Data Warehouse, Bronze, Gold, STEC, Software

Doug Cooper
Center for Research in Computational Linguistics
doug.cooper.thailand@gmail.com

Abstract

We are building an analytical *data warehouse* for linguistic data – primarily lexicons and phonological data – for languages in the Asia-Pacific region. This paper briefly outlines the project, making the point that the need for improved technology for endangered and low-density language data extends well beyond completion of fieldwork. We suggest that *shared task evaluation challenges* (STECs) are an appropriate model to follow for creating this technology, and that stocking data warehouses with clean *bronze-standard* data and baseline tools – no mean task – is an effective way to elicit the broad collaboration from linguists and computer scientists needed to create the *gold-standard* data that STECs require.

1 Introduction

The call for this workshop mentions the first step of the language documentation process, pointing out that the promise of new technology in documenting endangered languages remains unfulfilled, particularly in the context of modern recording technologies.

But lack of tools extends far beyond this first step. It encompasses the accessibility of data long since gathered and (usually, but not always) published, as well as applications *for* the data by its most voracious consumer: the study of comparative and historical linguistics.

We encounter these problems daily in preliminary development of data and software resources for a planned *Asia-Pacific Linguistic Data Warehouse*. Briefly, our initial focus is on five phyla (~2,000 languages): Austroasiatic, Austronesian, Hmong-Mien, Kra-Dai, and Sino-Tibetan, which form a Southeast Asian convergence area, and individually extend well into China, India, the Himalayas, and the Pacific. Data for languages of Australia and New Guinea will follow.

Not all of these languages are endangered, but many are; not all are low-density, but most are.

Our data are preferentially drawn from the sort of lexicography gathered for comparative purposes (ideally 2,500 items per language), and the phonological, semantic, and phylogenetic data that can be found for, or inferred from, them. These are the only kind of data for which we are likely to find near-complete language representation. We include smaller lexicons when necessary, and intra-language dialect surveys when available. All available metadata are incorporated, including typological and phonotactic features, (phylogenetic) character sets, geo-physical and demographic data, details of lexicon coverage, extent, or quality, and bibliographic or source data.

Such data are not always easily found. Their delivery packages – primarily books and journals – may be discoverable via bibliographic metadata, but details of the datasets themselves are not. As a result, traditional bibliographic documentation, accessed via portals like OLAC (Simons and Bird, 2000) and Glottolog **(**Nordhoff an**d** Hammarström, 2011), tends to have low recall and precision in regard to data resource discovery.

Our experience in acquiring and performing methodical *data audits* of large quantities of published and unpublished materials reveals sets of lexical, grammatical, phonological, corpus, and other materials that are regular enough in form, and extensive enough in content, to comprise aggregable *linguistic data supersets* for the Asia-Pacific region.

These ongoing data audits take a three-tiered approach, separately documenting texts (to enable source recovery), their abstract data content (to enable high-recall *resource discovery*), and any concrete, transcribed data instances (to enable high-precision *data aggregation*).

Discovery and aggregation only open the door. Many datasets are hand-crafted for a researcher's specific needs and interests, even if they fall into larger research categories. Yet far from having reliable algorithms for central concerns (such as proto-language reconstruction, or subgrouping of linguistic phyla in family trees or networks) the field has not yet had to grapple with basic problems – such as normalizing phonological transcription or gloss semantics, or accurately assembling large-scale cognate sets – that will be

Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 91–99,
Baltimore, Maryland, USA, 26 June 2014.

presented by datasets that include millions of data items for thousands of languages, and many more thousands of dialectal variants.

The central issue we face is the gap between:

- the results of published and unpublished fieldwork, and
- their usability in downstream research and reference applications.

In some cases this gap is painfully obvious – as in the backlog of carefully elicited wordlists still awaiting phonetic transcription. In others, the gap becomes evident when we begin to assemble large comparable datasets from published data; deceptively difficult, and never accomplished for collections broader than a single language family, or larger than about 200 words per language. Such tasks are still basically hand work; often requiring the specialized knowledge of the field researcher.

