Cyberling 2009 Workshop:
Towards a Cyberinfrastructure for Linguistics
Workshop Report

Emily M. Bender, PI
1. Introduction

The Cyberling 2009 workshop was held in Berkeley, CA, from July 16-18, 2009, in conjunction with the LSA Summer Institute. The workshop brought together 43 researchers, representing a variety of different perspectives on the field of linguistics (including the perspective of non-linguists, from information science, computer science, and archiving), to discuss the way forward towards a cyberinfrastructure for our field. The workshop participants were invited with an eye towards supporting a conceptualization of infrastructure that was as inclusive as possible of different approaches to the study of language and towards connecting as many geographically dispersed research communities as possible. The list of participants, their research fields, and their affiliations are given in Section 10.

The workshop format alternated between plenary sessions and working-group break-outs. Each working-group was tasked with exploring and writing a white paper about a particular topic related to the development of a cyberinfrastructure for linguistics. The plenary sessions included an organizational introductory session, short sessions during which the groups presented updates on their discussions, and three open presentations to the general linguistics community, on data sharing, computational methods, and the results of the working groups.¹

The questions posed to the working groups were all interrelated, to varying extents, and it was impossible to avoid some duplication of effort. Nonetheless, we found that the working group sizes (5-7 participants each) were very conducive to productive collaboration, and the workshop wiki and plenary update sessions allowed a chance for cross-working group discussion. In some cases, the working groups used some of the break-out time to hold joint sessions. We are encouraged by the degree of convergence in the recommendations of the different subgroups in the cases of overlap. To a certain extent, this was achieved through cross-working group collaboration, but more broadly it suggests that there are real answers to be had to the questions we are asking, and validates the work done by the participants.

It was clear from the discussions at the workshop that there is widespread interest, across subfields and among individuals as well as scholarly organizations and funding agencies, in creating, promoting and using cyberinfrastructure in linguistics. It was also clear that communication is a real problem. Even among people actively involved in cyberinfrastructure-related projects, individuals were not aware of other large-scale projects in the field. Thus one of the primary impacts of the workshop was to open up the lines of communication across these disparate projects. We hope to continue that broadening of communication through our nascent blog, http://blog.cyberling.org, to be launched by the end of 2009.

The working groups were asked to use the workshop wiki (http://cyberling.elanguage.net) to prepare and publish their white papers. Some groups produced a single page, while others produced a set of linked pages. Our original plan had been to cultivate this wiki as a repository of information about cyberinfrastructure for linguistics and an online community for collaboration on matters relating to building our cyberinfrastructure. However, in the weeks following the workshop, the wiki did not take off. We have thus decided to “freeze” the wiki as

¹ These open sessions did not bring in as many external participants as we had hoped, perhaps due to insufficient advertising, and perhaps due to competing events.
an artifact that was produced by the workshop, and focus our efforts at online community building and information dissemination on the blog mentioned above.

The remainder of this report consists of executive summaries of the white papers from each of the seven working groups, followed by the white papers themselves. Where there are multiple web pages in a single white paper, they have been linearized as subsections or appendices to the paper. Hyperlinks have been turned into footnotes with URLs.
2. Summary of White Papers

2.1. Working Group 1

Working Group 1 was tasked with identifying and documenting existing and needed standards for annotation of data. In answering these questions, they produced a working definition of “annotation” and “annotation standard”, a statement of best practices in the development of good annotation standards, and a handful of case studies of existing annotation standards.

WG1’s working definition of “annotation” is “the act of adding, to primary linguistic data, information representing analyses or models of aspects of the data.” Their working definition of “annotation standard” is “is a set of conventions that is associated with a commitment to adhere to the conventions by a community of users.” Each of these definitions is buttressed with examples and further discussion, including discussion of what annotation and annotation standards are good for. The best practices identified are consistency/reliability, usability, resilience, accountability/responsibility, interoperability and extensibility/adaptability.

A major take-home lesson is that annotation standards are not static objects, but rather dynamic conventions that evolve as they are applied to growing bodies of data. As such, defining an annotation standard requires not only the initial scientific work of creating it, but also the organizational work of ensuring its long-term stability. This work includes mechanisms for approving additions/changes to the standard, as well as community support in the form of teaching and mentoring, and the development (and maintenance) of standard-compliant tools.

The WG1 white paper concludes with a list of links to and descriptions of existing standards and resources, across the subfields of phonetics/phonology, morphosyntax, syntax and semantics, pragmatics and discourse structure, gesture, and others.

2.2. Working Group 2

Working Group 2 was tasked with looking into other standards, existing and needed, outside of annotation standards. They begin by refining that goal to looking at standards for storage, retrieval and search of data. Their white paper includes a discussion the distinction between standards and best practices, how to encourage wider adoption of standards, and information about participating in the ISO process, as well as case studies describing applications of standards for storage, retrieval and search, a set of seed lists of existing standards, both for linguistics in general and specific to certain subfields, and a discussion of not-yet-existent but needed standards.

The WG2 paper gives examples of specific areas where standards are needed, grouped under the general headings of storage, retrieval, search and access/reuse, as well as noting that standards for metadata are also needed.

- Storage: digitization of primary data and metadata, open access standards and formats, providing within-subfield models of best-practices for sharing legacy data, publication (cf. WG4), and archiving.
- Retrieval: versioning, digital fingerprinting, adaptive coding (the ability to adjust coding scheme as knowledge evolves), stable addressing of resources, web standards
• Search: linking to audio/video source materials, annotation conventions, consistency and quality of annotations
• Access/reuse: citation standards, privacy and legal issues, subfield-specific usability concerns, reusability

The case studies cover the creation of interlinearised data, reuse of linguistic data for NLP, character encoding (Unicode), web standards, version control in a sociolinguistics project, and genealogical classification using AutoTyp. The seed lists cover sociolinguistics, typology, and language documentation and description, and were written with the intention that others would add information to the wiki.

WG2 found that for web services provided by distributed, multi-organizational groups, the REST (Representational State Transfer) style of web standards is the best existing starting point, however, we would need to build discipline-specific extensions to existing architectures.

2.3. Working Group 3

Working Group 3 was tasked with identifying existing and needed tools which will encourage “ordinary working linguists” to both use the existing data available through a nascent cyberinfrastructure and, more importantly, contribute their own data to the effort. Such tools need to be standards-compliant, making it easy for linguists to contribute data while helping them do things they already want to do. Thus WG3 was asked to identify possible “killer apps”, tools which would be so compelling and useful that they would see widespread adoption and thus widespread migration and contribution to cyberinfrastructure.

The WG3 white paper is a narrative describing a “killer app” in this domain, the kind of research workflow and collaboration it could support, and then focusing in on what could be done as a first step. The central idea is that of web services, i.e., making use of cloud computing to provide benefits to the researcher (access to tools, computing power, and off-site back-up) while also setting the stage for seamless data sharing when researchers are ready for that. One thing that has not yet been explored is how this model of application(s) fits with the publication model of ensuring data reliability and provenance, as explored by WG4.

2.4. Working Group 4

Working Group 4 was tasked with addressing the twin issues of data reliability and data provenance, or how to ensure that data creators get credit for their work while data consumers can verify the quality of the data they are using. The solutions to these problems will need to scale to the case where a data set has multiple layers of annotation by different authors. WG4 points out in their white paper that the integrity and completeness of provenance data is also important for rights management, including the rights of the speakers who are the original source for the data.

The heart of WG4’s proposal is to promote curated data sets as publications. They note that the technology for doing this is already well-understood, but that substantial institutional and social engagement on the part of our field is required to establish publication of data sets as a community norm.

WG4 proposes the use of Persistent Identifiers (PIDs), which remain the same even as the URLs they point to change, as a key technology in building a provenance system. PIDs can be
associated with every individual and organization involved in the production of an annotated
data set, and associated with particular pieces of information within the dataset. PIDs can also be
used in access-management systems.

The WG4 white paper ends with suggested first steps, including training linguists (of all career
stages) in the practice and importance of data publication; senior linguists setting an example by
publishing data; providing information on how to cite data sets along with the data; publishers,
editors and reviewers requiring provenance information for all data cited; granting agencies
requiring publication of data, with real consequences for researchers and their institutions if there
is no follow through; and establishing electronic data publishing journals with peer review.

2.5. Working Group 5

Working Group 5 was tasked with investigating how cyberinfrastructure has been developed in
other fields, and abstracting what can be applied to cyberinfrastructure for linguistics. WG5’s
research into other fields focused on the natural sciences, as these have been more successful in
setting up cyberinfrastructure than the social sciences. In considering various approaches to
cyberinfrastructure, WG5 noted data handling practices, collaborative/organizational structure,
and the internet technologies and tools used.

Based on reviews of cyberinfrastructure efforts in astronomy, human geography, and biology,
WG5 concludes that a key practice in data handling is using a distributed model with a
decentralized data store, allowing many people to publish. In order to produce a cohesive
cyberinfrastructure out of distributed data and data providers, WG5 promotes the use of
terminologies and ontologies, which should allow some flexibility of data
presentation/visualization including “localization” to different term sets. Finally, successful
cyberinfrastructure requires open data (to the extent possible) and citability of data (cf. WG2 and
WG4). In terms of collaborative/organizational structure, WG5 briefly outlines a philosophy
they dub “collaborative modularity” in which multiple groups work independently, building tools
and standards and collecting data to meet local needs, but with an eye towards sharing those
tools and standards with others, including planning for sharing from the start, and seeking out
and using existing standards. Finally, under internet technologies and tools, WG5 recommends
the use of Linked Data (making the meaning of assertions in annotations explicit), cloud
computing (cf. WG3) and Web services protocols, including SOAP (Simple Object Access
Protocol) and REST (cf. WG2 and WG3).

2.6. Working Group 6

Working Group 6’s task was to investigate funding models for creating and maintaining the
various components of a cyberinfrastructure. The working group generalized from “funding” to
“support”, as support other than direct financial contributions can also be critical to the success
of the project. WG6 identified four dimensions of the problem: projects and activities that need
support, potential sources for support, organizational structures through which support can flow,
and general concepts and principles.

The activities in need of support include creation of data sets, tools, and standards, as well as
creation and maintenance of archives, development of instructional materials, and new forms of
scientific communication. Sources of support include government agencies, corporations,
private foundations, small private donors, productization, institutional buy-in, overlap with
pedagogical effort, and volunteer effort. Organizational structures include academic and
government entities, free-standing or affiliated labs and institutes, companies, consortiums, and
software/standards-centered communities. The key principles include interdisciplinary
cooperation (beyond linguistics), international cooperation, active community participation,
citability (cf. WGs 2, 4, and 5), and consideration for the life cycle of the project as well as the
public interest.

WG6 had several proposals for immediate action. The top-down initiatives proposed are a letter
to the head of the Linguistics Program at NSF, an open letter from the President of the LSA
suggesting a national initiative to rescue the cultural legacy of archival speech and language
material, and recommendations to the LSA’s Technology Advisory Committee to gather and
publicize links for graduate education pertaining to cyberinfrastructure and supporting skills.
WG6 identified three areas that could benefit from immediate bottom-up initiatives providing
good targets for funding: venues for publication of data, creation of a model for what
documentation of reasonably widely-spoken, under- or undocumented languages should look
like, and better tools for corpus-based distributed lexicography.

2.7. Working Group 7

Working group 7 considered the problem of collaboration structure, and how to coordinate the
efforts of people involved in cyberinfrastructure creation. WG7 distinguished two types of
collaboration: joint research and large-scale coordination of effort, and focused on how to bring
about the latter. They identified three key ingredients: cooperation, community building and
communication. Under cooperation, they identified a variety of models of coordination,
including exploiting existing infrastructure (e.g., ISO), peer-to-peer efforts such as TalkBank,
and building infrastructure from the ground up as in TEI. They identified several distinct means
of community building, including funding programs, evaluation campaigns, on-line fora, SIGs,
workshops, tutorials and journals. The white paper also lists several potential communication
vehicles, including wikis/blogs, workshops/tutorials, and reviewing guidelines and reviewer
feedback, among others.

In their discussion of content to be communicated about, WG7 echoes the other groups in calling
for interoperability, the promotion and adoption of standards, promoting researcher “buy in”,
establishing and publishing shared principles for managing issues connected to open access to
data, and building bridges to other fields working with speech and language data. The LSA and
other scholarly organizations stand out for their potential to facilitate this communication. We
were fortunate to have a representative of the LSA in WG7, and as a result, an informal working
group within LSA has already been created to explore how LSA can help bring out the necessary
“culture change”.

WG7 ends their white paper with a list of short term and long term action items. In the short
term, WG7 calls for recommendations to be drafted for funding agencies regarding use of
standards, data publication, etc., and for journal editors and conference organizers regarding the
citation of tools and resources and the publication of data; for teaching resources to be created a
curated, perhaps through the LSA; and for the conversation(s) begun at the workshop to continue.
In the long term, WG7 envisions stronger communication among projects and groups related to
cyberinfrastructure, encourages those developing standards to contribute to ISO TC37 SC4, and
for the field to work towards harmonization of data formats and annotation format and content.
3. White Paper: Working Group 1

Working group members: Mary Beckman (co-chair), Stuart Robinson (co-chair), Sarah Churung, Greville Corbett, Charles Fillmore, Richard Wright

3.1. Preamble

The Annotation Standards group was charged with identifying and documenting existing and needed standards for the annotation of linguistic data. Further, the group was asked to consider possible standards that may need to be developed in the future. Annotation standards support interoperability, aggregation of data, and (ideally) applications that help linguists address the research questions that they are interested in answering while creating consistently annotated data as a side-effect. Another consideration that the group took into account is that sometimes these goals may be in partial conflict with standards for the ethical treatment of human subjects.

This wiki page is the report (“white paper”) from the group, who acknowledge the helpful comments of the other attendees at the Cyberling09 workshop, particularly Emily Bender, Nancy Ide, and Mark Liberman.

3.2. What is annotation and what is it good for?

Annotation is the act of adding, to primary linguistic data, information representing analyses or models of aspects of the data. For example, if the primary linguistic data are an audio recording of a sequence of turns in a conversation between two speakers, then one type of annotation could be the marking of speaker-change points in the conversation within a layer of annotations related to the analysis of discourse structure. Another series of annotation layers could begin with an orthographic and/or a segmental transcription of the speech. Other annotation layers in this series might include a tokenization (segmentation) and glossing of the words or other similar units in the orthographic or segmental transcriptions of the recording. Other series of annotation layers on the morphosyntactic side of language could include a subsequent set of part-of-speech assignments to the words and/or a parsing of the syntactic structures of the sentences and other linguistic expressions in the recording. A parallel series of annotation layers on the phonetics/phonology side of language could include the tagging of linguistically significant events in the spectral patterns of the utterances (e.g., the release burst of each plosive and the transitions between different voice qualities), a parsing of prosodic structures that group segments into syllables and higher-order constituents, and the identification of salient points of coordination in the rhythms at different levels (e.g., marking of stressed or accented syllables).

3.2.1. What elements of an analysis can be annotated?

As the above example illustrates, elements of a linguistic analysis that can be annotated and for which annotation conventions can be codified separately are of at least three types: (1) tokenization/segmentation, (2) syntagmatic structure, (3) paradigmatic content of the events/tokens and structure. In the ontology of linguistic annotations, these aspects could be thought of as (1) the identification of instances or things, (2) the identification of relations among things, and (3) the identification of classes of things or relational functions. Bird and Liberman

1 http://en.wikipedia.org/wiki/Ontology_(information_science)
(2001) give an insightful discussion and a framework for formulating good ways of treating all three aspects. They propose to formalize this ontology in terms of the annotation graph — a directed acyclic graph, in which each annotation token is (minimally) a triple consisting of two nodes that point to the positions in the string of labels on any annotation tier, and the label for the arc connecting these points, as in Figure 1 and Figure 2 below.

Figure 1. Spectrogram (and Praat TextGrid merge of original label files) for the first three words in utterance train /dr1/fjsp0/sa1 from the TIMIT corpus (top) with a screenshot of the original phn and wrd label files (the first and third tiers of the Praat TextGrid) and the associated annotation graph snippet from Figure 2a in Bird and Liberman (2001).

Figure 2. Sample view for the first utterance in a Hayu narrative\(^2\) from the LACTITO archive (top) with a screenshot of the snippet of the annotation file and associated annotation graph from Figure 5 in Bird and Liberman (2001).

\(^2\) [http://lacito.vjf.cnrs.fr/archivage/tools/list_rsc.php?lg=Hayu](http://lacito.vjf.cnrs.fr/archivage/tools/list_rsc.php?lg=Hayu)
3.2.2. Examples

We expand on these three different aspects by illustrating each in reference to the type of example described above, where the primary linguistic data are an audio recording, and also in reference to cases where the primary linguistic data are instead a written text, which may or may not have begun as an orthographic transcription of an audio recording.

