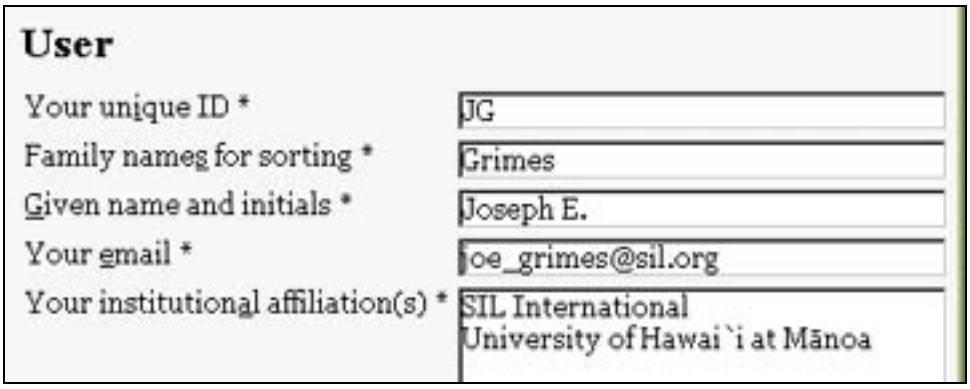
Designing Tools That Promote Archiving

Joseph E. Grimes
SIL International, University of Hawaii at Mānoa
EMELD 2006 Handout

(1) Tools for linguistics can be designed to promote consistent archiving, by making it obvious how and where to include the needed information.
(2) Not all linguistics tools cover the same range of metadata; some need to go beyond the metadata already defined by the Open Language Archives Consortium (OLAC).
(3) Experience with Wordcorr for comparative linguistics, design considerations for applications in sociolinguistics and lexicography.
(4) Three types of metadata in each: persons who collect and interpret the data, works they produce, and speech varieties they include.

Comparative Linguistics

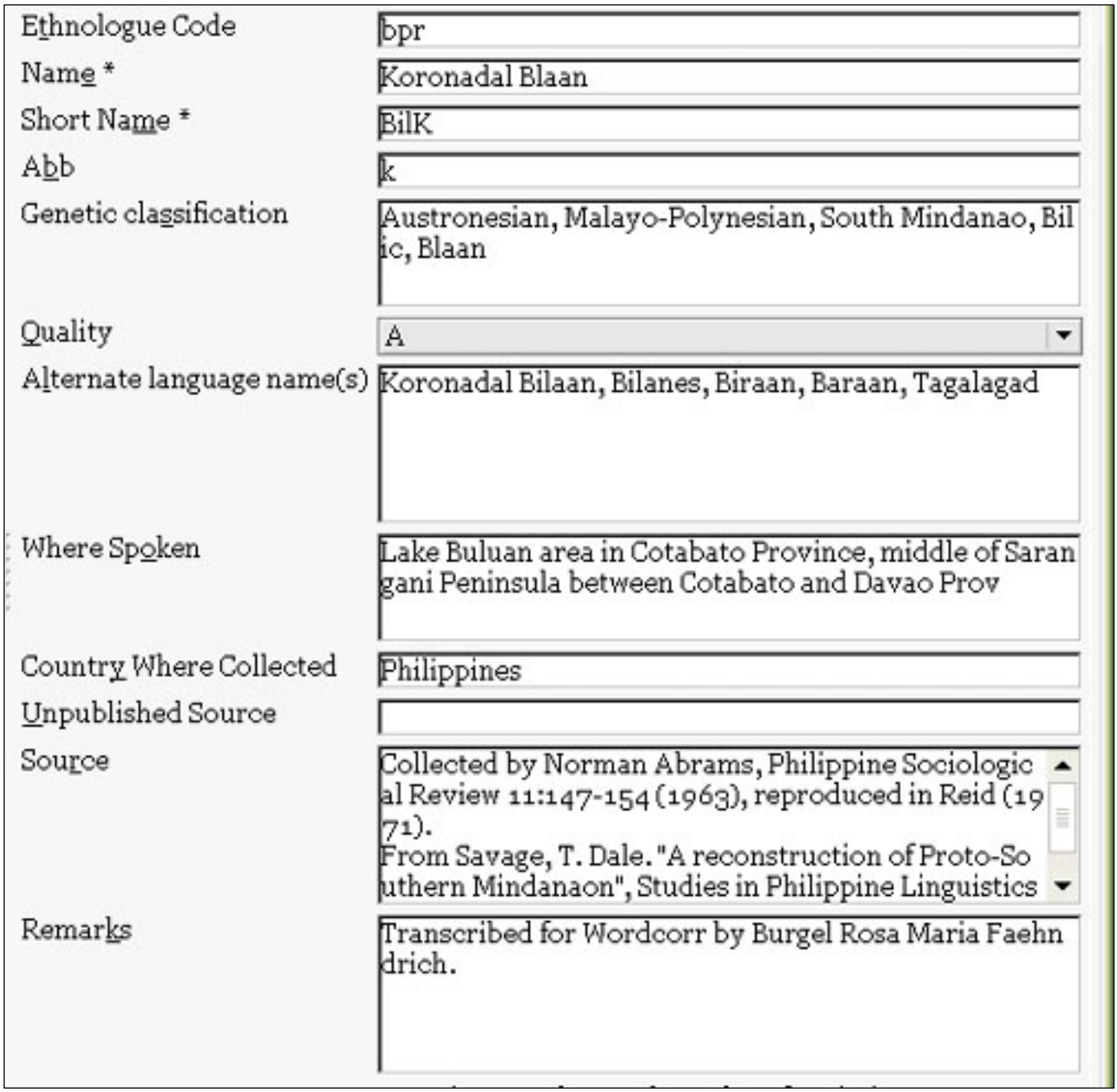
(5) Wordcorr, with 470 downloads, helps linguists apply the comparative method consistently to parallel word lists, without letting either data or analysis go astray.
(6) User metadata, a little bit less than is required for an article in Language:



(7) Collection metadata, similar to library information for an article or monograph:

Collection Title *	JG-Mindanao
Short Title *	Mind5
Your Role as Creator *	annotator
Contributor's ID or Name	Maria Faehndrich
Primary Gloss Language *	English
Its language code *	eng
Secondary Gloss Language	
Its language code	
Keywords for searching	Bilic, Mindanao, Southern Mindanaon
Description *	The three Bilic varieties from the southern tip of Mindanao and adjacent islands, plus one Subanen and one Manobo variety
Remarks	Five speech varieties of Mindanao, Philippines, transcribed using IPA notation, as a test data set for Wordcorr. Compiled for Wordcorr by Joseph E. Grimes and Maria Faehndrich. Order of entries follows Savage. Entries 1-27 have been annotated and tabulated in the Contemporary view by
Published Source(s)	T. Dale Savage. 1986. A reconstruction of Proto-Southern-Mindanaon. Studies in Philippine Linguistics 6:2.181-223. Lawrence A. Reid, editor, Philippine Minor Languages: Word Lists and P
Geographic Area Covered	Mindanao, two offshore islands in the south.
Stable Copy Located At	
Rights Management	Open Publishing
Year Copyright Asserted	
Creator	Joseph E. Grimes [JG]
Publisher	joe_grimes@sil.org

(8) Variety metadata, similar to a reduced Ethnologue entry for each speech variety in the collection, plus provenance of field data and other things relevant to comparativists but not yet in the OLAC schema:



(9) Wordcorr transforms its metadata into OLAC form for incorporation into Linguist List's OLAC repository. Not released yet.

Sociolinguistics

(10) Multilingual situations can be assessed using information on how proficient different segments of a community are. One test[1] is based on the observation that you have to

[1] Carla F. Radloff, *Sentence repetition testing for studies of community bilingualism.* Dallas: Summer Institute of Linguistics and University of Texas at Arlington, 1991.

know a second language quite well in order to repeat whole sentences in it immediately. The sentence repetition test discriminates lower degrees of proficiency well, higher degrees poorly.

(11) The test is easy to administer, hard to set up. One computational tool from the 1980s set it up correctly, but is completely user unfriendly.[2] So a redesign is needed with user interfaces that capture both public and internal metadata for

- Test designer and team for a particular L2[3] test
- Native speakers of the L2 as talkers and testers for candidate sentences
- Test protocols from the L2 speakers for every candidate sentence in the calibration, scored by trained testers such as the design team
- Test subjects for the calibrations whose proficiency in L2 has been calibrated independently for validation, such as by Round Table tests[4]
- Calibration results for every candidate sentence
- Equivalent sets of test sentences and the PDAs on which each is installed
- Test administrators trained to score the test, each using a separate PDA
- The sample of L2 speakers being tested: no names or addresses, but serial number, location, which test, administered by, using PDA, position in societal model space …
- Test protocols for every field test, collated from multiple PDAs
- Several kinds of summary results.

Lexicography

(12) Just beginning the design stage, a Web-based tool for investigating endangered and underdocumented natural languages by producing theoretically coherent dictionaries combining the insights of the Meaning-Text[5] and Natural Semantic Metalanguage[6] approaches. It will probably use a factory design pattern to accommodate diverse structures:

- Alphabetic vs. semantic arrangement of entries
- Internal structuring of entries by sense vs. by part of speech
- Policy for use of subentries
- Unforeseen structures.

(13) Different granularity is also needed for different presentations:

- The same example, or fragments of it, illustrate a number of entries
- The example source may or may not be included in every presentation; it might be cited in a reference dictionary, but not in a school dictionary or on the Web
- Multivalent lexical function values, linked in both directions, may have different levels of detail in different presentations
- The usual metadata for creator, collaborators, library search, subject language, language of description (if different), and possibly the protolanguage if reconstructions or known etyma are included
- Metadata for different presentation options need to be incorporated in the master document

---

[2] Joseph E. Grimes, *Language survey reference guide.* Dallas: Summer Institute of Linguistics, 1995. Pp. 46-59, also in the LinguaLinks Library.

[3] L2 is the standard abbreviation for "second language," which may actually be a person's third or fourth or … language.

[4] Interagency Round Table on Foreign Languages, begun by the Foreign Service Institute of the U.S. State Department. See Grimes 1995, pp. 34-45.

[5] Igor' Mel'čuk et al. Dictionnaire explicatif et combinatoire du français contemporain. Montréal: Les Presses de l'Université de Montréal, 1984 et seq.

[6] Cliff Goddard, Semantic analysis: A practical introduction. Oxford: Oxford University Press, 1984.

- If the dictionary covers more than one speech variety, the same variety metadata needed for comparative linguistics are needed as well.