1.1 Data life cycle: anticipate or participate

We see the need for tools as part of a new sort of *data life cycle management* that extends the concerns of content, format, discovery, access, citation, preservation, and rights as usually articulated, notably in Bird and Simons (2003).

Simply put, producing publishable or "correct" results is not sufficient to guarantee the downstream usability of data. Rather, data must undergo a series of transformations as it travels from one research specialty to the next. We hope there will be an increasing expectation that the data producer either anticipate or participate in this process.

At one end of the cycle, this often requires small, specialized datasets of the sort needed to support software development for tasks like automated transcription or phonemic analysis – still open problems in the context of under-resourced languages.

At the other, building massive datasets that are suitable for improving and extending quantitative comparative linguistic applications – or discovering the scales at which different methods might be most useful – has not been a priority for the linguistics community: if a few representative items demonstrate a relationship or support a reconstruction convincingly, then exhaustive coverage does not make the argument stronger.

We face a classic resource deadlock. High-quality "last-user" datasets are not constructed because traditional methods are too expensive and time-consuming. However, tools for refining "first-producer" data on an industrial scale are not built because the high-quality datasets needed to validate them do not exist. Development of computational methods for problems like subgrouping tends to focus on a small number of available datasets, while their results are criticized for precisely this.

2 STECs and gold-standard data

Log jams in natural language processing are nothing new. A *shared task evaluation challenge* (STEC) presents an open challenge to the field in the context of evaluating performance on a specific task. Originally developed in the context of the TIPSTER Text Program (which initiated the long-running MUC and TREC conference series) as discussed in Belz and Kilgarriff (2006), see also Hirschmann (1998) "*Over the past twenty years, virtually every field of research in human language technology (HLT) has introduced STECS.*"

The STEC is the culmination of a series of efforts intended to focus and advance progress by asking such questions as:

- what problems need to be solved in order to advance the field? Where are we trying to go, and what is standing in our way?
- what kinds of necessary data are not generally available? What kinds of datasets are too difficult for individual researchers to create?
- what kind of functional decomposition into simpler goals will help demonstrates and measure progress in quantitative and qualitative terms?

Both data, and evaluation metrics, are made available well before the STEC, which is often held in conjunction with a major conference. The task is typically initiated by the release of a dataset; results are submitted by some deadline, and the results of evaluation are announced before or at the conference.

The terms *gold-standard* and more recently, *silver-standard* (for machine-generated sets) are used to describe datasets created for use in STECs and NLP applications. These can be thought of as being "correct answers" for quantitative evaluation (Kilgarriff 1998).

Gold-standard datasets are built to enable comparable evaluation of alternative algorithms or implementations. Frequently, part of the set will be publicly released in advance to serve as training data, while part of it is held back to provide test data (and is released at a later date).

Gold-standard datasets reflect the state of the art in an area, such as the specification of word senses, delineation of word boundaries, or evaluation of message sentiment, for which there may not be any purely objective ground truth. We can reasonably expect to allow alternative formulations of gold-standard sets in areas in which the state of the art may be uncertain, even in the eyes of experts. And we can anticipate increased critical scrutiny of previously accepted judgments as more base data and better investigative tools become available; see e.g. Round (2013, 2014).

2.1 STECs for low-density languages

In our opinion, all of the reasons for which STECs are devised and gold-standard datasets defined apply equally to the low-density language problems we touched on in Section 1. These include:

- normalization and syllabification of transcribed data,
- phonetic transcription of audio and orthographic data,
- morphemic analysis of transcribed data,
- extraction of a phonemic analysis from phonetic data,
- identification of internal cognates and/or derivationally related forms, as well as loan-word identification and stratification,
- automated reglossing / translation (to a standardized gloss set) of glosses and/or definitions.
- automated inference of phylogenetic subgrouping.
- automated generation of proto-forms,

All are characterized by the same requirement for human judgment in processing, and lack of absolute certainty as to outcomes.

The critical difference is that (as far as we know) STECs in NLP invariably focus on high-density languages for which both data and expertise are readily available. In contrast, low-density languages – which presumably includes the entire range of endangered languages – are by their nature specialty realms, for which expertise, even within a single phylum, is often widely dispersed.