(1) For audio data, tokenization considerations can include not just the need to decide on the number of things that are instantiated at any given level of annotation, but also the need to agree upon where token boundaries (e.g., edges of segments or words) should be placed relative to the disparate spectral cues to the often asynchronous and/or smoothly changing postures of different articulatory systems. For text data, tokenization considerations similarly can include the need to introduce word boundaries (e.g., as spaces in text) for languages with writing systems that do not use space for word separation, or, for English, the need to separate punctuation that shows significant syntagmatic boundaries — by adding surrounding spaces — from punctuation that is a part of a name or word (e.g., “that’s that!” as opposed to “etc.”, “Mrs.” “U.S.A.”).

(2) For audio data, syntagmatic structure can include agreed-upon conventions for identifying time points on different annotation tiers that should be synchronized because they are the same event (e.g., the time stamp for the beginning edge of the first segment in a word is also the time stamp for the beginning edge of the word, and should change if/when the segment boundary is moved). It can also include the principles for differentiating different types of coordination across annotation tiers (e.g., when the time of a linguistically meaningful fundamental frequency maximum is identified relative to the time of a stop release). For text data, syntagmatic structure can include the bracketing of sequences of text that function as single constituents, or the indexing of anaphors to their antecedents, or the indexing of discontinuous collocates, as between the first and last words in “wreak this type of havoc”.

(3) For audio data, the development of annotation conventions inevitably includes work to agree on the set of contrasting types to distinguish at several levels. For example, in tagging consonant and vowel segments, should the labels include only “broad” phoneme classes, or should major allophones be distinguished, or even finer phonetic detail marked? Whose intonational analysis should be adopted in tagging utterance melody? For text data, similarly, annotation conventions can involve work to codify the set of paradigmatic contrasts at many levels, from conventions regarding the number of different types of filled pauses and how to spell them in an orthographic transcription of recorded speech, to conventions for identifying cells in morphological paradigms, which might need to be differentiated in significance across languages. For example, systems for morphosyntactic glossing intended for cross-language comparison might need to recognize that for one language ‘singular’ is understood as in paradigmatic contrast to ‘plural’ while in another language it might be in contrast with ‘dual’ and ‘plural’. This could conceivably be achieved by adding a “legend” to a given annotation layer, for a given language, linking each special annotation category to a description of the relevant contrast set, e.g., in a grammar.

3.2.3. The purpose of annotation and its relationship to the data

As the last example should make clear, conventions for each aspect of an annotation scheme cannot be established without thinking carefully about what the data are and what the annotations are for. In general, annotations are dedicated to specific purposes: it is hard to imagine a corpus development project that seeks to account for every detectable phenomenon in a language sample. Sociophoneticans who are interested in the increased use of creaky voice by young
female speakers of American English would want to have an annotation tier where beginning and endpoints of creaky voice are marked; phonologists who are interested in comparing the probability of finding a particular segment sequence within a word to the probability of finding that sequence across a word boundary obviously do not need such a tier. Researchers interested in preferred or habitual locations on the part of individual speakers obviously need to have speaker-ID information in the data they study; those interested in finding examples of syntactic phenomena obviously do not need such information. These different needs for more or fewer levels of annotation can be described as different points on a scale of granularity. But the different granularities can also involve the same level of annotation, but different degrees of specificity in the paradigmatic set.

The following is an example of levels of granularity in the syntactic description of English rate expressions such forty dollars an hour, forty miles an hour, forty miles a gallon, forty times a day, forty dollars an ounce, and the like. In dealing with such phrases, one purpose might be that of providing a preliminary mark-up for a parser, inasmuch as this pattern of two adjacent NPs is not a part of the ordinary grammar of the language. For this purpose, it would be enough to block off such phrases and mark them as NP: as such they then can fit into PPs (moving at forty miles an hour) and VPs (earns forty dollars an hour, gets forty miles a gallon), etc. A quite different purpose one might have for annotating such expressions is providing a mark-up that is usable for language understanding efforts: in such cases the type of unit (linear extent, money amount, time, weight, etc.) in each of the two parts of these expressions should be indicated (with information from a lexicon), allowing the automatic assignment of the phrases to such categories as Fuel-efficiency, Price-per-unit, Frequency, Speed and the like.

An obvious and important use of annotation is that of providing a layer of representation that is available for further analysis; in a sense, this amounts to regarding one person’s annotation as another person’s primary data. For example, phonological or orthographic transcription allows morphosyntactic analysis of a sample of speech more directly than the acoustic trace.

This property of layered annotation raises the issue, strongly associated with the late John Sinclair, that mistakes in one layer of annotation compound mistakes in higher layers. This point can be seen in the fact that descriptions of English syntax tend to accept the tokenization implied in the standard orthography, so that, for example, whose and another are treated as single units. Proposals about the clitic vs. suffix analysis of the possessive marker in English would be argued differently if whose were who’s, making who the hell’s fault is that? seem not so anomalous; and descriptions of the pattern that allows a mere twenty dollars, an extra five pages, an additional twenty dollars could be seen as incorporating another five pages (i.e., as an other five pages).

An analogous situation arises when phonological descriptions accept the tokenization or the set of paradigmatic categories implicit in the conventional segmental transcription for a language, even when describing speech produced by child speakers who have not yet acquired the phonological system of the language.

These pitfalls suggest that the sociology of developing annotation conventions might be an object of study in its own right. They also suggest that linguists should think flexibly about the types of things that should be considered annotations, so that conventions can be developed for how to link these things to the primary data. For example, it might be appropriate to think of responses from naïve judges, elicited over the web using Mechanical Turk or the like, as a kind of annotation, in which case, it could be useful to develop standards for eliciting these judgments
and tools for linking the responses back to the corpus of primary data that provided the stimuli. It might also be appropriate to think of skilled formant “correction” as a kind of annotation, in which case, there could be standards for “correcting” formants and associating the formant traces with the corpus, as in the development of the tools for the Origins of New Zealand English Project.³

3.3. What are annotation standards and what are they for?

Some annotation is created to serve a single researcher’s needs. If the annotation practices developed by this researcher, for a certain class of phenomena, are consistent, so much the better for this solo researcher. An issue of annotation standards arises when there is a need or opportunity for other researchers to work with the same data, or when researchers become interested in the same kinds of phenomena in other data samples, or in other languages, and want to be able to make generalizations.

An annotation standard, then, is a set of conventions that is associated with a commitment to adhere to the conventions by a community of users. A standard can evolve gradually in a community of researchers who are working on similar problems in some language domain, so that assumptions about the analytic space converge in some way that promotes the natural emergence of infrastructure for developing, transmitting, and codifying a standard. A standard can also arise from adopting a tool that brings with it assumptions about the data being analyzed that can be met by adhering to the standard.

No matter the path of convergence, however, annotation standards cannot be defined without reference to a shared set of assumptions and an associated community of analysts. As a corollary, a set of annotation conventions cannot be evaluated (or standardized) without at least an initial reference to a community of analysts and users.

Within such a community, annotation adds value by spreading the workload of providing agreed-upon analyses to a larger set of shared primary data than can be analyzed by a researcher working alone. Looking outward from the core community, annotation provides expert analyses for others who might not otherwise have access to the primary data.

This understanding of the relationship between the analyses of the data that are to be encoded in the annotations (the “model”) and the primary data themselves leads to the following characterization of what annotation standards are for and how they can be evaluated.

3.3.1. Within the original community of developers and users

- It is critically important to ground any annotation schema in terms of the particular model of the phenomena being annotated, and to develop it in relationship to the question being asked and the shared assumptions of the community about the phenomena being observed and modeled. Within this initial community, the annotations evolve as a set of “common law” rules about what the observed phenomena are. These rules will specify how the data should be segmented into tokens, how the tokens will be labeled in terms of an agreed upon inventory of contrasting types and relationships, and how relationships among tokens will be parsed and labeled.

³ http://www.ling.canterbury.ac.nz/onze/cc.html
A set of annotation conventions (rules), therefore, can only be evaluated first in relationship to the initial user community and their questions. While there might be domain-specific evaluation criteria, defined relative to independent observational tools, one critically important evaluation criterion that is common to all domains is the reliability of the annotations. Is the annotation consistent within and between annotators? If an annotator observes the same data twice in different independent annotation sessions, are the analyses (the annotations) the same? Similarly, if two different annotators observe the same data independently, do they arrive at the same analyses (tokenization and labelings)?

To achieve consistency typically requires a long iterative process of “common law” development, during which two or more users annotate some set of data separately, then convene to discuss and adjudicate the disagreements, formulate new principles to cover the cases discussed, and then start a new round of independent annotation, comparison, discussion. The initial users will need to agree on the degree of consistency that is needed to accomplish their goals. An ancillary set of “laws” will need to be developed to reliably differentiate between disagreements that arise from intrinsic ambiguity and disagreements that arise because the conventions and annotation tools are not yet at the point of required coverage/stability/usability.

It can be useful in the process of developing, evaluating, and using an annotation standard to work on the different aspects separately. For example, if at some stage of development (or in some subcommunity of users), the tokenization is more reliable than the identification of relationships, the annotations might be adequate for some subset of the initial purposes, but not others. It is then important to develop conventions for tagging corpora or parts of corpora for relevant facts such as which version of the annotation scheme was used or the level of experience and/or training of the annotator(s).

3.3.2. When extending to a new community of users…

Annotations developed within a particular community of initial developers and users might be extended to another community of users who are addressing different questions and may have different model assumptions. The goodness of a standard then becomes a product not just of the initial developers/users, but also of the flexibility/ingenuity of the later users of annotated data.

The needs of various communities are in some cases overlapping (both phoneticians and sociolinguists may seek standards for phonetic annotation) and in other cases conflicting (a fieldworker may want language-specific idiosyncratic part-of-speech categories in interlinear glossing whereas a typologist may want agreed-upon cross-linguistically motivated categories). To get a sense of the potential disparities among different communities of users, we listed the first sets that came to mind:

- sociolinguists
- computational linguists and NLP practitioners
- language acquisition specialists
- psycholinguists and laboratory phonologists
- specialists in speech and language disorder
- fieldworkers and language typologists
• stylometrists, disputed-author researchers, etc.
• educators evaluating text complexity, comprehensibility

Even within a single later-adopting community, however, the questions and needs may differ in relationship to different types of primary data. Here we can differentiate at least among (1) full video recordings, (2) audio-only recordings, and (3) spoken utterances that were recorded only as text in the first level of “annotation” of the fieldworkers’ written transcription. The initial tokenization/labeling of each of these primary data types may be an orthographic transcription, and in the case of type (3), the initial tokenization/labeling then becomes the only record. For some communities of users, the models and questions related to such transcriptions might differ dramatically from the models/questions that can be applied to data that are (4) originally written texts.

In adopting (and adapting) a set of annotation conventions to a new set of questions and applications, then, it is again useful to ask: What aspects of the annotations can we usefully tease apart and evaluate/adopt/develop separately?

3.4. What does it take to be a good annotation standard?

3.4.1. Best practices (themes)
The associated properties that define a good annotation standard can be grouped into a few overarching themes and associated questions about the annotation conventions:

• Consistency/Reliability
  ◦ What is the history of the annotation conventions? Did they evolve in careful, iterative rounds of (1) discussion of the goals of the annotation set, (2) independent annotation of a suitably diverse corpus of primary data by a large number of annotators, (3) calculation of inter-annotator agreement, and (4) discussion of points of agreement and disagreement and incremental revision?
  ◦ Are there standards / mechanisms for continued calibration of consistency within and between annotators?
  ◦ What are the published intra-annotator and inter-annotator consistency rates?
  ◦ Are the conventions designed to allow transparent, easy, reliable “back-tracking” to the primary data, via time stamps or via sequence position nodes within an annotation stream that has a reasonably fine-grained tokenization?

• Usability
  ◦ Is there good (accessible and extensible) documentation?
  ◦ Is there a suitably diverse and continuous community for teaching (and testing the ability of) new annotators / users?
  ◦ Are there good tools for annotating and using the annotations, and good community mechanisms for building / extending / sharing tools?
  ◦ Is there a reliable connection between the annotations and the primary data that allow the user to track back to the data to check a suitable subset of the annotations?
  ◦ Is the design of the annotation schema such that annotations can be used as reliable tags back into the primary data, for easy queries using standard query tools?
- **Resilience**
  - How does the standard deal with inter-annotator disagreement? Is information about disagreements preserved so that they can be analyzed in the course of developing the next version, to determine whether there are common cases of inherent ambiguity that need to be marked, or new cases that the conventions do not yet cover?
  - Are there principled mechanisms for marking degree of uncertainty about difficult or ambiguous cases? (See the CHAT manual for a thoughtful discussion of this question.)
  - Are there graceful ways of choosing to provide more or less specific degrees of analysis?
  - Are there good mechanisms for providing and getting the most out of partial annotations?
  - Are there robust ways of extending partial annotations to more of a corpus and of verifying and modifying the annotations of a corpus?
  - Relatedly, are there good mechanisms for keeping track of which parts of a corpus are in what state of annotation and verification / modification?

- **Accountability/Responsibility**
  - Again, are there robust mechanisms for maintaining transparent links back to the primary data? and are these mechanisms ethical? Do they insure the explicitly or implicitly agreed-upon degree of confidentiality of the person or people who produced the primary data (or who produced some subset of the annotations by providing naive judgments)? The issue of confidentiality is especially vexing when the primary data are video recordings. (See sections 4.3 and 4.4.)
  - Do the standards encourage (or even allow) later “consumers” to credit the annotations in publication?
  - Are the annotators (or the annotator level) for different parts of a corpus or different aspects of the annotation identified in a way that allows later users to partition the annotations -- e.g., into annotations by native speakers versus non-native speakers?

- **Interoperability**
  - Can the annotation be validated and used in different tools or computational models?
  - Is the logical structure of all three aspects of the annotation conventions transparent, and transparently related to the documented descriptions of the annotated phenomena?
  - Also, is it possible to translate to and from some other annotation conventions that have been used for this set of phenomena, in a way that makes it possible to share data across different analytic frameworks?
  - Are the formats for encoding the different aspects of the annotation conducive to using the annotations for purposes different from the originally intended ones?
  - Are the definitions of the annotation elements freely available and stored in an open format?
  - Are any requisite tools for annotating or using the annotations free open source?

- **Extensibility/Adaptability**
  - Can the annotation schema be extended to annotating utterances in other styles from the utterance sets for which it was developed? Can the annotation conventions be used for utterances produced by other speaker types? Can they be extended to (or readily adapted for) annotating data from other dialects, other languages, ....?
Is there a solid and suitably diverse core of users (and “maintainers”) to allow the standard to evolve and change in response to user feedback and/or to new needs?

Is there a sensible consensus or mechanism for deciding when to “publish” a new version?

Are there good standards and mechanisms for versioning? For example, is there a robust way to permanently associate meta-data about which version of the conventions was used in annotating (different parts of) any corpus? Are there tools for keeping track of who the taggers were at different levels / times, and are there tools (or at least a “crib”) for how to “translate” across corpora and/or across levels of annotation as the standard evolves and expands?

3.4.2. Best practices ("tangibles")

The hallmarks of an emerging annotation standard therefore begin with these two important social characteristics:

- community: There is a sustainably large and diverse community of core users/maintainers.
- history: There is a history of effective dissemination of the conventions and recruitment of new core users.

Other more tangible accoutrements of annotations standards that have exemplified the best practices identified above include:

- documentation: The conventions are adequately and fully documented, in a “reference manual” that can be consulted easily by experienced users.

- training manual: Ideally, there is also a separate, well-tested training manual (or a standard syllabus for training courses) that leads new users through a graduated sequence of more and more difficult examples culled from data that were annotated in developing the documentation and/or the reliability metrics.

- inter-annotator reliability metrics: There are published records of inter-annotator consistency tests. Ideally, these tests differentiate between disagreements that stem from intrinsic ambiguity and disagreements that have other, remediable sources such as inadequacies in coverage, deficiencies in the documentation, or the like.

- computational tools: Members of the community have invested in developing computational tools that increase the reliability of the annotations.

- conventions for metadata: The community has developed mechanisms for protecting confidentiality of the producers of the data, crediting of the provenance of the annotations, and so on.

- conventions for responsible maintenance: A set of conventions, or a more elaborate institutional framework, has also emerged, for responsible maintenance of the conventions, for continued elaboration of the documentation, and for updating of the training manual (or re-accreditation of the training courses).

