Resurrecting the Dead (Languages)

Documenting, archiving, and teaching at the Linguistics Research Center of the University of Texas

Hans C. Boas and Todd B. Krause
hcb@austin.utexas.edu, bobtodd@math.utexas.edu

Indo–European Lexicon

Documenting an entire language family presents some special challenges. In particular, it is important to highlight the unity of the family, showing what distinguishes the languages within the family from those outside. This is accomplished most compactly through an etymological dictionary, a collection of roots which, through well-defined phonological rules, provide shared ancestors for core vocabulary found among branches on the linguistic family tree. For the Indo–European family, the standard reference is [Pokorny].

Structure of the Lexicon

In essence, [Pokorny] is a numbered list of etyma, i.e. of vocabulary items reconstructed for the parent language Proto–Indo–European (PIE). Under each etymon, or root, the reader finds a list of the descendants, or reflexes, of that root in the various daughter languages. The LRC's website recapitulates this structure in a simple fashion:

- Each PIE root has a separate webpage
  - The URL actually contains the root's number in [Pokorny], providing the possibility of direct access
  - Alternately, a separate page provides a master list of all available roots for visual browsing
- On the page for each root, the user finds a list of all languages showing a reflex of that root
- Under each language, the page lists the major individual reflexes for that language

Semantic Indexing

Though specialists may already know which PIE root they seek, users from outside the subdiscipline may only have broad ideas of the type of word they are looking for. The LRC therefore provides a separate semantic index, allowing access to etyma based upon how they describe the world around us.

- Semantic categories derived from [Buck]
  - E.g. ANIMALS, DWELLINGS, EMOTIONS, BELIEFS
- Each general category has further subcategories
  - Subcategories under EMOTIONS include PASSION, SURPRISE, JOY, ANXIETY
- Under each subcategory the user finds relevant etyma

Overarching Goals

- **Document an Entire Language Family**
  - **Provide cohesion to a web of distinct languages**
  - **Represent the diversity of linguistic and cultural phenomena**
- **Provide Multiple Points of Entry**
  - **Lexical**
  - **Semantic**
  - **Structural**
- **Provide Resources for Specialists and Non–Specialists**

Early Indo–European OnLine (EIEOL)

Each Indo–European language, regardless of its relation to other languages, is a unique entity with its own distinct characteristics, historical evolution, and textual tradition. The Early Indo–European OnLine (EIEOL) collection seeks to highlight the diversity exhibited by the Indo–European family. EIEOL supplies structural overviews of major representatives from each branch of the linguistic family tree.

- Structural overviews comprise series of 5–10 lessons for each language
- Each lesson conforms to a prescribed format
  - Introduction: background information on culture, language, and text
  - Reading: a textual selection with word–by–word grammatical analysis and contextual gloss
  - Grammar: a 5–point explanation of important grammatical features

The EIEOL Format

All language scholars maintain their own preferences as to how best to make progress in learning a language. The EIEOL format strikes a balance between the needs of the specialist and of the non–specialist. A linguist within the subdiscipline requires an introduction to the features a given language shares with its relatives. The same specialist may also need a succinct overview as to what other disciplines, such as anthropology and history, have contributed to defining the context surrounding the language and its speakers.

Non–specialists, by contrast, may wish to avoid dedicating months to a university course. They may nevertheless need more than a superficial description of the language. The EIEOL format seeks to fulfill these needs in a format derived from World War II military training and designed to allow soldiers to begin translating intercepted messages in a minimum amount of time.

- Self–standing introductions to the languages
  - Written with non–specialists in mind
  - Introduce technical terminology and highlight linguistic issues as needed
  - Provide bibliographies pointing to sources for further study

Lesson Processing

The LRC generates webpages automatically based on author content. `Perl` scripts read the ASCII files, substitute the betacode transcription with Unicode characters, and insert HTML formatting. This provides three main benefits:

- Design is separated from content
  - Webpages can be redesigned without modifying content files
  - Authors need not learn HTML
- The system is lightweight
  - All content files are ASCII
  - All processing files are `Perl` scripts: i.e. ASCII files
- Data can be reorganized in different fashions
  - Processing takes glossed texts to create glossaries of all surface forms, as well as all headwords
  - English meanings are extracted to create English–to–(Target Language) glossaries

Specialists in Language, not HTML

The content of the various EIEOL language series derives from the work of specialists in the respective languages. The LRC has designed an infrastructure that requires minimal knowledge of computers and no knowledge of computer programming languages. This frees lesson writers to focus on the content, and it allows a minimal LRC staff to integrate that content seamlessly into the greater LRC ecosystem. This power derives from two basic features:

- All content is written in pure ASCII files
- Computer scripts automatically convert ASCII content to linked HTML pages

Lesson Writing

In order to minimize issues of fonts differing between computers, lesson writers create a betacode map. This is a file that describes how the lesson writer will represent a character from the indigenous script (say the *ash*: æ, a Unicode symbol) in terms of an ASCII sequence (perhaps a| on the keyboard). Wherever the author wishes an ash to appear, she writes a| in the content files. The glossed readings in each lesson require a separate file with special formatting. The LRC stipulates some guidelines:

- Betacode sequences appear in lesson content between paired @–signs: e.g. @a|@
- In glossed readings, each word is entered on a separate line with a rigid format

  surface_form @ analysis of <headword> meaning @ [contextual_translation]

References

- Buck, Carl Darling **A Dictionary of Selected Synonyms in the Principal Indo–European Languages** (1949)
- Pokorny, Julius **Indogermanisches etymologisches Wörterbuch** (1959)

Current Overviews