Thus, the problem we face in creating successful STECs for documentary linguistics is not simply a matter of thinking up tasks, and relying on in-house expertise to develop gold-standard datasets. Rather, advancing development of computational tools requires participation from a large community of independently working linguists as well.

3 Cast bronze to net gold

Our approach to achieving this begins by laying the groundwork for collaboration between:

- computer scientists who recognize the need for better data, and will join the challenge of solving practical problems in building massive, comparable datasets, and
- linguists willing to help create and validate the gold-standard reference sets and training data needed to establish quality metrics for improving software tools.

We think this collaboration is best motivated in the old-fashioned way: reduce participants'

vapor	no data could be located (useful when documenting data availability by ISO code)
water	untranscribed audio recording only
paper	print/image/PDF data are in hand, but not transcribed or extracted
tin	raw e-orthography and definitions (as in typical documentary dictionaries)
copper	raw e-forms and glosses (as in purpose-collected comparative lexicons; e.g. Holle lists)
bronze	clean electronic data and metadata, ready for hand or machine processing, naive normalization of forms and glosses, cognate sets partially specified, capable of demonstrating preliminary data warehouse functionality (**Software**: baseline vanilla algorithms)
silver	machine-normalized or grouped data, not yet verified by humans (**Software:** better than baseline)
gold	human-verified/accepted, machine-usable comparable datasets (**Software:** (best) able to produce gold-standard results)

Table 1. Data quality standards re lexicons, cognate sets, reconstructions, and subgrouping, with parallels to software tools. **Silver-** and **gold-standard** are the only terms commonly used in this context.

startup costs, flatten their learning curves, highlight expected outcomes that will advance collaborators' self-interests, and help provide the data, tools, and/or metrics that collaborators will need to seek funding themselves.

This in itself as a long-term effort – easily 5–8 years for our region, with optimal funding – whose thrust can be summarized as **cast bronze to net gold** (see **Table 1**).

Locating data, and bringing it to the minimal state required for computer applications requires a massive amount of work. Consider just the discovery aspect, for which the data audit mentioned earlier entails an ongoing, two-pronged effort.

On one hand, we identify potential data content by acquiring as much published and unpublished print material as possible, including complete journal runs, monograph series, informally published "gray literature," extensive sets of unpublished field notes, and regular publication backlists (notably, a half-century of works from *Pacific Linguistics*, which will be added to our on-line repository later this year).[1]

On the other, we systematically work through the complete ISO 639-3 inventory (as a proxy for the on-the ground truth, and as a means of helping to perfect the standard, as well as identifying documentary shortfalls that might be short-listed for fieldwork) of our region, attempting to find at least lexical content for every language.

Overall, our summary project development plan has four steps, which relate to content and scale, and determined our choice of a regional focus – for which we could take responsibility – rather than either working at greater depth on a single phylum, or attempting to build a global framework, and then relying primarily on outside contributors.

First, define an area that is broad enough to be of wide linguistic interest, and able to supply a range of control and alternative test conditions for both traditional and computational methods. Even allowing for typological variation that may be found in individual phyla, we think this usually requires a regional perspective.

[1] For New Guinea, this required a special sub-project dubbed *INGA*, dedicated to tracking down "invisible" New Guinea archives held in libraries and file cabinets around the world! As implied, when possible we negotiate rights to scan and make all materials freely available in an on-line repository, and will begin to register DOI names (when appropriate) for texts and data this year.

Second, locate and prepare raw data of sufficient breadth and depth. We think that aiming for blanket rather than selective coverage is appropriate – it enables the broadest range of research agendas by reflecting the natural state of human migration and constant language contact.

Third, establish research goals that capture the interest of both fields – documentary / comparative / historical linguistics and computer science. This extends the argument for complete regional coverage, especially in convergence areas. But it also argues for *limiting* scope to an area in which it is realistically possible to actively recruit involvement, conference by conference.

Finally, we need to lower barriers to participation We think we can do this by providing a framework that allows data owners to take advantage of existing software tools, and which provides software developers with easily customized data test beds – the analytical *data warehouse*.