3.5. The state of the art (with some case studies)

In this section, we illustrate the considerations outlined in Section 3 by briefly reviewing the development and current state of annotation standards in four very broad areas. These reviews highlight two factors that promote or hinder the development of reliable and resilient standards.
The first is the degree to which the “semantics” of the target phenomena are naturally constrained. At one extreme is the case of phonological annotation of consonants and vowels of spoken languages. Here, the aerodynamics of the vocal-auditory channel tightly constrains the tokenization and types of relationships at the lowest level of the prosodic hierarchy in an extremely robust way. An example at the other extreme is the analysis of grammatical constructions, where it is difficult to even imagine what boundaries could be imposed by nature on what a language-specific morphosyntactic construction can mean.

The second factor is the age of the language type and/or of the systematic linguistic investigation of the phenomena across languages. Spoken languages may have existed as long as there have been modern homo sapiens, and the “annotation” of consonants and vowels goes back to the first alphabetic writing systems. By contrast, the systematic study of signed languages has a much shorter history.

3.5.1. Phonology of spoken languages

3.5.1.1. Annotation systems for vowels and consonants

As noted above, tokenization and other aspects of the analysis of categories at the lowest level of the prosodic hierarchy for spoken languages is tightly constrained by the (psycho)physics of the human articulatory and auditory systems. As a result, it has been relatively easy to develop conventions for annotating utterances of spoken languages at this level of this part of the grammar, using an alphabetic analysis, and the International Phonetic Alphabet⁴ is a premier example of a well-developed annotation standard. It is maintained and updated by the International Phonetic Association,⁵ which was established in 1886 and today is associated with the International Congress of Phonetic Sciences, a meeting held every four years which attracts several thousand attendees.

The International Phonetic Alphabet has a handbook,⁶ which documents the annotation conventions and specifies a well-codified format for presenting a catalog of the consonant and vowel inventories of a spoken language variety using the IPA. (There is a long-standing section of the Journal of the International Phonetic Association devoted to publishing such language-specific schema.) The most recent version of the handbook was published in 1999, after a conference convened in 1989 to review the coverage of the categories in the IPA consonant and vowel charts and the lists of other symbols for categories that do not fit neatly into the tokenization and paradigmatic features that are encoded in the consonant and vowel charts. In between the conference and the publication of the revised Handbook, the Journal of the International Phonetic Association published a report on deliberations of the conference as a whole (JIPA, 19: 67-80) as well as reports from working subgroups charged with deliberating on various more focused issues such as “Computer Coding of IPA symbols” (JIPA, 19: 81-82) and “the best means of transcription of disordered speech” (JIPA, 24:95-98) and correspondence from members of the association commenting on the proposed revisions and deliberations at the conference (e.g., JIPA, 20: 22-32).

⁴ http://www.langsci.ucl.ac.uk/ipa/ipachart.html
⁵ http://www.langsci.ucl.ac.uk/ipa/
⁶ http://www.langsci.ucl.ac.uk/ipa/handbook.html
While there is no official training manual, there is a long history of teaching the annotation conventions (and the phonological analyses on which they are based), which predates the founding of the International Phonetic Association. For example, Henry Sweet’s *A handbook of phonetics*, published in 1877, includes vowel and consonant tables that are organized in terms of the same dimensions of analysis as the modern IPA chart — i.e., openness, place, rounding for vowels and place, manner, laryngeal properties for consonants. The same basic approach is also taken in most subsequent textbooks, including Peter Ladefoged’s well-known *A course in phonetics*, which is still used in training students in the annotation of vowels and consonants in many departments of phonetics, linguistics, logopedics, and speech & hearing science.

There is also a history of research on inter-annotator consistency rates for segmental transcription, and on the factors that affect transcription consistency. In general, it is easier to be consistent the closer the annotation is to a “broad” phonemic transcription. For example, Eisen (1993) reports complete agreement among three transcribers of only 50% for a “narrow” transcription, even when distinguishing among only ten “major class” categories such as “voiced plosive”. When the same transcribers were asked instead to note only segments that deviated from an automatically inserted broad “dictionary” form transcription, consistency improved to 85%.

3.5.1.2. Where the analysis breaks down

The different levels of reliability for “broad” versus “narrow” transcription hint at some of the things that affect inter-annotator reliability. Reliability is highest when the primary data are clean recordings of careful fluent utterances produced by adult native speakers of a dialect for which there is a consensus phonemic analysis that can be the basis for the tokenization and consonant and vowel label set, and when the goal of the annotation is to produce a “broad” phonemic transcription as the basis for morphological analysis or the like. Reliability is lower when recordings are noisy, when the primary data are casual or dysfluent utterances, when the dialect of the speaker(s) is an understudied variety that differs from the dialect on which the IPA description is based, or when the goal is to produce a “narrow” transcription as the basis for sociophonetic analysis of variation in the speech community or the like. In the latter cases, reliability is improved if annotators are well trained in phonetics (not just in “classical phonemics”) and have recourse to tools such as the interactive spectrographic display window in the Praat signal analysis tool. However, no amount of phonetic training will resolve the inherent unreliability of shoehorning “sub-phonemic” paradigmatic variation and “suprasegmental” parsing differences into a phonemic segmental model.

Alphabetic annotation of pre-school children’s speech poses special challenges, then, because it assumes (paradigmatic and syntagmatic) phonological structures that may not be in place until the child is much older. Very cohesive research groups, such as the Stanford Child Phonology project that collected and annotated a cross-linguistic longitudinal corpus between 1967 and 1992, can achieve published inter-annotator consistency rates as high as 90-95% agreement. However, this agreement is typically achieved by regular (at least weekly) meetings among the primary transcribers, during which inconsistent tags are “corrected” to a consensus category (or to an “expert” tie-breaker category when consensus is impossible). Pye, Wilcox, and Siren (1988) suggest that this practice hides the true nature of the difficulty, since points of low inter-transcriber reliability can indicate places where the standard phonemic analysis of the target language is particularly inappropriate for the child’s developing phonological system. Hewlett and Waters (2004) make a similar point, suggesting that the problem of obscuring “sub-
phonemic” variability is compounded in most large-scale cross-sectional norming studies, in which fairly “broad” transcription is used in on-the-fly observations without a permanent audio recording, in order to be able to collect data from a large number of children. Edwards and Beckman (2008) suggest that even in cross-sectional studies where “narrow” transcription of recordings is done, transcription should be supplemented by experiments eliciting (potentially continuous) perceptual responses from phonetically untrained native speaker/listeners. All of these researchers remind us that tokenization and paradigmatic differentiation at the level of vowels and consonants is a product of the interaction of the natural constraints from the aerodynamics with the exigencies of lexical contrast in dense neighborhoods. The phonemic analysis that is the basis for the IPA conventions is less compelling for a speaker whose lexicon is still too small to have very dense phonological neighborhoods.

Prosodic phenomena such as stress and syllable structure, and intonational phenomena such as the melodies that group syllables together into phrases and the like, pose a related challenge for phonological annotation. Because there is no comparably compelling natural basis for tokenization of melodic events, spoken languages are much more diverse in the ways in which utterances are structured above the leaf nodes of the prosodic hierarchy. The Working Group on Labeling of Suprasegmentals at the 1989 conference that led to the current IPA handbook recognized this by deciding to recommend no standard annotation conventions for intonation (Bruce, 1989). A basic principle of the ToBI annotation framework also says that “phonetic transcription” of prosody and intonation is impossible. See Beckman, Hirschberg, and Shattuck-Hufnagel (2005) for further explication of this point and the implications for the development of annotation conventions for prosody and intonation. See also Pitrelli et al. (1994) for the remarkably good inter-annotator reliability rates that nonetheless can be achieved when conventions are specific to a particular dialect.

3.5.2. Morphosyntax and semantics

In the preceding section we described how phonological and phonetic annotation is easier for broad class distinctions at the leaf nodes of the prosodic hierarchy, where tokenization is naturally constrained by the psychophysics of speech production. There may be an analogous difference in the degree of difficulty of annotation for morphosyntactic structures and semantics of spoken languages. Specifically, it again seems easier to agree on tokenization of elements that are closer to the bottom of the constituent hierarchy, where the psychophysics of the vocal-auditory channel interact with more general cognitive considerations of attention and memory to promote a “temporal” or sequential (as opposed to a “spatial” or simultaneous) decomposition of the form/meaning mapping. We illustrate by describing standards for annotating aspects of morphology, syntax, and semantics that have emerged in three different communities, before touching on some of the challenges in these areas of the grammar.

3.5.2.1. The Penn Treebank

The Penn Treebank is a collection of English texts that have been grammatically annotated for part of speech, grammatical function (predicate argument relations), and constituency (Marcus et al., 1993; Marcus et al. 1994; Taylor et al., 2001). For each sentence there is an accompanying syntactic analysis (a tree), hence the term treebank (a bank of syntactic trees). Although the Penn Treebank contains material taken from multiple sources (the Wall Street Journal, the Brown Corpus, the Switchboard corpus, and ATIS), its collection of annotated Wall Street Journal newspaper articles is so well known that many researchers think of it the Penn Treebank as the
Wall Street Journal corpus. It is available under a commercial license from and distributed by the Language Data Consortium (http://www.cis.upenn.edu/~treebank/). The corpus is distributed as a collection of texts with accompanying stand-off annotation. The Wall Street Journal section of the corpus, for example, consists of 2499 articles (totaling approximately a million words) published in the Wall Street Journal during a three year period. For each article, there are a number of plain text files that contain various types of annotation as well as a “master” annotation file that integrates all of the annotation.

The Penn Treebank is an important point of reference in grammatical annotation given its success. Not only has it become an important resource in computational linguistics (like WordNet or CELEX), as any search of the literature will reveal, but it has also inspired a large number of similar projects for other languages--e.g., Chinese (Xue et al., 2002), Czech (Hajicova, 1998), Spanish (Navarro et al., 2003), and German (Brants et al., 2002).

The Penn Treebank Project has a number of strengths that help explain its popularity. Chief among these, of course, is that by providing a non-trivial amount of annotated newspaper text it managed to scratch an itch felt by the community of researchers interested in computational linguistics, natural language processing, and related fields. But in addition the Penn Treebank Project provided good documentation and designed the corpus in such a way that it could be easily used: consisting only of plain text files (easily processed), providing good documentation, versioning the corpus, etc.

Despite its success, the Penn Treebank has its weaknesses. One of these is the absence of a standard toolkit or user application for its viewing and/or manipulation. This is unfortunate since such a toolkit could have improved the pace of development and helped improve annotation quality (by eliminating errors that could be detected through automated validation). The under- availability of tools also impedes its adoption by those lacking the resources to develop their own tool. (The creation of various open source toolkits has ameliorated this problem to some extent, but it is a comparatively recent development compared to the age of the Penn Treebank Project.) Another potential weakness is the grammatical model used to describe grammatical functions, which adheres to an old-fashioned Government and Binding analysis that posits, among other things, traces for movement. However, even though the annotation is couched in a multistratal theory of grammar, this has not hindered its use in monostratal theories of grammar, such as LFG (Frank, 2000).

In fact, it is unrealistic to expect a great deal of standardization of the content of annotation for grammatical information given the highly contentious nature of grammatical theory itself. The difficulties inherent in this problem can be seen in attempts to develop treebanks in languages with more flexible word order and discontinuous constituency, such as German. Although it is possible to annotate German using a grammatical theory that posits traces, it leads to inelegance and German treebanks have as a result departed from its grammatical model. (The NeGra and TiGer annotation schemes use graphs with crossing edges rather than simple context-free trees.)

### 3.5.2.2. Leipzig Glossing Rules

The Leipzig Glossing Rules, which build on Lehmann (1983) are a de facto standard for morphosyntactic glossing. They represent a light-touch codification of previous best practice.

---

7 http://www.eva.mpg.de/lingua/resources/glossing-rules.php
They allow for analyses of varying levels of granularity, subject only to the requirement that the segmentation in the glossing line must match that in the source language line. There is a standard set of abbreviations, which could usefully be extended. Documentation consists of a short document, with examples, freely available on-line.

In their current version, the Leipzig Glossing Rules provide the means and the options for linguists of different persuasions to give adequate morphosyntactic glosses. A next step would be to suggest how users should characterize their own usage in a particular publication. At the most obvious level, all additional abbreviations should be specified. Then, we should note that the abbreviations are mainly for feature values (for ‘singular’, ‘feminine’ and so on); it would be good practice to specify which feature each is a value of. Normally this is obvious, occasionally it can cause confusion. Finally, since the rules can be used for different purposes, it is helpful if users clarify and spell out their assumptions. For most purposes, when glossing yesterday we bid three hundred pounds for that horse, we would gloss bid as past tense. The information is derived not from the form itself but from the time adverbial, since the form bid could be present or even imperative. For writing about tense, word order, argument structure, and so on, this solution is fine. If writing about syncretism, however, this theoretical ambiguity would matter, and would need to be indicated appropriately. More generally, the annotator frequently has to be selective in the level of detail included in the morphosyntactic glossing. This means that we cannot expect that different linguists would provide identical annotations. But we should aim for a greater level of consistency than we often find. The rules offer the alternatives, but good practice requires us to choose consciously, to specify the choices made and to apply them consistently.

In terms of tools, it would be useful to have a tool that would check annotations for internal consistency and for any unintentional departures from the conventions.

### 3.5.2.3. FrameNet Annotation Criteria

The Berkeley FrameNet project is using annotation of written corpora to build an on-line lexical resource for English, based on frame semantics and supported by corpus evidence.

Certain annotation practices followed in FrameNet can be seen as arbitrary but motivated. One of the central notions of FrameNet is that of Valence: general descriptions of the combinatorial possibilities of individual lexical heads (verbs, nouns, adjectives, and some prepositions), expressed in both syntactic and semantic-role terms. Valence descriptions are derived automatically from a body of annotated sentences, so it is obviously necessary to agree on how the annotations are structured. The need to pair syntactic arguments with semantic roles motivates our decision to include the “markers” of a phrase with the constituent. For example, if we wish to identify the speaker and the content of an announcement the **announcement by the governor of her decision to resign**, those two elements are blocked off as [by the governor] and [of her decision to resign]: some projects would ignore the prepositions and select only the NPs in those expressions. Similarly, in **the governor’s announcement that she intended to resign** the labels would be assigned to [the governor’s] and [that she intended to resign], including both the genitive suffix and the that-clause. And similarly, then, in **the governor announced to the world that she intended to resign**. A differently motivated project might leave out the structure-markers in order to represent the “content” more faithfully; FrameNet includes these in order to match semantic and syntactic segmentation, allowing users to recognize the structural elements.
It will be seen that the annotations themselves do not distinguish markers that are determined by
the grammar (the possessive ending in the governor’s announcement), from those that are
determined by the meaning of the PP as a whole (the preposition in under the table), or by the
governing lexical head (the preposition on in we can depend on Harry). Such information is
recoverable from the grammar and the lexicon, but it is not part of the annotation.

The standardization of this FrameNet practice has various consequences. All annotators on the
Berkeley project agree to use it in their work, and various FrameNet or FrameNet-like projects in
other languages have agreed to follow the same, or analogous conventions. These are Spanish
FrameNet, Japanese FrameNet and the SALSA project in Germany. Furthermore, Professor
Hiroaki Sato of Senshu University in Tokyo manages a browser of FrameNet data8 and he is
developing a way of pairing valence patterns across the various languages that have FrameNet
databases; the comparisons work best if all users treat function markers in the same way.

3.5.2.4. The difficulty of developing annotation systems for grammatical constructions

There are well-known tree-banks that offer syntactic parses of all of the sentences in a sample,
meeting certain levels of adequacy, such as the Penn Treebank. There are various levels of part-
of-speech tagging for large corpora, such as the British National Corpus, that are for the most
part successful. But a complete record of the special grammatical constructions in a text does not
seem feasible.

For research purposes it should be possible to tag (say) all comparative sentences in a text,
identifying the scales and the phrases that directly or indirectly indicate the entities being
compared. It should be in principle possible to identify all idioms or tight collocations in a text,
however long this might take. It should be possible to notice constructions with certain
peculiarities for the sake of assembling examples for further study, such as, for English, the
pattern that has a degree-modified adjective followed by an indefinite NP marked by the
preposition of (Do you need {this big} [of a box]?). But the expressions that represent
individual constructions are frequently tightly intertwined, and the effort to work out the nature
of such integration on a large scale is not likely to be possible. A sentence like He’s in no bigger
of a hurry than you are exhibits a comparative structure (bigger ... than you are), a collocational
idiom (in ... a hurry, one of the few uses of hurry as a noun), the puzzling structure with the of-
phrase (bigger of a hurry), a special minimizing use of the word no with a compared adjective
(consider the difference between he’s not smarter than your mother and he’s no smarter than
your mother), and the particular form of the than-clause (than you are vs. than you, than
expected, than ever, etc.). Representing the working of all of these constructions and their
articulation is not to be expected.

