety of fonts and encodings have been used by different researchers using different tools, and not all fonts are currently available. In any case, it is unrealistic to assume that future users of archived Ega resources will have access to appropriate software and expertise for interpreting or reverse engineering the character encodings correctly.

4.2. Approach

We addressed these problems by defining and implementing a mapping between the glyphs used in the Ega documentation and Unicode IPA code-points (Unicode Consortium, 2003). The correspondence table is expressed in

[3] See www.spectrum.uni-bielefeld.de/langdoc/EGA/, www.cs.mu.oz.au/research/lt/projects/ega/.

[4] See www.sil.org/computing/shoebox/.

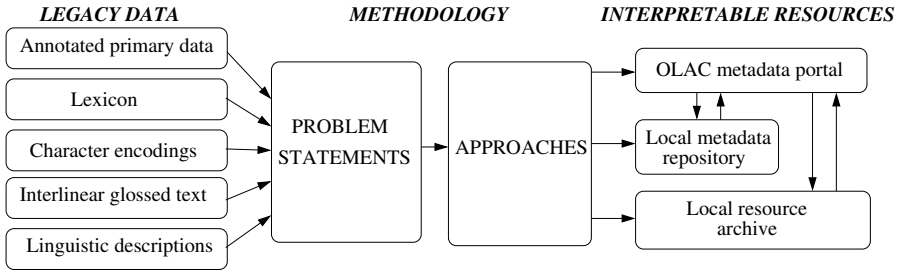

Figure 1: Securing interpretability: procedure.

XML, which allows comprehensive support for exotic character representation and encodings. The database is automatically augmented with the fields containing Unicode representations. This simple step greatly enhances the durability of the resource. Furthermore, by applying the mapping table to the data it is possible to identify unmapped characters and fill any gaps in the table. The table itself is archived as formal documentation of the legacy encodings, i.e. as fine-grained metadata.

Our procedure is specific to Shoebox at this point, but we claim that the approach is generalisable, though in many cases the solutions will be more complex. In the general case, we need to add an additional font analysis step: each font needs to be analysed on a per glyph basis and each glyph mapped to a code-point in Unicode. This second step is required regardless of whether or not a glyph is actually used in the documentation itself (in essence, this is securing the interpretability of the font structure itself, rather than the interpretability of specific characters, which is the focus of the present discussion).

5. Securing interpretability of interlinear glossed text

5.1. Problem statement

Interlinear text is a common presentation format for the expression of linguistic information. Although a range of specialised tools for creating and manipulating interlinear text are under discussion (Hughes et al., 2004), much legacy interlinear material is fundamentally unstructured in the sense that alignment of interlinear text is purely visual through use of spaces, tabs, line breaks and page breaks, rather than through use of well-defined text objects such as tables. It is a common experience that in format conversion of such documents, even with supposedly compatible applications, significant loss of pertinent alignment information occurs. In the case of Ega, only very basic instances of interlinear material exist, with phrases and free translations of the elicited responses of the West African Language Data Sheets, WALDS (Kropp-Dakubu, 1980). Very little detailed morphosyntactic markup is available. Consequently, deriving consistency and understanding of structural relations is at best difficult, and at worst, unsustainable over the longer term.

5.2. Approach

We addressed the problem by designing and implementing a translation of the Ega interlinear sources into the EMELD model for interlinear text, a four level XML based representation (Bow et al., 2003). This process has three distinct stages:

1. converting the original interlinear rendering into a tabular format,
2. converting the tabular format into a tree structure,
3. expressing this tree structure conventionally in XML.

In order to handle the specific typological properties of Ega, the 'morph' tier of the EMELD model was extended to include tiers for *lexical tone*, *morphosyntactic tone*, *morphosyntactic category* and *morphological paradigm*.

Additionally an expanded lexical database containing all morphs and glosses found in the texts was created for reference and consistency checking.

6. Securing interpretability of annotated recordings

6.1. Problem statement

The available Ega audio and video recordings are of questionnaire-based interviews, narratives and other interactions. These primary data are already digitally formatted on DAT and miniDV, and as conversions to WAV, MP3 and AVI formats on CD-ROM. In some conversions across signal data formats it is possible to lose information on *temporal resolution* (and consequently precision of alignment), and on *frequency resolution* and *spectral faithfulness* (especially by converting into lossy formats such as MP3). All of these potentially damage prospects of interpretability for future linguistic and phonetic analysis and use in computational applications such as information retrieval or speech and text technology. The files have been annotated by different annotators using a number of Open Source tools such as Praat, Transcriber, TASX, and the proprietary esps/waves+ ('XWaves'),[5] resulting in a variety of annotation formats and thus potentially different degrees of precision of alignment: some annotation formats use point-based single time-stamps, while others use interval-based time-stamp pairs; the former require additional conventions stating how they apply to intervals.