4 The data warehouse

A *data warehouse* is an integrated collection of databases that incorporates tools for sampling, analyzing, and visualizing query results. Unlike repository databases intended for storage and retrieval of prepared values (perhaps for off-line processing), data warehouses assume that data filtering, transformation, and analysis are essential to satisfying every query. In the context of comparative lexicons, such tasks are well beyond the scope of existing *virtual research environments* such as WebLicht (Hinrichs et al 2010) and TextGrid (Neuroth et al 2011), which focus primarily on text corpora.

Because sampling filters allow selection of homogeneous or representative subsamples, we can be as inclusive as possible in regard to data acquisition. We are not talking about data quality; rather (working within our overall criterion of comparative lexical data) we want to avoid excluding sets because of concerns about dataset size or content disparity, or over-representation of dialect survey data.

Many operations we wish to perform on or with data involve open research questions. Although users may perceive the warehouse as providing access to tools, we intend to present it to tool developers as a tunable test bed of data that does not require them to deal with data management, as well as a means of using, and encouraging development of, open-source toolkits such as the pioneering work of Kleiweg (2009)

and List and Moran (2013). We return to the idea of plug-and-play operations on lexicons in Section 6.

The warehouse also helps provide added value to potential data contributors. Even if software is freely available, preparing data or setting up tools can impose substantial, even insurmountable, burdens on data creators, particularly in regions in which cooperation between linguists and computer scientists is less common than in the US or Europe.

4.1 Data warehouse query-flow

In our test warehouse implementation, functionality is divided as follows:

- *filter*: define a *search universe* based on phylogenetic or phonotactic properties, geophysical or proximal location, lexicon characteristics, or other data or metadata features.
- *frame*: specify data and/or metadata to be returned, e.g. specific aspects of the form and/or gloss, or metadata details that might be useful for correlation testing.
- *analyze*: extract phone inventories, calculate functional load, investigate lexical neighborhoods, cluster data by phonological similarity, etc.
- *visualize*: provide alternatives to tables as appropriate, e.g. tree/graph/map layouts.
- *recycle*: search within returned data, use faceting to extend searches, or let the visualization serve as a chooser for a new search.

For brevity we discuss just one feature: filtering. This lets the search universe be defined in as much detail as possible, and is partly common sense: our overall data universe is decidedly lumpy due to the decision to include small samples (some <100 items) when necessary, and dialect surveys (perhaps with only minor differences between doculects) when possible.

It is also intended to take advantage of the large quantities of available metadata, whether it is *explicit* / *external* – that is, related to the language or doculect, or *implicit* / *internal*, i.e. can be derived from individual datasets or samples.

Such metadata includes proposed phylogenetic relations, typological features, geophysical and demographic data, characteristics of lexicon composition, extent, or quality, bibliographic or source data, and phonological properties of the doculect itself. Some of this metadata may be returned with individual items as part of the *data frame*.

Filter targets may be specified if appropriate. For example, a filter might limit a search to languages that contain sesquisyllables, or instead require that returned *items* be sesquisyllabic.

4.2 An example query and result

Figure 1 shows the result of a relatively simple warehouse query (using our unreleased exploratory implementation): a *geo-constrained phylogenetic tree* for Trans-New Guinea languages. Tree topology follows Ethnologue 16 (Lewis, 2009) as provided by the MultiTree project (Aristar and Ratliff, 2005); other analyses are