Research that collects and explores examples of grammatical constructions, idioms, collocations,
and multiword expressions in general, and illustrates their properties one at a time, has got to be
an essential task for linguistics and computational linguistics, both for grammar writing and as a
way of producing learning corpora for machine-learning techniques to improve syntactic parsers.
But since many of the most important constructions cannot easily be associated with individual
words in a sentence, or with specific nodes in a parse tree, there is little likelihood of acquiring
large-scale accurate annotations of grammatical constructions, beyond familiar parsing and
chunking of nonproblematic sentences, any time soon. The problem is further compounded by

8 “FrameSQL” http://sato.fm.senshu-u.ac.jp/fn2_13/notes/index.html
the fact that cross-framework agreement on syntactic phenomena in general is not easy to achieve: dependency-based and constituency-based treatments are not always interconvertible; theoreticians who seek to minimize redundancy in their analyses would not see the same number of construction types in a given text as the grammarian who wishes to work with structures of finer granularity.

The proposal in this section favors rich analysis of small texts, together with extensive sampling of given constructional phenomena one at a time, or in small families of constructions. Such a combined approach should eventually lead to understanding the importance of non-core constructions and multiword expressions, classifying their variety, estimating how many of them there are, determining their relevance in profiling different genres, estimating their “density” in different kinds of texts, exploring the manner in which they are learned, and evaluating their contribution to measures of language complexity.

A sample of constructional annotations prepared within the FrameNet project can be seen on http://www.icsi.berkeley.edu/~hsato/cxn00/21colorTag/index.html.

3.5.3. Annotation of gesture

The two factors relevant for annotation developments, as discussed above in the beginning paragraphs of Section 4, are especially salient for the progression of standards in gesture annotation. We discuss gesture here as inclusive of gestures in signed languages as well as discourse-related gestures of the spontaneous type which accompanies natural spoken language narratives.

First, in contrast to the tightly constrained audio-articulatory modality of spoken language systems, which are articulated using the vocal tract, gesture systems involve the visual-gestural modality. Traditionally, gestures are understood to be articulated using the hands in movements. In the case of sign languages, however, the category of gestures has recently been expanded to include certain movements involving the head, face, and shoulders. (See Neidle et al. 2000 for some discussion of nonmanual gestures in ASL, and Boyes-Braem, 2001 for the descriptions of “mouth gestures” in multiple European sign languages.) Moreover, the visual space of gestures in sign languages exists as a complex continuum which involves the signed phonemic structures in the lowest level of the prosodic hierarchy, larger signed morphemes which engage in spatially bound agreement relationships with other signs, and “nonmanual” gestures of higher-order prosodic structures. So while the conventions for annotating gestures in the traditional sense may be developed with relatively straightforwardness using video analysis, developing annotation standards for gestures which handle the complexity of these relationships must also rely on multiple and dynamic layers of annotation of the types illustrated in the figures and examples in Section 1.

The Language Archiving Technology tool, ELAN,9 is a professional tool for the complex annotation of video and audio sources. “Tiers” are implemented for simultaneously displaying and annotating parallel levels of analysis. These can be nested for dependencies between, say, an independent parent annotation of morpheme-by-morpheme transcription, and referring tiers for

9 http://www.lat-mpi.eu/tools/elan/
the varying gestural articulators (e.g. hand vs. mouth). A full manual for ELAN is available online.\footnote{http://www.lat-mpi.eu/tools/elan/manual/}

Second, the systematic investigation gestures as a linguistic phenomena is a relatively new pursuit. This is true for both the annotation of gestures in spoken language narrative studies (such as with the McNeill Lab project\footnote{http://mcneilllab.uchicago.edu/topics/annotation.html}) as well as the annotation of sign languages (see Neidle et al.’s Sign Stream Project,\footnote{http://www.bu.edu/asllrp/SignStream/} and the Berkeley Transcription System manual\footnote{http://childes.psy.cmu.edu/manuals/bts.pdf}). The next section presents some considerations for developing unified conventions in the annotation of sign languages.

### 3.5.4. A unified annotation standard for signed languages

Finally, we address the need for a unified annotation of sign languages, as identified in this workshop group and codified within the desiderata for qualities of annotation standards in general. First, however, some prior discussion concerning the dissemination of tools and standards among the communities of practice (sign language linguists) is necessary. For although several tools currently exist for the scientific annotation of video data (see above for gesture annotation using Anvil, ELAN, and Cross-Modal Analysis of Signal and Sense, for example), and while the target users are a close-grained community, widespread standard for sign language transcription and annotation is lacking.

#### 3.5.4.1. What sign language annotation is, and what it is not

We begin by clarifying the purpose of a unified standard of sign languages. We emphasize that we do not aim to advocate a writing system of signs, nor do we intend for annotation to replace the primary linguistic video data with a derived set of data. Rather, annotation of sign languages should complement the primary data record as a way of tagging and searching the data. And the goals of a unified sign language annotation standard are to provide a shared platform of convention for collaborating across the various linguistic domains.

#### 3.5.4.2. What a unified standard provides

For all linguists, annotation is paramount, and standards promote convergence. For sign language linguists, the annotation of primary video data poses several challenges. Few standards exist, for example, when it comes to annotating sign languages for fundamental linguistic phenomena such as pronominalization or indexicalization within the interlinear gloss. A more complex issue is the matter of transcribing certain non-manual features that are coarticulated with the manual signs--as functional labels (neg), abbreviations of the action (head shake), or even further break-down of the correlates involved. On the practical side, high-quality video data can require large loads of memory, and utilizing tools for analyzing video requires higher processor speeds and memory load.

These issues stand to hinder the standardization of data annotation. Through further practice and dissemination, however, advances in sign language annotation standards provide the potential for consistency, conversation, and conventionalized practices among a growing community. The
ideal situation (projected solution) is one where sign language linguists, whether collaborating in
an international workshop setting or via remote communications, have access to one mutually
accessible standard that is extensible for all sign languages, interoperable across varying domains
and models of interest, granular across levels of linguistic analysis, and practical for continuous
usability.

3.6. Existing annotation standards and resources
This section lists the various annotation conventions and other resources for developing and
discussing annotation standards that were suggested by participants of Cyberling09.

3.6.1. Phonetics and phonology
- Phone segment tagging symbols, including both:
  - IPA and its various ASCII-fications, such as Sampa, WorldBet
  - and language-specific phoneme-segment encodings such as Arpabet for American English
    and the CSJ encoding for Japanese
- The various ToBI conventions and similar conventions in other frameworks such as the
  ToDI conventions
- The PhonBank conventions and tools

3.6.2. Morphosyntax
- Leipzig Glossing rules
- ISO Morphosyntactic Annotation Format (MAF)
- ISO Lexical Markup Framework (LMF): Homepage (with Publications and Tools)
- Typecraft: a labeling system which, for any verb construction of a given language, provides a
template for that construction type displaying its argument structure, in a fashion as
transparent as possible. The template is constructed from a universally established inventory of
labeling primitives.
- tags for short-unit word (SUW) and long-unit word (LUW) in the CSJ

14 http://www.phon.ucl.ac.uk/home/sampa/
15 http://en.wikipedia.org/wiki/WorldBet
16 http://en.wikipedia.org/wiki/Arpabet
18 http://www.ling.ohio-state.edu/~tobi/
19 http://todi.let.kun.nl/
20 http://childes.psy.cmu.edu/phon/
22 http://www.lexicalmarkupframework.org/
23 http://www.typecraft.org/tc2wiki/Verbconstructions_cross-linguistically_-_Introduction
3.6.3. Syntax and semantics
- FrameNet annotation manual\(^{25}\)
- The tokenization guidelines, part-of-speech tags, and bracketing conventions for the Penn TreeBank project\(^{26}\)
- ISO Syntactic Annotation Format (SynAF)\(^{27}\)
- ISO Semantic Annotation Format - Time and Events (SemAF-TIME)\(^{28}\) (formerly TimeML\(^{29}\))

3.6.4. Pragmatics and discourse structure
- CHAT\(^{30}\) conventions for segmenting turns and identifying the participant and the setting
- DAMSL\(^{31}\) and other schemes documented and discussed at the 1998 DRI meeting\(^{32}\) such as:
  - intentional structure annotation (see Nakatani, Grosz, Ahn, and Hirschberg, 1995)

3.6.5. Gesture
- David MacNeil’s Gesture annotation\(^{33}\)
- BTS sign transcription system\(^{34}\)
- NEUROGES-ELAN system: workshop series homepage\(^{35}\)
- Michael Kipp’s Anvil\(^{36}\)
- Carol Neidle’s SignStream\(^{37}\)
- Vislab’s (Francis Quek’s) Cross-Modal Analysis of Signal and Sense for multi-modal human discourse

3.6.6. Other resources
- The EMU Speech Database System (see Cassidy and Harrington, 2001): sourceforge page\(^{38}\)

\(^{25}\) http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126
\(^{26}\) http://www.cis.upenn.edu/~treebank/
\(^{28}\) http://www.tc37sc4.org/new_doc/iso_tc37_sc4_n269_ver10 wg2_24617-1_semaf-time_utf8.pdf
\(^{29}\) http://www.timeml.org/
\(^{30}\) http://childes.psy.cmu.edu/
\(^{32}\) http://www.cs.umd.edu/users/traum/DSD/schemes.html
\(^{33}\) http://mcneilllab.uchicago.edu/topics/annotation.html
\(^{34}\) http://childes.psy.cmu.edu/manuals/bts.pdf
\(^{35}\) http://www.berlingesturecenter.de/seminare/neurogeselan/neurogeselan.html
\(^{36}\) http://www.dfki.de/~kipp/anvil/
\(^{37}\) http://www.bu.edu/sslrp/SignStream/
\(^{38}\) http://emu.sourceforge.net/
• ISO TC37 SC4 - Language Resource Management: Homepage\textsuperscript{39}
  ◦ ISO Data Category Registry for linguistic concepts\textsuperscript{40}
• GOLD (General Ontology for Linguistic Description)\textsuperscript{41}
• various (lineages of) POS tagging systems such as:
  ◦ the ones assumed in the taggers linked into the Stanford NLP resources page\textsuperscript{42}
• text transcription symbols promoted on the LDC Corpus Cookbook page for transcription/symbols\textsuperscript{43}
• International Standards for Language Engineering (ISLE)\textsuperscript{44}
• ISO Linguistic Annotation Framework (LAF) Overview\textsuperscript{45} and Draft Standard (under revision)\textsuperscript{46}
• XML Serialization for LAF: ISO Graph Annotation Format (GrAF)\textsuperscript{47}
• SoundIndex and related search and view tools developed in the LACITO\textsuperscript{48} Linguistic Data Archiving Project (see Jakobson, Michailovsky, and Lowe, 2001).
• XML Corpus Encoding Standard (XCES)\textsuperscript{49}
• SIGAnn : ACL Special Interest Group on Annotations\textsuperscript{50}
• NXT System
  ◦ Switchboard in NXT\textsuperscript{51}
  ◦ NITE XML Toolkit\textsuperscript{52}
  ◦ Video lecture by Jean Carletta: the NITE XML Toolkit Meets the ICSI Meeting Corpus\textsuperscript{53} tool and framework for building

\textsuperscript{39} http://www.tc37.sc4.org/
\textsuperscript{40} http://www.isocat.org/
\textsuperscript{41} http://linguistics-ontology.org/
\textsuperscript{42} http://www-nlp.stanford.edu/links/statnlp.html#Taggers
\textsuperscript{43} http://projects.ldc.upenn.edu/Corpus_Cookbook/transcription/symbols.html
\textsuperscript{44} http://www.ilc.cnr.it/EAGLES/isle/ISLE_Home_Page.htm
\textsuperscript{45} http://www.cs.vassar.edu/~ide/papers/LAF-LREC06.pdf
\textsuperscript{46} http://www.tc37sc4.org/new_doc/iso_tc37_sc4_N463_rev00_wg1_wd_LAF.pdf
\textsuperscript{47} http://www.cs.vassar.edu/~ide/papers/LAW.pdf
\textsuperscript{48} http://lacito.vjf.cnrs.fr/archivage/description.htm
\textsuperscript{49} http://www.xces.org/
\textsuperscript{50} http://www.cs.vassar.edu/sigann/
\textsuperscript{51} http://groups.inf.ed.ac.uk/switchboard/links.html
\textsuperscript{52} http://www.amiproject.org/showcase/standards-and-toolkits/nite-xml-toolkit-for-annotations
\textsuperscript{53} http://videolectures.net/mlmi04ch_carletta_iab/
• E-MELD (Electronic Metastructure for Endangered Language Data) Homepage
• All of the annotation systems listed on the COCOSDA Corpus Annotation Tools page
• The Automatic Mapping Among Lexico-Grammatical Annotation Models project resources

3.7. References


54 http://emeld.org/index.cfm
55 http://www.ldc.upenn.edu/annotation/
56 http://www.scs.leeds.ac.uk/amalgam/amalgam/amalghome.htm


Alexis Palmer (chair), Debbie Anderson, Eric Kansa, Pavel Mihaylov, Johanna Nichols, Alicia Wassink

4.1. Introduction
As part of the Cyberling2009 workshop at Berkeley, this working group was charged with identifying and documenting existing and needed standards for the digital storage, retrieval, and search of linguistic data. Another concern was the potential for reuse of language data by parties other than the original creators of the data. We present the results of our working sessions as a set of wiki pages, as outlined below.

4.1.1. Process
While standards of some kinds for storage, retrieval, and search of linguistic data do exist in linguistics, many subfields of linguistics talk more about “best practices” and “common practices” than they do about “standards”. We discussed the ways these terms are used on our Big Ideas page. We were not tasked with discussion of the related and important issue of annotation standards. For discussion of this issue, please see Group 1: Annotation Standards.

4.1.2. Summary of Big Ideas regarding data sharing (storage, retrieval and search):
- How do we define ‘other standards’?
- Standards vs. Best Practices
- How do we encourage adoption of standards in linguistics?
  - Data sharing: the publication model (for more on this, see the white paper from Group 4)
  - Standards are great, now how do I use them?
  - How can I participate in the creation of ISO standards?

4.1.3. Issues addressed within the WG2 wiki pages:
- Unicode character encoding standards for increasing stable display, readability, and sharing of data
- Relational database storage
- Wiki-based sharing of research
- Metadata tags for increased transparency and usability of data
- Version control
- Web standards for sharing datasets
- Machine reusability of data (under construction)

4.1.4. Results
- A set of examples/case studies demonstrating applications of standards for storage, retrieval, and search and their utility for linguistic research.
A set of subfield-specific seed lists of common practices, requirements, conventions, etc. The purpose of creating these lists is twofold. First, the lists should be helpful to individual linguists working in the subfield in question. Second, they can serve as reference material to linguists from other areas who might wish to annotate beyond their individual research concern.

- A seed list of existing standards for storage, retrieval, and search of linguistic data.
- A handful of recommendations regarding not-yet-existent but needed standards for linguistics cyberinfrastructure.
- Additional resources: relevant links, papers, etc.

### 4.2. Big Ideas

#### 4.2.1. How do we define ‘other standards’?

For the purposes of this workshop, we take our domain of interest to be standards related to the sharing of language data within the linguistics community. The discussion is organized around a questionable\(^1\) division of data sharing into four subtopics:

- Storage of digital data
- Retrieval and discoverability of digital data (i.e. discoverability at the document or resource level)
- Search of digital data (i.e. discoverability at the within-document level)
- Access and reusability of digital data

---

\(^1\) We call this a questionable division because the four are deeply interrelated.
<table>
<thead>
<tr>
<th>Storage</th>
<th>Retrieval</th>
<th>Search</th>
<th>Access / Reuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata</td>
<td>Metadata</td>
<td>Metadata</td>
<td>Metadata</td>
</tr>
<tr>
<td>Digitization of both primary data &amp; metadata</td>
<td>Versioning tracking major changes / decisions made and motivations for same</td>
<td>Source Materials linking to audio/video source</td>
<td>Citation Standards for how to cite datasets</td>
</tr>
<tr>
<td>Formats &amp; Standards open access standards &amp; formats</td>
<td>Digital Fingerprinting</td>
<td>Annotation Conventions** collection and dissemination of conventions used in existing data collections</td>
<td>Privacy / Legal Issues related to user access, privilege assignment, copyright and data ownership</td>
</tr>
<tr>
<td>Legacy Data providing within-subfield model for best practices data sharing</td>
<td>Adaptive Coding ability to adjust data coding scheme as knowledge evolves</td>
<td>Consistency consistency and quality of data annotations</td>
<td>Subfield-Specific Usability Concerns** specialized standards, metadata sets, ontologies, etc.</td>
</tr>
<tr>
<td>Publication</td>
<td>Stable Addressing of Resources</td>
<td></td>
<td>Reusability repurposing of data for use in addressing new research questions by both humans and machine</td>
</tr>
<tr>
<td>Archiving</td>
<td>Web Standards</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Some of the key areas of concern for each of the four topics we take to fall within our domain of interest. ** indicates action items listed in the table that may be implemented immediately or in the near term.