6.2. Approach

Securing the long-term interpretability of varied and proprietary binary audio, video and other phonetic signal formats on magnetic and optical media is a complex and

[5] See www.praat.org, www.etca.fr/CTA/gip/Projets/Transcriber, tasxforce.lili.uni-bielefeld.de, www.entropic.com/esps.html.

specialised task which is being addressed by engineers and archivists worldwide.[6] We did not address this task, except to prefer non-compressed data formats and to preserve the available temporal, frequency and amplitude resolutions. Our focus is on securing the annotations rather than the signal files themselves. For this purpose the generic TASX XML format was used, and tools for inter-converting between files and into TASX XML format was developed. Results were validated by reverse conversion and file comparison. Losses between existing formats only occurred with metadata (header information); all available metadata information was retained in the TASX format.

7. Securing interpretability of linguistic descriptions

7.1. Problem statement

A set of serious interpretability problems arises with morphosyntactic descriptions in general and those of Ega in particular:

1. The nomenclature of different linguistic traditions often has a common core (e.g. N = noun, V = verb, S = subject, O = object), but details vary greatly.
2. Different terminologies associated with different national languages compound the problem, as in the Ega case where grammatical descriptions are available in both French (Bole-Richard, 1982) and English (Gibbon et al., 2003).
3. There is currently little guidance or flexibility for the individual linguist in deciding which terms to adopt, how to express them in their data, or how to relate these terms to higher level cross-linguistic ontologies which allow for consistency of semantic content.

7.2. Approach

A number of proposals for practical morphosyntactic categorisation are available, notably the well-known EAGLES and EUROTYP taxonomies. We selected GOLD, the General Ontology for Linguistic Description (Farrar and Langendoen, 2003) as a newer source of consistent morphosyntactic terminology. The process of assessing the nomenclature of linguistic annotation terminology and creating the relevant mapping to a higher level ontology has several stages:

1. Assessment of the extant terminology within the language data set is gathered.
2. Consideration of idiosyncracies of the descriptive metalanguages (in the case of Ega, nomenclature from both English and French must be considered).
3. Consideration of the overall expressivity of the linguistic annotation terminology and comparison with the general categories enumerated in GOLD.

Where correspondences exist, we can utilise the GOLD namespace to express these concepts directly in the annotated Ega text. There remain a number of problems where gaps are found in GOLD.

[6] See www.dcs.shef.ac.uk/spandh/projects/swag/, www.iasa-web.org/index.htm.

8. Conclusion

We have addressed the problem of conserving legacy data from the new perspective of securing interpretability, describing a scenario in which five types of legacy resource are processed with restricted time and personnel. The goal of securing archive interpretability of these resources was achieved, though recently mooted crucial issues such as metadata consistency (Trippel et al., 2004) have not been considered here. The resulting archive in the Bielefeld Ega repository is accessible via the OLAC (Open Language Archive Community) metadata portal.[7] We suggest that our our approach is generalisable and particularly appropriate for efficient endangered language documentation.

9. References

Bird, Steven and Gary Simons, 2003. Seven dimensions of portability for language documentation and description. *Language*, 79:557–582.

Bole-Richard, Rémy, 1982. Esquisse de grammaire: L'Ega. In Georges Hérault (ed.), *Atlas des Langues Kwa, Tome 2*. Abidjan: ILA.

Bow, Catherine, Baden Hughes, and Steven Bird, 2003. Towards a general model of interlinear text. In *Proc. EMELD Conference 2003: Digitizing and Annotating Texts and Field Recordings*.

Farrar, Scott and Terry Langendoen, 2003. An ontology for linguistic concepts. *GLOT International*, 7(3):97–100.

Gibbon, Dafydd, Firmin Ahoua, and Sophie Salffner, 2003. Outline of Ega Morphosyntax. Technical report, U Bielefeld.

Hughes, Baden, Catherine Bow, and Steven Bird, 2004. Functional requirements for an interlinear text editor. In *Proc. LREC2004*. Paris: ELRA.

Kropp-Dakubu, Mary Esther, 1980. West African Language Data Sheets Vol. 1. Technical report, Accra: University of Legon & West African Linguistic Society.

Trippel, Thorsten, Dafydd Gibbon, and Felix Sasaki, 2004. Consistent storage of metadata in inference lexica: the metalex approach. In *Proc. LREC2004*. Paris: ELRA.

Unicode Consortium, 2003. *The Unicode Standard, Version 4.0*. Reading, MA: Addison-Wesley.

Yergeau, Francois, Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, and Eve Maler, 2003. *Extensible Markup Language (XML) 1.0*. Boston MA: World Wide Web Consortium, 3rd edition.

Acknowledgments

We are indebted to our Ega partners Lucien Bazey and Towe Cyprien Oko for language expertise, the late Eddy Gbéry for lexicographic documentation, Firmin Ahoua and Bruce Connell for fieldwork on and analyses of Ega, Sophie Salffner for Ega annotation, and Thorsten Trippel for format conversion tools. This research has been supported by National Science Foundation Grant Number 0094934 (Electronic Metastructure for Endangered Languages Data) and by the German Academic Exchange Service (DAAD) grant *M.A. in Documentation of Local Languages*.

[7] See www.language-archives.org/.