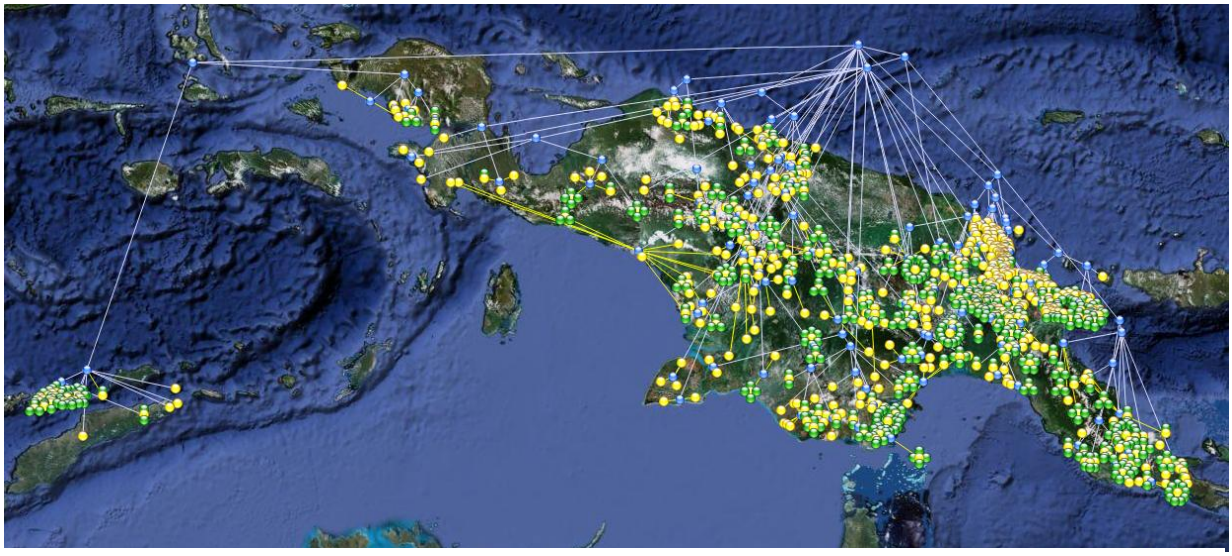

Figure 1 A geo-constrained phylogenetic tree (analysis by Ethnologue via MultiTree). This *cluster tree* keeps low-level group nodes near their daughters, but raises the root nodes. Dialects are green, languages yellow, and groups blue

Figure 2 A search for "bone" in ABVD Austronesian data (again, relations by Ethnologue via MultiTree), constrained to locations in Indonesia, and projected onto a map

readily specified. In this example dialects (from the same sources) are arranged in a circular pattern around the ISO 639-3 hub language (and, again, other analyses could be used instead). The same filtering and visualization routines are used in a different manner in **Figure 2**, which shows words for "bone" in Austronesian languages as provided by ABVD (Greenhill et al, 2008).

5 Data comparability and reusability

We will finish the discussion of data warehouses with a quick look at data comparability and reuse. Comparability or equivalence of datasets can be looked at in two ways

- at the *content* level, e.g. to ensure that the same systems of transcription and glossing are used for all datasets, and
- at the *structural* level, in identifying datasets of comparable complexity, structure, or available detail.

At the content level, normalization of forms and glosses is the critical transformation in the journey to gold-standard quality. We will briefly describe our systems for normalization, ***Metagloss*** and ***Metaphon***, because they are ripe for computational assistance. The discussion ends with a quick introduction to ***Etyset***, the framework we intend to use to describe and distribute structured datasets, such as those that incorporate subgroup and cognate detail.

5.1 Gloss, Metagloss, Etygloss

In most of our applications, a *gloss* is semantic annotation provided by the wordlist author in order to index phonological forms. Unfortunately, these may be elicitation terms rather than glosses (*green*? "grue." *blue*? "grue"), or local vernacular rather than common or scientific terms for flora and fauna. Phrasing varies wildly, and proper reading may depend on having the list context available (*short/tall, short/long*). Translation may be lossy (*strew* or *scatter* as nouns) due to differences in grammaticalization or lexification. All of these undermine comparability.

We have begun to define an intermediate, standardized ***metagloss*** layer to express the author's intent (if discernable). A third layer, the ***etygloss***, will help account for semantic shift in labeling cognate groups; i.e. glossing empty placeholders for proto-language reconstructions. In the simple case all three layers are identical.

Metagloss provides a controlled vocabulary for re-annotating or translating existing lexicon glosses; it foregrounds the critical design link between glossing and searching. We map this to *WordNet* senses, creating a low-overhead tool for word-sense disambiguation and facet generation.