4.2.2. An aside... Standards vs. Best Practices (a distinction with a difference?)

We weren’t certain that linguistics, as a field or academic culture, has a tradition of clearly differentiating standards from best practices. We do not have an organizational body within the field that sets standards. Subfields are autonomous and vary to the extent that standards or best practices are discussed, named and adhered to. We nonetheless considered a few possible distinctions that might generally be made between the two, so we can be as clear as possible about what we mean when we use the terms “standard” and “best practice” in these wiki pages:

4.2.2.1. Standards

- Often, these are theory-neutral conventional systems for accomplishing some task (often related to analysis, description, or publication) in linguistics (e.g., the IPA system for phonetic transcription)
- Named (so practitioners may name standards to which their practices adhere in published work, for example)
- Official (new standards will explicitly obsolete prior or existing ones) handed down from a high-level organization charged with regulating usage, nomenclature, etc.
- Use is subject to sanction or mandate
Developed over time via a process involving the deliberations of an organizational body of experts, after discussion and consensus

Follow from best practices, ranked and subjected to selection

links: discussions regarding standards

links: political issues

4.2.2.2. **Best Practices**

- Often, principled practices rather than mandated systems for accomplishing some analytical, descriptive, or publication-related task in linguistics
- Recommended, but not strictly enforced
- Generated by practitioners in a bottom-up process, who wish to build consensus in practice and are often interested in motivating the need for a particular practice

4.2.3. **How do we encourage adoption of standards in linguistics?**

4.2.3.1. **Data sharing: the publication model**

Data sharing: the publication model offers some possibilities with regard to building incentives for adopting standards, acknowledging use of annotated corpora, receiving and giving credit for the use of annotated and marked-up data (as a scholarly practice of value to the field). Working group 5 explored ways that other disciplines are sharing data, so we may learn from these examples.

Publication mechanisms for linguistic data collections are one possibility for encouraging adoption of standards.

- Receiving academic credit for publication of data would provide a needed incentive for doing the extra work needed to be sure that standards are followed.
- Peer review will improve the quality of shared data.
- Publication and proper citation of data facilitate demonstrating the scholarly contribution made by providing the data.
- Publication of legacy data would provide a valuable training ground for young researchers as well as providing a model for preparation of data according to best practices.

4.2.3.2. **Standards are great, now how do I use them?**

Widespread use of standards and/or best practices just won’t happen unless it is easy for people to:

- Locate information re: standards and what they entail.
- Learn how to apply the standards to their own data.

Of course, a commitment to communication, collaboration, coordination, community building and open access to data are crucial for supporting the use of standards. Working Group 7 discussed this issue.
4.2.3.3. **How can I participate in the creation of ISO standards?**

ISO is home to a wide array of standards, and the process of standardization can appear to be opaque and daunting to the outsider. A short page devoted to how linguists can participate in ISO standards development is located at: How to Get Involved in ISO Standards Development.² (This page also includes a short section on how to participate in the development of the Unicode Standard.)

By actively participating in ISO standardization, linguists will have a vested interest in the use of such standards. Involvement by linguists also has the result of making sure standards are suited to current needs, and haven’t become fossilized. Ideally, participants should get recognition from their host institution for work on standards development, a job that often requires many hours of time and (at times) considerable personal expense.

4.3. **Case Studies**

The case studies developed by Working Group 2 are intended to serve as illustrations of the following topics as these have arisen and been addressed in specific subfields of linguistics:

- Unicode character encoding standards for increasing stable display, readability, and sharing of data
- Relational database storage
- Wiki-based sharing of research
- Metadata tags for increased transparency and usability of data
- Version control
- Web standards for sharing datasets
- Machine reusability of language data

4.3.1. **Case Study: Interlinearisation**

As an example of real-world application of standards and an attempt to set standards in interlinearisation, we can look at TypeCraft³ (henceforth TC).

From the TypeCraft project home page:

TypeCraft a multi-lingual online database of linguistically-annotated natural language text, embedded in a collaboration and information tool. This set-up allows users (projects as well as individuals) to create their own domains, to invite others, as well as share their data with the public. The kernel of TypeCraft is morphological word-level annotation in a relational database setting, wrapped into a communication system, not unlike popular online community sites. TypeCraft allows you to import raw text for annotation and export annotated data to MS Word, OpenOffice.org, LaTeX or XML for further use.

---

² [http://cyberling.elanguage.net/page/Getting+Involved+in+ISO+Standards+Development](http://cyberling.elanguage.net/page/Getting+Involved+in+ISO+Standards+Development)

One of the primary goals of TC is to provide a standards-based tool for interlinear glossing and data sharing. Some of the standards (existing ones or those developed specifically for TC) in respect to WG2’s expertise are:

**Encoding**
- TC uses Unicode everywhere from internal storage, through user interaction to external export formats.

**Storage**
- The internal storage is a relational database but the external one is XML, with an openly defined DTD (migration to XSD schema planned). This allows us to adjust the internal database in any way we see fit to make the tool better but at the same time keep the external XML format static and allow collaboration with other tools/systems. If we need to extend the XML format, we will strive to preserve existing structures and only add what is new. This ensures backwards-compatibility.

**Retrieval**
- TC is an online tool and every annotated phrase or collection of phrases can be accessed with a single URL. TC is also wiki-based and annotated phrases can be embedded in wiki articles. This makes the data accessible to everyone reading a TCwiki article (think wiki-based papers on linguistics!) and since it’s a wiki, they can be commented on the talk page.

**Metadata**
- TC keeps track of the language of annotated resources by using the ISO 639-3 language codes. The ISO standard is not perfect (language names not always accurate and sometimes the official name for a particular language code is unacceptable to some speakers of the language) but nevertheless ISO language codes are definitely a good start and much better than having users supply language names themselves.
- A fixed set of glossing and part-of-speech tags. This prevents typos when entering tag names (the system will refuse to save anything with misspelled tags) and enforces consistent usage (e.g. singular should be called only SG and not both S and SG). Although in some cases we allow several tags for one and the same grammatical feature to reflect parallel standards.
- Related glossing tags are grouped together into larger virtual tags (e.g. DAT and ABL would be the tags for the dative and the ablative case and they both belong to CASE).
- Rudimentary support for text-level metadata with planned support for OLAC.4

**Access**
- Annotated interlinear glosses can be exported to XML as mentioned above. The power of XML is often underestimated. XML is reusable--a good XML representation can be transformed into many other formats. TC uses the same XML format and different XSL-transformations5 (EXtensible Stylesheet Language) to produce the same data in other formats: HTML (visually appealing to humans when rendered by a browser, can easily be imported into MS Word/OpenOffice.org as part of a paper), LaTeX (to include in a paper). XSLs are

---


powerful enough to combine exported data from TC with data from other systems (e.g. a lexicon tool) to for example produce a lexicon in PDF (via LaTeX), where every entry is illustrated by an example annotated in TC.

- TC is community based, centered around a wiki website and allows for data sharing and collaborative work (groups of people can work on the same texts/phrases).

- Sharing and usage rights:
  - Every annotated phrase can be either private or published. Published phrases are accessible (read-only) to all users of the system.
  - Users of the system can belong to different groups and phrases can be shared with a particular group. All users in that group will have both read and write access to the shared phrases.
  - Future plans: manage usage rights by using access control lists (ACLs). ACLs are a powerful way to restrict or grant rights with fine granularity.

Search

- Searching within TC is consistent because of enforced standardization of metadata. E.g. when a user searches for class markers in Bantu languages (or any other similar morphemes in other languages), he or she does not have to guess what the annotator might have used but can just look through the list of glossing tags and see those that correspond to class markers. The same is valid for languages as all languages are stored as references to ISO codes so it is impossible to annotate for the same language and refer to it with different names.

- TC makes every bit of the annotation searchable. The search is powerful, information from different levels can be combined together and searched for. E.g. all phrases where a NOUN is annotated for both CASE and ANIMACY. This makes it easy to access data on particular linguistic phenomena.

4.3.2. Case Study: Machine Reusability of Data

This page discusses one particular manner in which linguistic data may be re-purposed: as training data for statistical machine learning approaches in computational linguistics and/or natural language processing. More specifically, we’re talking about training data for supervised or semi-supervised methods — methods that learn from labeled data.

4.3.2.1. Linguistic data as training material for machine learning

Goal of this case study: To highlight how decisions related to the annotation and storage of linguistic data (in this case, interlinear glossed texts from a language documentation project) can make the data more or less useful as training data for statistical machine learning methods.

4.3.2.2. What makes data good training data?

The key word here is consistency. Roughly put, the machine model learns generalizations over observed data and uses those to predict analyses for previously-unseen data. In order for it generalize well, the collection of data must be as internally-consistent as possible in the way that it is coded/labeled.

---

6 http://en.wikipedia.org/wiki/Access_control_list
A second important consideration is the underlying data structure. For efficient machine processing, there must be some explicit indication of relations between the text and its annotations. These two points are illustrated below with examples of interlinear glossed text, a common way of representing language data.

### 4.3.2.3. Labeling consistency

**Typographic consistency**

It is important in any data collection and annotation effort that all annotators work from one agreed-upon set of labels. For the sake of the machine learner, it is also important to adhere to capitalization and punctuation conventions. For example, ‘PST’ and ‘pst’ may both be intended to indicate a past tense morpheme, but the machine will see them as two distinct labels. Of course, many such issues can be handled by processing the data post-annotation and pre-model training, but to do so efficiently requires text manipulation skills that those producing the original data may or may not have.

One way for projects to maintain labeling consistency is by use of an annotation interface which restricts the space of allowed labels.

**Analytic consistency**

Maintaining analytic consistency is a much more difficult task. In cases where the analysis is reasonably well-understood at the outset of annotation, agreed-upon conventions for analysis and annotation may be made available to annotators in the form of a detailed annotation manual. It is often the case, however, that analysis and annotation proceed in parallel. In documentation and description of less-studied (or previously unstudied) languages, this is in fact the normal situation. Several bits of record-keeping can help to deal with changing analyses:

- tracking the source of each label (i.e. the specific annotator) as well as the time and date of annotation
- documenting changes in analysis and/or labeling conventions, indicating the nature and source of the change, how the change should be manifested in the annotation (in other words, what did the previous analysis look like? what does the new analysis look like?), the date and time at which the decision to change the analysis was made, and whether or not the change has been back-propagated to previously-labeled data
- using annotation tools and/or data formats which are able to maintain a historical record of changes in the data (along with the metadata associated with those changes)

We recognize that some of these desiderata are not easily attainable with currently-available systems for text glossing and interlinearisation, particularly in the language documentation context. We thus add our voice to those calling for development of an open source, updated, general-purpose system for text interlinearisation and glossing.

### 4.3.2.4. Data structures

First, we point to the pages of WG1: Annotation Standards as well as the Existing Resources page⁷ for many valuable resources pertaining to standardization of data structures for annotation.

---

⁷ http://cyberling.elanguage.net/page/Existing+Standards+and+Technologies
The resources presented on these pages include links to proposed standards and extensive bibliographic references related to this topic.

**Interlinear glossed text (IGT)**

The particular concern in this case study is the use of interlinear glossed text (IGT) as training data for a machine learner. First, here’s an example of IGT from the Mayan language Uspanteko (Pixabaj et al.).

<table>
<thead>
<tr>
<th><strong>Full text</strong></th>
<th>Kita’ tinch’ab’ej laj inyolj iin.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td>Kita’</td>
</tr>
<tr>
<td></td>
<td>tinch’ab’ej</td>
</tr>
<tr>
<td></td>
<td>laj</td>
</tr>
<tr>
<td></td>
<td>inyolj</td>
</tr>
<tr>
<td></td>
<td>iin</td>
</tr>
<tr>
<td><strong>Morpheme</strong></td>
<td>Kita’ t-</td>
</tr>
<tr>
<td></td>
<td>in-</td>
</tr>
<tr>
<td></td>
<td>ch’abe -j</td>
</tr>
<tr>
<td></td>
<td>laj</td>
</tr>
<tr>
<td></td>
<td>in-</td>
</tr>
<tr>
<td></td>
<td>yolj iin</td>
</tr>
<tr>
<td><strong>Gloss</strong></td>
<td>NEG</td>
</tr>
<tr>
<td></td>
<td>INC-</td>
</tr>
<tr>
<td></td>
<td>E1S-</td>
</tr>
<tr>
<td></td>
<td>hablar -SC</td>
</tr>
<tr>
<td></td>
<td>PREP E1S- idima yo</td>
</tr>
<tr>
<td><strong>Spanish Translation</strong></td>
<td>No le hablo in mi idioma.</td>
</tr>
<tr>
<td><strong>English Translation</strong></td>
<td>I don’t speak to him in my language.</td>
</tr>
</tbody>
</table>

*Table 2: Interlinear glossing example.*

**Links between annotation tiers**

The table above shows three tiers of annotation for this Uspanteko clause. The ‘TEXT’ tier contains each word of the clause (word boundaries are indicated by double-line cell borders). The ‘MORPHEME’ tier shows a segmentation of each word into its component morphemes, and the ‘GLOSS’ tier shows a morpheme-by-morpheme gloss of the clause, including both gloss labels for non-stem morphemes (e.g. NEG for kita’) and lemma translations for stem morphemes (e.g. hablar for ch’abe).

Two NLP tasks we might imagine learning from such data are morphological segmentation (producing the ‘MORPHEME’ tier, given at least the ‘TEXT’ tier and perhaps the translation(s) as well) and morpheme glossing (roughly, given the ‘MORPHEME’ tier, produce the ‘GLOSS’ tier). This is where the data structure used to represent the interlinear text becomes crucial!

Most often when we encounter IGT -- as, in fact, in the table above -- the links between annotation tiers are conveyed through visual aspects of the presentation. Here, for example, the association of morphemes with the words they belong to is communicated using double-line borders at word boundaries. Visually-oriented presentations of IGT do not generally provide the explicit encoding of these relationships that a machine learner needs to make sense of the data. In order to use IGT as training data, it must be presented to the machine learner in a format that directly encodes links between elements from one annotation tier to those on another.

**Structured representational formats**

What is needed to address this concern is a format which preserves structured links between annotation tiers. XML formats are one way of preserving said links. At the same time, using XML follows current recommendations regarding longevity and portability of data (for example, Bird and Simons 2003, EMELD School of Best Practices). Several XML formats for IGT have been proposed, including EthnoER’s EOPAS (Schroeter and Thieberger 2006), IGT-XML (Palmer and Erk 2007), and an earlier model outlined in Bow, Hughes, and Bird 2003 (Bow et al.}.
Another approach is the use of Annotation Graphs (e.g. Bird and Liberman 2001, Maeda et al. 2002).

4.3.3. Case Study: Character Encodings

4.3.3.1. Why use Unicode?

Linguistic data should be created and stored using standards. For written text, the character encoding standard is Unicode®/ISO 10646.

Linguists who create text using a non-Unicode font run the risk of jeopardizing their data, for it won’t be easily found by widely used search processes, nor will it be saved in a stable, standardized format that will guarantee longevity.

PDFs, Word documents, webpages, and other documents that are put online but are created using a non-Unicode font will cause problems when doing searching. Using images for missing letters or symbols also present a problem, as in the following example.

The snippets below come from the article “A Preliminary Study of Jaw Movement in Arrernte Consonant Production” by Marija Tabaina (Journal of the International Phonetic Association 2009, 39: 33-51). There are two versions available: an HTML version and a PDF.

The HTML version uses images for certain IPA symbols, including one for the voiceless retroflex stop “ʈ”.

It should be noted that some speaker SI’s productions of the apico-post-alveolar /ɭ/ were of a particular variant observed in some speakers of the language, namely, a strong palatalization before the stop proper, i.e. /ʃə/ is produced as [ʃɛʃ] or [ʃɛʃ], with the ‘retroflex’ quality of the apical consonant not always being clear. However, not all of SI’s productions of /ɭ/ were pre-palatalized; as a result, this speaker’s post-alveolar data should be treated with caution.

Unfortunately, it is not possible to search on the images in the HTML document itself, or by doing a “Google” search across the Internet on such an image, so one won’t be able to locate f “ʈ” in the HTML document.

In the PDF version of the same text, a non-Unicode font has been used for the “ʈ”. At first glance, the “ʈ” appears fine in the PDF:

It should be noted that some speaker SI’s productions of the apico-post-alveolar /ɭ/ were of a particular variant observed in some speakers of the language, namely, a strong palatalization before the stop proper, i.e. /ʃə/ is produced as [ʃɛʃ] or [ʃɛʃ], with the ‘retroflex’ quality of the apical consonant not always being clear. However, not all of SI’s productions of /ɭ/ were pre-palatalized; as a result, this speaker’s post-alveolar data should be treated with caution.

However, the font used has put the glyph for “ʈ” (Unicode LATIN SMALL LETTER T WITH RETROFLEX HOOK, U+0288) in the spot that is properly allocated in Unicode for the DAGGER (†, U+2020). As a result, it is not possible to search in this document for the “ʈ” (Unicode U+0288), because the “ʈ” has been overlaid on the dagger character. (In a similar way, many old Greek fonts would put the lowercase alpha on top of Latin “a” in the font, lowercase beta on top of “b”, etc. To search for alpha, one had to search on the Latin letter “a”. With

8 http://unicode.org/
Unicode, alpha has its own number [codepoint], which is different from the Latin lowercase “a”, so now it is possible to search for alpha separately from Latin “a”.

A non-Unicode font will present other problems for the user: If one copies and pastes the letter “ʈ” from the above PDF into a Unicode-compliant word-processing document, it appears as a dagger.

The above example demonstrates that using a non-Unicode font can prevent search engines from finding documents and text properly.

Another important factor is that old documents created with non-Unicode fonts will be hard to read in the future, since they won’t be based on an international standard. If an elderly linguist were to die and leave his data on his computer (which he had keyed in with his own non-standard font), it may take considerable time and effort to convert the data into a standardized format that is usable by others, with the possibility that the data could be lost forever.

4.3.3.2. Tools (Fonts/Keyboards/etc.)

The listing below is not comprehensive, but is only intended to provide a few reliable webpages that provide tools and other useful information on Unicode-enabled products.

Fonts

Most core fonts that come with recent operating systems are Unicode-based, but they may not include all the special characters required by linguists. (Note: A new font that will be released with Windows 7, Ebrima, will include improved support for many African languages that use the Latin script. It also include the Vai and N’ko scripts.)

- Unicode-enabled fonts with IPA from SIL: http://scripts.sil.org/IPAhome (Note: SIL fonts that are prefixed with “SIL IPA” are non-Unicode fonts)
- John Well’s webpage with a listing of IPA fonts (and Unicode input info):
  http://www.phon.ucl.ac.uk/home/wells/ipa-unicode.htm
- LinguistList info on fonts: http://linguistlist.org/sp/Fonts.html

Character pickers

These enable users to select the letters or symbols they wish, and cut and paste them into documents.

- http://people.w3.org/rishida/scripts/pickers/

Keyboards and Inputting Methods

- Keyboard info from SIL
- Keyboards and Inputting Tips from E-MELD
- IPA Keyboards for the PC from the Speech, Hearing, and Phonetic Sciences Dept., University College London: http://www.phon.ucl.ac.uk/resource/phonetics/

9 http://scripts.sil.org/UniIPAKeyboard
10 http://linguistlist.org/cfdocs/emeld/school/classroom/unicode/ipafont.htm
Other Script Standards and Useful Resources:

- ISO Script Codes (ISO 15924)\(^\text{11}\)
- Recommendations on the Development of New Orthographies:\(^\text{12}\) this is a set of guidelines for linguists who are devising orthographies, so the orthography can be accessible to users on computers.
- Unicode for Language Documentation:\(^\text{13}\) An informative page from E-MELD with sections on (a) adding characters to Unicode, (b) precomposed forms (letters made up of a base character + one or more diacritic, and why these are not in Unicode), and (c) IPA and Unicode.
- How to Get Involved in ISO Standards Development:\(^\text{14}\) This document also includes a short section on how to get involved in the development of the Unicode Standard.

4.3.4. Case Study: Web Standards

4.3.4.1. Design styles for effective use of the Web

The Web has a distributed architecture that sees application in a vast array of domains. Linguistics specific data sharing systems should follow “best-practices” for Web design so that linguistics data can be used effectively in conjunction with datasets from other disciplines.

RESTful\(^\text{15}\) design patterns are useful for multidisciplinary data sharing and using the Web as a publishing platform. Some important elements of RESTful design include:

- Data are located at specific addresses (URIs).
- Data can have different representations for different purposes. For instance, a resource may have a “human-readable” representation (a Web-page with attractive formatting for use in a browser) versus “machine-readable” representation such as XML for easy parsing by software.
- Getting data is very simple and only requires following a hyperlink to a specific address. There is no need to send more complex message to obtain data.
- In the same vein, there is a very limited set of general actions that can be performed on data. The most common is retrieving data (“GET”). Other verbs that can be performed include: Creating a wholly new resource (“PUT”), creating / updating a subordinate resource (“POST”), and deleting data (“DELETE”).

The simplicity of RESTful designs helps account for the explosive growth of the Web. However, some efforts at data sharing and design of shared infrastructure in the sciences and in industry deviate from RESTful design principles. These other efforts added layers of complexity (requiring additional verbs or other complex messages to interact with data, requiring “state” to be tracked in data exchanges, etc.). As a result, some cyberinfrastructure and enterprise systems

\(^{11}\) http://www.unicode.org/iso15924/

\(^{12}\) http://www.unicode.org/notes/n19/

\(^{13}\) http://linguistlist.org/cfdocs/emeld/school/classroom/unicode/documentation.html

\(^{14}\) http://cyberling.elanguage.net/page/Getting+Involved+in+ISO+Standards+Development

\(^{15}\) http://en.wikipedia.org/wiki/Representational_State_Transfer
have been more expensive to develop and maintain. These more complex systems are also harder to extend, use in unintended applications, and harder to make interoperable.

REST isn’t really a standard, but more of a design style. In addition, there are few official standards for designing REST styled Web-services (systems designed for the exchange of machine-readable data). From the perspective of building a cyberinfrastructure where multidisciplinary data-sharing is a goal, it is probably more important adhere to RESTful design styles for data retrieval than other operations (creating, updating and deleting data). In other words, if one wants to make linguistic data available for a multidisciplinary community, these data should be made through no more fuss and bother than simply following hyperlinks. (Note: there is nothing about REST that precludes security and authorization systems. Such systems can and do work perfectly well in RESTful systems).

4.3.4.2. Atom Standards and REST

Multi-disciplinary research may require mixing of data and services in ways that cannot be easily anticipated. Making this easy to do experimentally would be useful. This requires styles of service design (such as RESTful architectures) that lower costs and barriers to entry and use.

- **Atom Syndication Format**\(^\text{16}\) (widely implemented). Atom can serve as a convenient “standard container” for more specialized XML payloads (such as an XML document expressing interlinearised annotation of a text). Atom’s simple standard metadata that can make more specialized XML payloads more intelligible. The Atom standard is well designed and it can be extended to include additional metadata (most commonly with GeoRSS, a standard for expressing geographic data). Atom can be used as a common format to express the results of queries (as feeds, with feed entries as records). It can be extended to support more specialized applications.

- **Atom Publication Protocol**\(^\text{17}\) for updating and contributing to a collection (implemented by the SWORDS\(^\text{18}\) project). SWORDS lets you deposit content into a repository without worrying about what kind of repository it is.

4.3.4.3. Examples of Atom and REST

The screenshot below illustrates why REST and Atom-based web services are useful. The following example comes from aggregating data from the “Portable Antiquities Scheme,”\(^\text{19}\) an online database of antiquities found by the public and registered with museums and heritage organizations in the UK. The Portable Antiquities Scheme has a service that expresses the results of queries as a feed. The feed has geographic metadata expressed as GeoRSS.\(^\text{20}\) This feed is combined with a similar feed from Open Context.\(^\text{21}\) Yahoo Pipes,\(^\text{22}\) a feed manipulation and processing service, was used to combine the feeds from these two sources.

---

20 [http://www.georss.org/Main_Page](http://www.georss.org/Main_Page)
The above example is of interest because the Portable Antiquities Scheme and Open Context have very different underlying data structures, vocabularies, and schema. Yet data from these two sources can be aggregated to a limited, though still useful, extent. In Open Context’s case, specialized XML data (using the ArchaeoML global schema) for each record is available. Open Context’s Atom feeds point to these ArchaeoML data, making the specialized data easily available for more sophisticated applications than illustrated here.

The main point of this discussion is that useful and meaningful cooperation across different data sources (potentially across disciplinary boundaries) is possible even when using very simple common standards.
4.3.5. Case Study: Sociolinguistics Version Control

<table>
<thead>
<tr>
<th>Subdiscipline</th>
<th>Sociolinguistics</th>
</tr>
</thead>
</table>
| **Project Title**      | Dialect Evolution and Ongoing Variable Linguistic Input: English in the Pacific Northwest 200 Years After Lewis and Clark  
                          National Science Foundation Award: BCS-0643374  
                          (Alicia Beckford Wassink, Principle Investigator, University of Washington, Department of Linguistics) |
| **Software**           | Microsoft Sharepoint Server 2007 (for version control and remote collaboration) |
| **Goal of this Case Study** | Demonstrate the use of versioning software in an online research collaboration area (ORCA) to register the changes made to spoken language recordings and associated data to enable tracking of modifications, and make transparent the nature of and motivations for the changes |

4.3.5.1. Project description (from the project homepage23)

The Pacific Northwest English project investigates the features of English spoken in the Pacific Northwestern region of the United States (PNW), two hundred years after the introduction of non-indigenous speakers to the region. The Pacific Northwest English (PNWE) project explores the extent of English dialect development in the Pacific Northwest region of the United States. It also documents the stories of families with deep roots in the Pacific Northwest region. The project outcomes include: acoustic analysis of vowel speakers of speakers indigenous to the PNW (formant data, amplitude and duration measures), orthographic transcriptions of 200 hours of unscripted speech, detailed social network measures and analyses for judgment sample speakers. Sociolinguistic investigation includes analysis of speaker participation in several phonological changes affecting North American English, examination of questions related to interethnic contact and dialect focusing across three generations. Oral histories for participating families are being produced in conjunction with the Washington State Historical Society, and Prof. Jean Harris, Highline Community College.

4.3.5.2. Data Elicitation

- Elicitation of data involves the utilization of a hybrid methodology, combining phonetic analysis with a standard multi-part variationist sociolinguistic interview schedule allowing collection of data in different spoken registers (unscripted conversation, one-on-one interview, reading passage, word lists, semantic differentials, syntactic diagnostic prompts).

- While we cannot make all elicitation instruments available in this wiki (to avoid exposing materials to potential respondents), similar elicitation materials are publicly available in the Elicitation Materials Clearinghouse,24 Sociolinguistics Laboratory, University of Washington.

- Two-part sample includes data recorded in the field for a judgment sample and in the laboratory using telephony devices to acquire data for a complementary random sample.

---

23 http://www.artsci.washington.edu/nwenglish/index.asp

4.3.5.3. Storage

- Original recordings (recorded at a 44.1kHz sampling rate in uncompressed form to compact flash media, using M-Audio MicroTrack digital flash recording devices) are stored in four locations (as required by IRB protocols): 1) on a file server subjected to regular, incremental backups, 2) in ISO-9660 formatted compact disks, locked in a cabinet accessible only to the principal investigator, 2) in redacted form on compact disks in a CD archive, 3) in redacted form on a file server in an online research collaboration area (Microsoft SharePoint ORCA).

- Redacted formats have been edited in Praat software for the removal of potentially identifying subject information, i.e., the acoustic signal has been attenuated to zero, while leaving the time dimension intact. This allows all versions of the sound files to retain original timings, enabling location of temporal events of interest across versions of the recordings and transcriptions (which have been time-stamped based upon the non-redacted versions of the signal).

Version control

- Version control is provided via an online research collaboration workspace created using Microsoft SharePoint 2007. SharePoint runs on any platform (Researchers in our team are currently using MAC, Windows and Linux operating systems). Versioning is particularly useful in the document libraries where sound files, transcriptions, and Praat text tiers are stored.

Figure 6: Screenshot showing organization of ORCA main page.

- Version control requires (in this case, although other versioning software varies) that each user check out a sound file or transcript from the document library. SharePoint allows only one user to check out a file at a time, but other software (such as CSV, Subversion, etc) does not have this limitation.

- The file is modified by the user.

- At the end of a work session, the user uploads the modified version of the file to the document library. The software prompts the user to provide comments regarding what changes were made to the document, and automatically timestamps the new file with the upload time and version (Figure 7, Figure 8). Figure 7 shows the SharePoint pulldown for a file called
SR2CF2A_non-conversational. This is an orthographic transcription file that exists in several versions because it has been subjected to a process of anonymization. We desire to view the version history for this file. Figure 8 shows the version history for this file. The current iteration is version 6, which has been wiped of information that potentially identifies a study participant. Clicking on any version (from 1-6) will result in a prompt by the system to view or restore that iteration.

- Crucially, all prior versions are available to the user. This allows full control and comparison of different versions of the documents stored in the library without overwriting data.
- A discussion area within the ORCA allows discussion of substantive changes to collection, analysis, and other protocols so that important decisions may be registered as part of the project history.
4.3.5.4. Metadata

- Akustyk software is used for associating project, speaker and token level metadata with events in the sound file.
- A project handbook registers methodology and decisions made.
  - The metadata associated with all recordings is here²⁵

4.3.5.5. Access

- Sharepoint allows for restriction of access depending on permissions criteria for each member of the research team. It is possible, in principle, to share redacted versions of the recordings with all members of the team with data analysis functions, and restrict access to the non-redacted versions to the PI. Permissions criteria are set by principal investigator.
- The public face of the project includes: 1) the project website, 2) exemplifying sound files that may be played out or downloaded from maps on the project website, 3) individuals and organizations may download data files for particular speakers from the project website, for the set of speakers who have consented that their materials be made available in this way (see Human Subjects consent form sample²⁶).

Versioning software

- Concurrent Versions System (CVS): An open-source revision control system (http://www.nongnu.org/cvs/)

---

²⁵ http://cyberling.elanguage.net/page/Sociolinguistics+metadata+tags
²⁶ http://www.artsci.washington.edu/nwenglish/jointhestudy.asp
• Subversion: An open-source revision control system (http://subversion.tigris.org/)
• Microsoft Sharepoint Server, 2007 (http://sharepoint.microsoft.com/Pages/Default.aspx)

Benefits of utilizing version control software
• Very easy to use
• Allows research team to avoid the pitfall of saving numerous copies of the same file(s) in the same, or worse, different locations, and having to remember those locations.
• Allows research team to keep track of the current version of a working sound file, spreadsheet (containing acoustic measures and demographic data in this case) between different users and/or different machines
• Offers the capability to revert to earlier versions of some or all of the files in a given workspace
• Some versioning software (e.g., Subversion), offers the ability to merge work done on the same file by different users
• Members of the working team located remotely may all access common elements in the same workspace when they *do* talk together (e.g. video or teleconferencing); and contribute without the risk of overwriting each other’s work

Considerations
Researchers will have to make value judgments about exactly what changes or decisions are meaningful to them, as well as to potential users. This means that different types of users will find versions of the data useful to differing degrees. In addition, version comments will always be somewhat subjective. Some subfields may have developed best practices regarding comments useful for versioning.

Considerations for different types of user (assuming resource uses a graded permissions structure):
• external users accessing a large corpus may not need the comments offered or content of different versions. They want to know the version number.
• researchers picking up the corpus to use for addressing new research questions are the group most likely to benefit from access to versions of the data (to know what version of the data they are working with, see the structure and format of this dataset as opposed to earlier versions, understand changes made to the resource after they have used it).

We want version control to be easy for the researcher, and useful. Once we begin to use such a tool, there may be more information than some researchers need, but version control INCREASES the utility of the data resource for others partly because it provides a minimal amount of metadata for a data resource (when the resource in its current form was produced, by whom and what the iteration of the version is). If a researcher has even a small set of basic principles for judging what types of comments are meaningful, supplying version comments can require little user effort. In short, a little goes a long way.
4.3.6. Case Study: Genealogical Classification with AutoTyp

<table>
<thead>
<tr>
<th>Subdiscipline</th>
<th>Typology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Title</td>
<td>Autotyp typological databases</td>
</tr>
<tr>
<td>Software</td>
<td>FileMaker Pro™</td>
</tr>
<tr>
<td>Goal of this Case Study</td>
<td>Demonstrate the use of notes and a log to track changes in the genealogical classification in the database.</td>
</tr>
</tbody>
</table>

See slideshow: jn_cyberling_reanalysis_trail.ppt. (The genealogical classification can also be accessed online at http://uni-leipzig.de/~autotyp).