The Metagloss controlled vocabulary can be extended; it uses attributes to specify predictable relationships (*sheep:male:castrated* for *wether*) and solve lexicalization problems that arise in gloss translation (e.g. *n@strew* is the noun form of *strew*). Additionally, it allows definition of *lightweight ontologies*; relations between Metaglosses that clarify semantic relations and improve search fallback performance.

5.2 Phon, Metaphon, Etyphon

Phonological forms present similarly difficult search problems; these go beyond easily fixed notational convention. For example, absence of marked syllable boundaries can make phonological searches difficult when we are interested in the phoneme's *role* (such as pre-nasalization) rather than its *sign* (/n/, /m/ etc.).

The same holds true for other context-sensitive symbols (e.g. "h" as /h/, /ʰ/, or as a prepended indicator of unvoiced phonemes). A greater problem arises from parsimonious notations that rely on commentary to clarify unwritten content, e.g. predictable vowel insertion – these must be made explicit.

We define an intermediate layer of standardized notation called ***metaphon***: a conventional notation that allows consistent search, while clearly documenting (and minimizing, in comparison to wild-card searches) the scope of any unavoidable approximation. A third layer, the ***etyphon***, allows temporary specification of a (possibly sub-lexical) phonemic rendition prior to any formal reconstruction.

Metaphon, like metagloss, is intimately tied to search functionality. Normalized transcription enables consistent extraction of phonological and phonotactic data. It lets the search universe be restricted to languages (or items) that have particular phonemes or features. This dynamic, data-driven process lets us weigh relative significance – frequency, salience, functional load – of features in sets that are themselves results drawn from a restricted search universe; e.g. to consider the functional load of tones in sesquisyllables.

5.3 Structural comparability: EtySet

The discussion thus far has focused on the form and quality of data *items*. We are equally concerned with what might be called structural comparability of data *sets*, because this determine the approach we take to systematic description, dissemination, and re-use of cognate sets, phylogenetic trees, or sets of proto-form reconstructions.

This has nothing to do with tagging or interchange standards, which can be handled with borrowed schemes designed for similar purposes, e.g. Newick notation (Felsenstein, 1986) or successors (Nakhleh, 2003). Rather, we require nomenclature that might be used to describe their contents, or to enable identification of sets of comparable complexity, structure, or detail.

We think such comparison is crucial to help research in quantitative historical linguistics move beyond its current state, which many linguists view as interesting but nevertheless ad hoc experimentation. In other words, we would like to see computational approaches to cognate identification, subgrouping, and proto-language reconstruction be developed and tested in environments for which the controlled variable is *linguistic typology*, with as many other factors as possible held equal.

Similarly, we would like to be able to vary starting conditions. For example Bouchard-Côté et al (2013) report on a computational approach to reconstruction given (assumed) prior knowledge of subgrouping in Austronesian. However, any one or two variables from amongst cognate grouping, reconstruction, and phylogenetic subgrouping may be used to test approaches to inferring or generating the third.

We refer to cognate sets, phylogenetic trees, and reconstructed proto-forms as *etysets*. The key terms of our working descriptive nomenclature are outlined in **Table 2**.

Etysets may be *bare* (links only), or *supported* by reconstructed forms or semantics; note that the phylogenetic analyses provided by Ethnologue, Glottolog, or MultiTree may be represent with bare etysets. An internal cognate etyset has depth (number of internal sets) and size (number of forms in each set). A regular cognate etyset has depth (the number of sets / implicit number of root proto-forms) and breadth (the number of lects represented in each cognate set).

For example a *bare cognate etyset o*f *Bahnaric, breadth Eth:80% / depth MSEA:90% depth* includes data from 32 (of 40. according to the Ethnologue analysis) Bahnaric languages, and at least 450 of the 500-odd terms in the MSEA (SIL 2002) elicitation list. Cognate groupings are provided, but not reconstructions or etyglosses.

breadth	number of nodes or leaves at any level of a phylogenetic tree.
depth	number of branch levels supplied.
degree	branchy-ness – the number of branches / degree of diversity at a given node.
density	a joint measure of breadth, depth, and degree.
size	# of cited or reconstructed forms associated with a leaf or branch node.
coverage	describes the extent of an etyset in terms of a fixed reference inventory.
phylogenetic etyset	described in term of breadth, depth, degree, and size.
documented node	includes metadata for approximate time depth and geographic location.
cognate etyset	may be *internal* or *regular*, and contains *internal* or *regular cognate sets.*

Table 2. Outline of the *EtySet* descriptive vocabulary.