This is a series of screenshots of records in the Autotyp typological database. One of the modules there is a genealogy for every language we have in the database (about 2600 languages so far). By now the genealogy amounts to a near-complete list of all the world’s language families and most of their major subgrouping. It is under constant revision as fieldwork and comparative work discover new languages and families, join families into larger groups, and otherwise change classification. The literature is also filled with speculative proposals about genealogical relationships that do not have adequate support. Keeping track of the revisions is not particularly burdensome and is absolutely essential for documenting the grounds for the classification. (Unlike most other classifications of the world’s languages, ours is based on evidence and criteria for non-chance degrees of resemblance, so every decision does need to be described.)

We have a Notes field on all of our genealogical classification records, and this is used to record the history of editing and reclassifying. The slides in this slideshow show:

Slide 3. The levels of classification we consider determinate. Stock and language entries are required for every record.

Slide 4. The search interface as seen for a language search from the on-line search interface.

Slide 6. Sample language record for Arabic, with its classification at different levels. Note that Afroasiatic, though a bona fide family, is not a stock but a higher-level grouping (demonstrable but not reconstructable).

Slides 8-15 (see commentary slides 7 and 14): These are a set of screenshots of records for languages of the Penutian macrogroup. They document our changing decisions on the status of this group and the subgrouping of some of its components.

Slides 16-17 Page from the database log outlining the changes made to Penutian in 2008.

Slides 17-18 Slide showing evidence and counterevidence presented in Notes comments, on the subclassification of Western Malayo-Polynesian.

4.4. Subfield-specific Practices

4.4.1. Sociolinguistics

While there has been little in the way of broad-based “standards” setting in the field of sociolinguistics, there are common practices that have been developed in research laboratories.

http://attachments.wetpaintserv.us/eDqDdED%2BQ8HcePckwBsq4w%3D%3D1680384
over years of community-based research in variationist and interactional sociolinguistics, course materials that have provided a foundation for graduate level training, and online curricular resources prepared and shared by individual practitioners who have chosen to make their materials and protocols available to the wider community. This page is intended to provide a partial, expandable list of such resources. There is also a growing list of published writings on best practices. These links are shared not as standards, but as aids for those interested in best practices (or just common practices) around data sharing, storage, and retrieval in sociolinguistics.

4.4.1.1. **Recommended Readings:**

4.4.1.2. **Software:**
Akustyk:28 Free, open-source vowel analysis software package. It installs as an add-on to the popular software, Praat.29 Provides relational database functionality for storing project, speaker, and token level metadata.

4.4.1.3. **The members of the WG2 were aware of common practices/best practices regarding:**
- Ordering of elicitation tasks in a conventional variationist sociolinguistic interview
- Standards for metadata tags for associating sociodemographic data with audio files30
- Conventions for labeling sound files
- Data analysis time estimations (time required to orthographically transcribe one hour of recorded speech, analyze one vowel at midpoint; summarize the demographic information contained in some database)

4.4.1.4. **Needed standards and data-sharing resources in Sociolinguistics:**
- Establishing a common, widely-used set of metadata tags for associating sociodemographic data with audio- and video-recordings that will render data useful for a range of types of sociolinguistic analysis addressing a range of research questions and may “scale up” in usefulness for other types of linguistic analysis
- Standards and models for the sharing of recordings, social data (assuming adherence to required human subjects protection protocols), transcriptions (orthographic, IGT, phone-level) and measurements drawn on that data (e.g. to support cross-language, cross-dialect study or language change research)
- What is data? Understanding the layering of the notion of data (or the continuum from data to products of research including full length or excerpted audio/video-recordings, transcriptions, written texts, measures, pages of IGT, summarized data, primary vs. secondary data).

30 [example: http://cyberling.elanguage.net/page/Sociolinguistics+metadata+tags](http://cyberling.elanguage.net/page/Sociolinguistics+metadata+tags)
What needs to be protected? What may be copyrighted by the researcher? What belongs to the community? How do we serve community interests so that research has broader impact in both scholarly and lay communities

- Levels of representation of the social data as distinct from the linguistic data

- Training in how to conduct appropriate inferential statistical tests for the range of different data structure types (from phonetic data, to syntactic data, to subjective reaction/attitudinal data)

- Training in generating human subjects applications that will enable sharing of data within the wider research community (at different levels within an institution or more broadly in the field, and in the lay community, as appropriate)

- Versioning practices that will enable tracking changes to all types of data and associated products of research, and which tracks the motivations for changes made to audio or text files

- Online clearinghouses for elicitation materials, research instruments, tools, and recording device configurations for various types of study. These should, at a minimum, be supplied with information regarding proper crediting of the originator of the tool, and instructions regarding how to cite the tool in a bibliographic record (example at UW Sociolinguistics Laboratory)
  - Praat scripts
  - Commutation tests
  - Reading passages
  - Word lists

4.4.1.5. **Needed standards for data-analysis in sociophonetics**

- Best practices for recording data in formats that will ensure sufficient fidelity for acoustic analysis of various kinds (amplitude, pitch, formant frequency, jitter, shimmer, duration)

- Representational conventions vis-à-vis what qualifies as narrow vs. broad transcriptions

- Standards regarding inter-measurer and inter-transcriber reliability (verification of measures)

- Training in use of determining appropriate inferential statistical tests for data structure types

- Greater transparency with regard to documenting transcription conventions (word class categories and memberships, explanations for use of phonetic symbols and diacritics).

4.4.2. **Typology**

4.4.2.1. **Common practices:**

Several institutions and projects have large databases with cross-linguistic typological information on a number of languages. (List with links to be added.) The information has generally been gathered over some years from published grammars, one’s own fieldwork, consultations with experts, etc. and represents an extremely large outlay of time and resources. In some cases language data is solicited in return for authorship credit for that portion of the database. Ways that database owners might provide peer review for these contributions have occasionally been discussed but probably never implemented.
Several projects make some or all of their data available, via a search interface or as downloadable data files. There are no field-wide standards for content of fields, data format, etc. but this does not seem to be an obstacle to finding and using data. It is standard to link data to language names using ISO codes. There is a growing consensus that data is most useful if it is free of one’s own lumpings (thus, e.g., not “large/medium/small” for inventories of elements such as morphological cases or phonemes, but the actual number; not “VO/OV” but an actual specific basic word order).

4.4.2.2. **Needed practices:**
Consensus on how to give database authors and owners credit for their work while also making data publicly available. Practices include: Make data publicly available and request a citation. Offer data on request in exchange for citation. Offer data on request in exchange for coauthorship. (All three models are used in other fields.)

Peer review for databases as a whole. (This is separate from the question of how the database owner obtains peer review for individual contributions to the database.) Accuracy of entries and usefulness, comprehensiveness, empirical and theoretical adequacy, etc. of data categories need review.

Appropriate credit (on one’s CV and with one’s employing institution or prospective employer) for creation and maintenance of databases.

4.4.3. **Language Documentation**

4.4.3.1. **Introduction**
The fieldworker (and this includes documentary linguists) needs to learn the structure of the language and produce a grammatical description, dictionary, and texts usable by researchers and others (including interested members of the speech community). He or she is usually also singlehandedly responsible for producing theoretical and comparative work on the language. Not much will be said here about producing grammars (though see Dryer 2006) or the general scholarly work, other than to note that the whole set of responsibilities gives the fieldworker a considerably heavier workload than the average for linguists, so tools and standards that save time are especially needed. The tasks in need of standards and tools for sharing and access are: recording and/or digitizing; transcribing; creating texts; creating a lexicon; and archival storage and access of recordings, texts, and lexicon (and perhaps other materials).

4.4.3.2. **Recommended Readings:**

4.4.3.3. **Existing standards, common practices, and/or best practices:**

*Recording*
Information available from the DoBeS project\(^{31}\) of the Max Planck Institute for Psycholinguistics, Nijmegen; the Hans Rausing Endangered Languages Project,\(^{32}\) SOAS, London; and others.

\(^{31}\) [http://www.mpi.nl/DOBES/documents](http://www.mpi.nl/DOBES/documents)
Transcribing
Many field linguists use Transcriber, available from SourceForge.net.

Text work
The main tasks are inputting or importing transcribed material, morphological interlinearising, syntactic annotation, parsing, lemmatization, concordance building, and preparation for corpus searching and archiving. For standards for some kinds of interlinearisation and annotation see the report of Working Group 1. To my knowledge there are no current widely used software tools that could be described as best practice, though there is a growing consensus about what they need to do (e.g. multiple annotation tiers; link to recordings; enable non-interlinear annotation for discontinuous, non-compositional, multi-word, and non-linear categories and functions; enable printout of text segments with standard publishable interlinears; enable corpus searches of all kinds; enable other access). There is also very little knowledge of what might go into theory-neutral syntactic annotation.

A list of tools dated 2004 and including some text tools is available from the Melbourne School of Engineering. I am aware of these updates and additions: Kura and a glossing tool under development by Thomas Mayer (presented at ALT8, July 2009). The DoBeS project has elaborate and fairly specialized tools for text and dictionary work.

Dictionary compilation
The most common practice for compiling dictionaries of all kinds (descriptive, defining, etymological, etc.) seems to be use of commercial database software to create a self-standing database that does not, e.g., link to a text corpus.

Archival storage and access
Archives have their own standards for metadata and data formats. The DoBeS project has an extensive list of archives with links.

4.4.3.4. Needed standards and data-sharing resources
Easy-to-use tools for text work are badly needed. Basic research, leading ultimately to standards, for theory-neutral syntactic annotation is needed.

32 http://www.hrelp.org/archive/resources/index.html
33 http://sourceforge.net/projects/trans/
35 http://www.kura.ats.lmu.de/index.php
36 http://lsa2009.berkeley.edu/alt8/
37 http://www.lat-mpi.eu/tools/
38 http://www.mpi.nl/DOBES/language_archives
4.5. Existing Standards

4.5.1. Storage

4.5.1.1. text encoding standards
- Unicode\(^{39}\) (properly a character encoding standard)
- Text Encoding Initiative (TEI) P5 Standard\(^{40}\) “These guidelines make recommendations about suitable ways of representing those features of textual resources which need to be identified explicitly in order to facilitate processing by computer programs. In particular, they specify a set of markers (or tags) which may be inserted in the electronic representation of the text, in order to mark the text structure and other features of interest.”

4.5.1.2. (domain-specific) terminological standards
terminological standards (e.g., the GOLD ontology\(^{41}\))

4.5.1.3. storage and retrieval standards
- repository systems (e.g. eSciDoc\(^{42}\) and Fedora Commons\(^{43}\) might be relevant in relation to long-term archiving)
- DELAMAN\(^{44}\) is ‘as international umbrella body for archives and other initiatives with the goal of documenting and archiving endangered languages and cultures worldwide. Our aim is to stimulate interaction about practical matters that result from the experiences of fieldworkers and archivists, and to act as an information clearinghouse.’
- The Rosetta Project\(^{45}\) is another archive actively promoting a set of best practices for storage of language data.

4.5.2. Retrieval

4.5.2.1. reference/identification standards (i.e. metadata)
- ISO language codes\(^{46}\)
- ISO script codes\(^{47}\) (ISO 15924)
- OLAC metadata\(^{48}\)

\(^{39}\) http://unicode.org/
\(^{41}\) http://linguistics-ontology.org/
\(^{42}\) https://www.escidoc.org/
\(^{43}\) http://www.fedora-commons.org/
\(^{44}\) http://www.delaman.org/
\(^{45}\) http://www.rosettaproject.org/
\(^{47}\) http://www.unicode.org/iso15924/
\(^{48}\) http://www.language-archives.org/tools.html
• Dublin Core\textsuperscript{49} metadata standard

4.5.2.2. Citation Standards

• COINS\textsuperscript{50} (a simple standard for embedding Dublin Core citation metadata in a web page)

4.5.3. Search

• Open Search,\textsuperscript{51} A very simple standard for sharing search results, usually by expressing such search results in the Atom Syndication Format.\textsuperscript{52} Although this is not the best standard (in terms of design, extensibility), it is relatively easy to adopt. Open Search also has proposed geographic extensions to describe how to query a collection based on geographic parameters.

4.5.4. Access/Reuse

4.5.4.1. Cultural Heritage Global Schema and Ontologies

• CIDOC\textsuperscript{53} (an ontology mainly applied by European museums and other heritage organizations that is nicely abstracted and very generalized, but is complex and has some difficulties in application)

• OCHRE/ArchaeoML\textsuperscript{54} (a somewhat more simple global schema / ontology for cultural heritage applications, including archaeology, epigraphy and philology. It is highly abstract so that projects and collections retain native descriptive terminologies but some degree of interoperability and shared services are facilitated.

4.5.4.2. Copyright and Intellectual Property

Creative Commons\textsuperscript{55} provides a series of standard copyright licenses and associated metadata to explicitly give certain permissions and conditions for use/reuse of copyrighted content. These are useful to define how content can be used. However, these are complicated to apply with scientific data, since US copyright law makes a distinction between “facts” (ideas, concepts, objective data) and “expressions”. Since many scientific datasets contain factual measurements and observation, they may not be protected by copyright. To make matters more complicated, the determination of what’s a fact and what’s an expression is ambiguous and a blurred distinction. This legal ambiguity and complexity makes it harder to use and reuse scientific data. Therefore, Creative Common’s scientific arm, “Science Commons,”\textsuperscript{56} recommends that scientists do not use Creative Commons copyright licenses for scientific data. Instead, Science Commons recommends that scientific application explicitly dedicate data to the public domain using the

\textsuperscript{49} http://dublincore.org/
\textsuperscript{50} http://ocoins.info/
\textsuperscript{51} http://www.opensearch.org/Home
\textsuperscript{52} http://www.atomenabled.org/developers/syndication/atom-format-spec.php
\textsuperscript{53} http://cidoc.ics.forth.gr/
\textsuperscript{54} http://ochre.lib.uchicago.edu/
\textsuperscript{55} http://creativecommons.org/
\textsuperscript{56} http://sciencecommons.org/
“CC-Zero” declaration. CC-Zero removes legal ambiguity around data, removes all restrictions for reuse, and in theory, maximizes the scientific value of data.

4.5.4.3. **APIs/standards for interfaces with other resources (e.g. corpora, lexica/lexical resources, treebanks)**

WordNet “This document presents a standard conversion of Princeton WordNet to RDF/OWL. It describes how it was converted and gives examples of how it may be queried for use in Semantic Web applications.”

4.6. **Needed Standards**

4.6.1. **Storage**

4.6.1.1. **Standard Software interface to contribute to repository systems.**

SWORDS (based on the Atom Publishing Protocol) lets users and other systems to contribute content to different repository systems. The linguistics community should look into building off of SWORDS for linguistics specific needs.

4.6.2. **Search**

4.6.2.1. **RESTful ways lower barriers to entry, but standards for REST-style architectures need more development**

Standards for Web-services are most developed for (WS/SOAP) style web services. These types of web-services see extensive application in enterprise, government, and some cyberinfrastructure environments (typically where there is a high-degree of centralization). The standards for SOAP web services are highly developed. However, these are more difficult to scale in distributed, multi-organizational, and multidisciplinary environments. REST-style approaches would be better for cross-disciplinary cyberinfrastructure. However, REST style approaches suffer from the lack of more highly developed standards (as described below).

4.6.2.2. **RESTful ways for repository systems to declare how they can be queried.**

The Atom Syndication Format can be very useful for expressing results of queries. Atom entries can be good containers to specify where individual records can be retrieved (and these individual records can be expressed in discipline specific XML formats). However, it will be good if an Atom feed could describe in a machine-readable way how its source collection can be queried (what are the query parameters available, what values can these query parameters take, etc.). To a limited degree, the Open Search standard accomplishes this, but this standard is mainly geared toward simple full-text searches (like Google). It would be better to use a better designed more general standard than Open Search.

---

57 http://wiki.creativecommons.org/CC0
58 http://www.w3.org/TR/wordnet-rdf/
59 http://www.swordapp.org/
60 http://atompub.org/rfc4287.html
61 http://en.wikipedia.org/wiki/SOAP
62 http://www.opensearch.org/Home
Also, sometimes it may be more useful to obtain the summarized results of a query more than the individual records. I’ve been thinking about this with my work on Open Context. One of the things I find useful as a service is to expose faceted metadata in machine readable formats (I’m using Atom). Exposing metadata like this provides a simple way for the system to define how it can be queried, as well as offering useful summary information about the contents of the collection. But, in the case of Open Context\(^63\), no widely adopted standard exists to support these kinds of capabilities.