6 Operations on lexicons

We end with a brief note about computational tasks for and by a data warehouse that is:

- stocked primarily with lexical, phonological, and phylogenetic data and relevant metadata,
- intended to support research in comparative and historical linguistics.

These fall under the general heading of *operations on lexicons*. We do not draw a strict dividing line between software employed to prepare data for use *in* a warehouse, and software used *by* the warehouse. We do exclude operations whose implementation is likely to be closely tied to a particular database implementation.

All would benefit from being implemented as plug-and-play functions, requiring some, but not excessive, programmer effort. This:

- allows head-to-head comparison of alternative algorithms, implementations, or interpretations of how measurements or actions should be carried out,
- allows encapsulation and offloading of computationally expensive algorithms; this is an important issue for some quantitative or statistical comparative methods, and
- encourages re-use of code in building new, alternative platforms for linguistic research.

We assume that all of these can be specified in terms of functionality, required data inputs, and expected data outputs, sticking to a Unix-like model in which data can be minimally formatted plain-text streams which, with the assistance of tabs, parentheses, and newlines, can be interpreted as bags, lists, vectors, matrices, trees, and the like. Higher-level streams (JSON, XML, RDF, HTML) are also reasonable outputs.

For brevity's sake, we limit examples to operations on phonological forms. We could easily list similar sets of operations – some straightforward, some not – on morphology, semantics, alternatives for visualization, cognate identification, phylogenetic subgrouping, proto-form generation, and the like.

Operations on phonological strings / lists

Conversion and markup of transcription

- between standardized and/or special-purpose notations,
- to novel notations, e.g. gestural scores,
- unambiguous conversion of notation from historical (e.g. Americanist) to IPA,
- potentially ambiguous normalization (e.g. interpretation of /h/),
- phonetic to phonemic conversion,
- marking of syllable boundaries,
- marking of syllable-internal features (e.g. onset, nucleus, coda),
- marking of morpheme boundaries.

Extraction / recognition of phonological features

- sonority sequence tagging.
- extraction/recognition of phones, phonation, co-articulatory, suprasegmental features,
- count/extraction of phone/feature n-grams,
- extraction or identification of arbitrary collocational features (e.g. sesquisyllable+tone),

Calculation of distance/similarity measures between strings, lists, and vectors

- weighted and unweighted edit distances,
- substring matching measures,
- vector cosine distance,
- phonologically based distance/similarity,
- language-internal distance/similarity,
- information content distance/similarity.

Clustering

- subgrouping list contents,
- "sounds like..." search (for very large sets).

Neighborhood measures

- generation of phonological neighborhoods,
- identification of neighbors,
- calculation of neighborhood size, density, clustering coefficients.

Load measures

- calculation of functional load of phonemes, features, collocations,
- calculation of salience of phonemes, features, collocations,
- use in pseudo-word generation.

7 Conclusion

The call for this workshop foregrounds development of software to aid in initial documentation of endangered languages, seeks models for collection and management of endangered-language data, and means of encouraging productive interaction between documentary linguists and computer scientists.

We suggest that these same needs exist all down the line, encompassing low-resource languages in general, documentation long-since completed, and analytical applications far removed from fieldwork settings. We propose that addressing them in downstream environments, such as data warehouses and STECs, may be an effective way to meet our common "preeminent grand challenge:" integration of linguistic theories and analyses, relying on massive scaling up of datasets and new computational methods, as articulated by Bender and Good (2010).

References

Anthony Aristar and Martha Ratliff. 2005. *MultiTree: A digital library of language relationships*. Institute for Language Information and Technology: Ypsilanti, MI. http://multitree.org.

Anja Belz and Adam Kilgarriff. 2006. *Shared-task evaluations in HLT: Lessons for NLG*. In Proceedings of INLG-2006.

Emily Bender and Jeff Good. 2010. *A Grand Challenge for Linguistics: Scaling Up and Integrating Models*. White paper contributed to NSF SBE 2020 initiative. http://www.nsf.gov/sbe/sbe_2020/-2020_pdfs/Bender_Emily_81.pdf

Steven Bird and Gary Simons. 2003. *Seven Dimensions of Portability for Language Documentation and Description*. Language 79:2003, 557-5822.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths and Dan Klein. 2013. *Automated reconstruction of ancient languages using probabilistic models of sound change.* Proceedings of the National Academy of Sciences. http://-www.pnas.org/content/110/11/4224

Joseph Felsenstein. 1986. *The newick tree format*. http://evolution.genetics.washington.edu/phylip/newicktree.html

Simon Greenhill, Robert Blust, and Russell D.. Gray. 2008. *The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics.* Evolutionary Bioinformatics, 4:271-283. http://-language.psy.auckland.ac.nz/austronesian

Erhard W. Hinrichs, Marie Hinrichs and Thomas Zastrow. 2010. WebLicht: Web-Based LRT Services for German. In: *Proceedings of the ACL 2010 System Demonstrations*. pages 25–29.

Lynette Hirschman. 1998. *The evolution of evaluation: Lessons from the Message Understanding Conferences*. Computer Speech and Language, 12:283–285.

Adam Kilgarriff. 1998. *Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs*. Computer Speech and Language, 12 (3) Special Issue on Evaluation of Speech and Language Technology, edited by Robert Gaizauskas. 453-472. http://www.kilgarriff.co.uk/Publications/1998-K-CompSL.pdf For TREC see http://trec.nist.gov. The TIPSTER site has been preserved here: http://www.nist.gov/itl/div894/894.02/related_projects/tipster/

Peter Kleiweg. 2006. *RuG/L^{04} Software for dialectometrics and cartography*. Rijksuniversiteit Groningen. Faculteit der Letteren. http://www.let.rug.nl-/kleiweg/L04/

M. Paul Lewis. 2009. *Ethnologue: Languages of the World, Sixteenth Edition*. SIL International, Dallas, Texas.

Johann-Mattis List and Steven Moran. 2013. *An Open-Source Toolkit for Quantitative Historical Linguistics*. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 13–18, Sofia, Bulgaria. *http://-*www.zora.uzh.ch/84667/1/P13-4003.pdf

Luay Nakhleh, Daniel Miranker, and Francois Barbancon. 2003. *Requirements of Phylogenetic Databases*. In Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03). 141-148. IEEE Press. http://www.cs.rice.edu/~nakhleh/Papers/-bibe03_final.pdf

Heike Neuroth, Felix Lohmeier, Kathleen Marie Smith: *TextGrid. Virtual Research Environment for the Humanities.* In: *The International Journal of Digital Curation.* 6, Nr. 2, 2011, S. 222–231.

Sebastian Nordhoff and Harald Hammarström. 2011. *Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources*. 783 CEUR Workshop, Proceedings of the First International Workshop on Linked Science 2011 http://iswc2011.semanticweb.org/fileadmin/iswc/-Papers/Workshops/LISC/nordhoff.pdf

Erich R. Round. 2013. *'Big data' typology and linguistic phylogenetics: Design principles for valid datasets*. Presented 25 May 2013 at 21st Manchester Phonology Meeting. Accessible via https://uq.academia.edu/ErichRound.

Erich R. Round. 2014. *The performance of STRUCTURE* on linguistic datasets & 'researcher degrees of freedom'. Presented 15 Jan 2014 at TaSil, Aarhus, Denmark. Accessible via https://uq.academia.edu/ErichRound

SIL Mainland Southeast Asia Group. 2002. *Southeast Asia 436 Word List* revised November 2002. http://msea-ling.info/digidata/495/b11824.pdf

Gary Simons and Steven Bird. 2000. *The seven pillars of open language archiving:A vision statement*. http://www.language-archives.org/docs/vision.-html.