4.6.3. Access/Reuse

4.6.3.1. **standards for using and developing basic natural language processing (NLP) tools, or for creating training data from existing linguistically-annotated data**

This may be getting a bit ‘out there’, but I’m thinking about the extreme prevalence of statistical machine learning methods in computational linguistics and the value of annotated linguistic data for training tools using such methods. Are there ways of making this data more easily adapted to be training data for developing NLP tools?

4.7. Resources

4.7.1. Encoding and Annotation Links:

4.7.1.1. **Autotyp**
- http://uni-leipzig.de/~autotyp/

4.7.1.2. **ISO (http://www.iso.org/)**
- My ISO Job: Guidance for Delegates and Experts (accessible document describing ISO and how experts can participate)
- Joining In: Participating in International Standardization (another easy-to-read document on participating in ISO standards. development)
- Standards Development (webpage on standards development with links)
  - http://www.iso.org/iso/standards_development.htm

ISO Committees of interest to linguists
- TC 37/SC 4 - Language resource management
  - http://www.tc37sc4.org/
- TC 37/SC 4 Working Groups
  - http://www.tc37sc4.org/working_group.htm
- TC 37/SC 2 - Terminographical and lexicographical working methods

\(^63\) http://opencontext.org/
• JTC 1/SC 2/WG2: Universal character set
  ◦ http://std.dkuug.dk/jtc1/sc2/wg2/

ISO Script Codes (ISO 15924)
• http://www.unicode.org/iso15924/

4.7.1.3. **OLAC**
• http://www.language-archives.org/OLAC/metadata.html

4.7.1.4. **TypeCraft**
• http://www.typecraft.org/

4.7.1.5. **Unicode**
  
  *Unicode code charts*
  • http://www.unicode.org/charts/

  *Fonts*
  • http://scripts.sil.org/IPAhome (Unicode-enabled fonts with IPA and other resources)
  • http://www.phon.ucl.ac.uk/home/wells/ipa-unicode.htm (John Wells’ website on fonts with IPA and Unicode info)
  • http://linguistlist.org/sp/Fonts.html (LinguistList info on fonts)

  *Character pickers*
  • http://people.w3.org/rishida/scripts/pickers/
  • http://weston.ruter.net/projects/ipa-chart/view/keyboard/

  *Keyboard and Input Methods*
  • http://scripts.sil.org/UniPAKeyboard
  • http://linguistlist.org/cfdocs/emeld/school/classroom/unicode/ipafont.htm (Tips on how to assign a keystroke for input and how to create a keyboard for character input)
  • http://www.phon.ucl.ac.uk/resource/phonetics/ (Unicode phonetic keyboard for PC with SIL fonts)

  **Recommendations on developing new orthographies**
  • http://www.unicode.org/notes/tn19/

  *Unicode for Language Documentation*
  • E-MELD document with useful information for linguists on Unicode:
    ◦ http://linguistlist.org/cfdocs/emeld/school/classroom/unicode/documentation.html

4.7.1.6. **XSL-transformations (EXtensible Stylesheet Language)**
• http://en.wikipedia.org/wiki/XSL_Transformations

4.7.2. **Data Collection Instruments Links**
• Elicitation Materials Clearinghouse (Sociolinguistics elicitation instruments, U of Washington, Sociolinguistics Laboratory)
  ◦ http://faculty.washington.edu/wassink/Brown+Bag/Elicitation+materials.htm
• Typological tools for field linguists (questionnaires, etc., Max Planck Institute for Ev. Anthro., Leipzig)

4.7.3. Metadata Tagging Links
• Akustyk
  ◦ http://bartus.org/akustyk/
• Sociophonetics metadata tags
  ◦ http://cyberling.elanguage.net/page/Sociolinguistics+common+practices

4.7.4. Web-Design Links
• RESTful design patterns (REpresentational State Transfer)
  ◦ http://en.wikipedia.org/wiki/Representational_State_Transfer

4.7.5. Storage Links
• SWORDS (based on the Atom Publishing Protocol)
  ◦ http://www.swordapp.org/
  ◦ http://atompub.org/rfc4287.html

4.7.5.1. Version control software
• Concurrent Versions System (CVS): An open-source revision control system
  ◦ http://www.nongnu.org/cvs/
• Subversion: An open-source revision control system
  ◦ http://subversion.tigris.org/
• Microsoft Sharepoint Server, 2007
  ◦ http://sharepoint.microsoft.com/Pages/Default.aspx

4.7.5.2. Infrastructure for Long-Term Archiving
• eSciDoc collaborative eResearch infrastructure
  ◦ http://www.escidoc.org/
• Fedora Commons Repository Software
  ◦ http://www.fedora-commons.org/

4.7.6. Sharing Links
• WS/SOAP (standards for web services)
  ◦ http://en.wikipedia.org/wiki/SOAP

4.7.7. Access Links
• Wiki access control lists (for assignment of usage rights and privileges)
  ◦ http://en.wikipedia.org/wiki/Access_control_list
4.8. Bibliography


Maeda, Kazuaki, Steven Bird, Xiaoyi Ma, and Haejoong Lee. ‘Creating Annotation Tools with the Annotation Graph Toolkit’. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC).


Pixabaj, Telma Can (coordinator), Miguel Angel Vicente Méndez, María Vicente Méndez, and Oswaldo Ajcot Damián. Uspanteko text collection, in Text Collections in Four Mayan Languages, 2003-2007. OKMA (Oxlajuuj Keej Maya’ Ajtz’iib’), Supported by Endangered Languages Documentation Programme (SOAS, University of London).

Schroeter, Ronald and Nicholas Thieberger. 2006. ‘EOPAS: the EthnoER online representation of interlinear text.’ In Sustainable Data from Digital Fieldwork (proceedings of conference held at the University of Sydney, 4-6 December 2006). Sydney University Press.

Bill Byrne (co-chair), Robert Forkel (co-chair), Dorothee Beermann, Arienne Dwyer, Florian Jaeger, Jeremy Kahn, Virach Sornlertlamvanich

5.1. Introduction

The Tools group is charged with identifying and documenting existing and needed tools which will be the face of the cyberinfrastructure for ordinary working linguists. These tools include both those used by data creators (e.g., linguists annotating data that they later share) and data consumers (e.g., linguists using the annotated data of others to create new kinds of data).

5.2. Existing Tools

- TypeCraft:1 Collaborative text annotation
- WALS:2 The World Atlas of Language Structures Online
- ODIN:3 Online Database of INterlinear glossed text
- TextGrid:4 “TextGrid aims to create a community grid for the collaborative editing, annotation, analysis and publication of specialist texts. It thus forms a cornerstone in the emerging e-Humanities.”
- Natural Language Toolkit5 (NLTK): “Open source Python modules, linguistic data and documentation for research and development in natural language processing, supporting dozens of NLP tasks, with distributions for Windows, Mac OSX and Linux.”
- eHumanities Desktop:6 (project is in alpha development stage, no description available yet)
- Roma TEI validation tool:7 “These pages will help you design your own TEI validator, as a DTD, RELAXNG or W3C Schema.”
- Chorus8 is a version control system designed to enable workflows appropriate for typical language development teams who are geographically distributed. Chorus is a Palaso Project.
- e-Linguistics:9 building a cyberinfrastructure for linguistics (including a Python toolkit for data migration; documentation is still being posted)
- Consistent Document Engineering Toolkit10

______________________________
1 http://www.typecraft.org/tc2wiki/Main_Page
2 http://wals.info/
3 http://www.csufresno.edu/odin/
4 http://www.textgrid.de/en.html
5 http://www.nltk.org/
6 http://hudesktop.hucompute.org/
7 http://www.tei-c.org/Roma/
8 http://projects.palaso.org/projects/show/chorus
9 http://purl.org/linguistics/e-linguistics
Thai language specific tools

- SWATH\textsuperscript{11} Thai word segmentation and POS tagging
- CU Thai word segmentation\textsuperscript{12}
- CU Thai Romanization\textsuperscript{13}
- iSpeech:\textsuperscript{14} Automatic speech recognition toolkit
- Vaja:\textsuperscript{15} Thai/English text-to-speech synthesis engine
- LIBS:\textsuperscript{16} String kernel based language-script-encoding (LSE) identifier for 85 LSEs
- KUI:\textsuperscript{17} Collaborative WordNet translation platform
- Web service and Open API\textsuperscript{18} for Asian WordNet, Multilingual Morphological Analyzer, and Thai Soundex

5.2.1. **General purpose tools in use by linguists:**

- R Project\textsuperscript{19} for statistical computing and the linguistics packages in EMU\textsuperscript{20}
- Praat:\textsuperscript{21} doing phonetics by computer
- Python\textsuperscript{22}
- ANVIL:\textsuperscript{23} video annotation research tool

5.3. **Needed Tools**

- A FOS aligner tool (or aligner development tool) at a grain finer than the intervals marked off fairly automatically in LDC’s transcriber tool

5.4. **What’s a killer application?**

Google Maps may be a good example for a killer app:

\textsuperscript{10} http://www.icsi.berkeley.edu/~jan/projects/CDET/
\textsuperscript{11} http://www.cs.cmu.edu/~paisarn/software.html
\textsuperscript{12} http://www.arts.chula.ac.th/~ling/wordseg/
\textsuperscript{13} http://www.arts.chula.ac.th/~ling/tts/
\textsuperscript{14} http://www.hlt.nectec.or.th/products/ispeech.php
\textsuperscript{15} http://www.hlt.nectec.or.th/products/vaja.php
\textsuperscript{16} http://www.tcllab.org/libs/
\textsuperscript{17} http://www.tcllab.org/kui/
\textsuperscript{18} http://www.tcllab.org/chumpol/sndws/api_regis.php
\textsuperscript{19} http://www.r-project.org/
\textsuperscript{20} http://emu.sourceforge.net/
\textsuperscript{21} http://www.fon.hum.uva.nl/praat/
\textsuperscript{22} http://python.org/
\textsuperscript{23} http://www.anvil-software.de/
• it’s killer in the way it brought mapping data to everyone.
• it actually killed, e.g. gml - at least gml’s hope for mass adoption.
• it didn’t piggyback on a standard, but set one: kml - and it turned out, creating xml files isn’t that much of a problem, if you want it badly enough.

But:
• can there be something like micro-killer-apps?
• can there be something like scientific killer apps? doesn’t “scientific” mean “too small to be killer”?

Following the Google Maps example a killer app would help pull data out of the drawers. This might happen in two ways:
• Make publishing data easier or
• provide big enough incentives to submit to tedious publishing.

5.5. What could killer apps for linguistics look like?
• search engines? or the semantic web (see this blog post\(^{24}\) for an idea of what this could mean)?
• data visualization?
• can “archiving” or “long-term preservation” be a killer app? (Does not sound like it - does it.)
• is reproducible research enough of an incentive to publish data?

may the killer app be something social/political - like a new model for scientific recognition on the web? and if so, what can we do to bring it about? Foster skills?\(^{25}\)

5.6. Killer applications are applications that are used lots and lots
Therefore a good question might be: Who are the linguists interested in finding and/or producing reusable data?
• computational linguists
• corpus linguists
• typologists
  ◦ encyclopedic works like the SIL guides are already interesting and useful, and there is growing interest in sharing data and linguistic ontology
• descriptive linguists
• theoretical linguists
  ◦ much theoretical work does not currently use reusable (or computationally-accessible) data, with some exceptions [e.g. the LFG and HPSG communities].

\(^{24}\) http://intern.blogs.mpdl.mpg.de/2009/07/07/lord-linked-open-research-data/
\(^{25}\) http://sgillies.net/blog/819/second-guessing-project-bamboo/
But data for the computational linguist is probably not quite the same as data for the typologist (or the theoretical linguist).

Likewise a killer app for a computational linguist is probably something very different from an application that a descriptive linguist, engaged into field work, would care to call a useful tool. Theoretical linguists might be interested in searching for data along yet another set of dimensions. Finally the generation of reusable resources, if considered important at all, must pay off academically to attract more than the occasional linguist. Perhaps we can conclude from this that we rather need a cluster of tools than this one application — together they might be a killer.

So following the definition above ("killer apps are apps that are used a lot"), we can probably assume that future killer apps will be on the web.

There seem to be two concerns here:

- what does ‘data’ look like for each field, and can we share specifications?
- what does an ‘application’ look like for linguists of various stripes?

These concerns need not pit “computational” linguists against other types of linguist.

5.7. **What does ‘data’ look like for each field, and can we share specifications?**

As any science linguistics is based on data, yet the form this data takes and the role it plays crucially depends on the way we perceive of language and the particular approach chosen in investigating its nature. Does that then mean that there is no such thing as “the empirical base of our field”. Not necessarily; it only means that this base must consist of a multitude of different types of linguistic data. If so, free access to and reusability of this data might be a commodity that is found useful by most of us.

Let’s assume we could agree on that point, what exactly does that mean for future linguistic tools?

All seems to come back to the same point, namely that we are chasing a ghost by looking for that one killer app; instead what we most likely need are several different tools, able to cater to the multitude of needs that define the linguistic field as a whole.

5.8. **Desirable Characteristics of Apps**

- No dead ends for data: While some apps (e.g. filemaker) may be “killer” in how they help organizing data, they also make reusing the data hard.

Peter Austin (co-chair), Tracy Holloway King (co-chair), Kurt Bollacker, Martin Haspelmath, Koenraad de Smedt, Paul Trilsbeek

6.1. Abstract

In this paper we discuss what data provenance and data reliability are, with special attention to the needs of the linguistics community when moving from simple data sets to more complex cyberinfrastructures. The integrity and completeness of provenance information as a basis for citation, rights management, etc. is crucial for those who record, annotate and compile materials, but also for the sources of raw materials, in particular for indigenous and minority communities, and for the scholars who use these materials as a basis for further scientific work. We suggest some first steps to promote data sharing and publication in the linguistics community.

6.2. Data Provenance

Provenance is the who, what, and when of metadata. When a data set is created, it is important to know where it comes from and who is responsible for its publication. Adequate information about provenance, i.e. information about how, where, when and by whom the data was collected, encoded and annotated, and who assumes responsibility for its publication, allows the quality of the data set to be assessed, provides a contact in case of questions, and establishes authorship of the data set, similar to the authorship of an academic paper.

The contents and status of a data set are not always clear from a quick inspection of the data. The data might come from native-speaker informants, from published works of literature, from the web, etc. Adequate metadata are needed for cataloguing data sets so as to make them searchable, and to secure that the data will be used in an appropriate way. Metadata should include a detailed description of the data set, including sufficient information on its provenance. In some cases provenance information should even be provided for separate sections of the data set wherever differences are relevant. For example, the data may have been collected over several years of field work; knowing when each section was collected and by who and under which circumstances may be important.

Provenance is extremely important to the scientific community which uses data sets as the basis of further analysis, hypothesis testing, etc. For the user, provenance is crucial assessing whether, and to what extent, a data set can form an appropriate basis for subsequent scientific work. Knowing how the data was collected and processed and who was responsible for it allows future users of the data set to judge the quality of the data, how it fits in with their research and how it compares to other data sets. Provenance is also important for the replicability of data and research results. Unless it is clear where the data came from and how it was created, data sets cannot be replicated. For example, if a data set has constituency trees annotated over it, it is important to know whether the trees were manually constructed, created automatically, or bootstrapped by manually correcting automatically constructed trees.

For the creators, provenance provides a way to get credit for the scientific work done in creating the data sets and allows the community to cite the data sets with their authors in academic works. Creating a high quality data set is extremely time consuming and requires highly skilled linguists: as such, it is important that those doing this work get credit for their contribution. Provenance is
also important in establishing and maintaining privacy rights for those who provided the data. There are many reasons that data may not be appropriate to publish freely in its entirety, e.g. individuals may be identifiable from videos or from the content of discussions, or private rituals may be recorded. In particular the rights of indigenous and minority communities should be respected and acknowledged.

Provenance tightly linked to the proper, standardized use of metadata and documentation in reliable environments. All data must be tagged with the appropriate metadata and linked to its documentation. This allows researchers to understand what is, and is not, included in the data set and its annotation and to correctly cite the data set, thereby acknowledging the work of the data set creators and allowing future researchers to reproduce their findings. Metadata and documentation is facilitated by adherence to standards and best practices. As students may not be familiar with these, more senior researchers and experts in natural language data curation need to share their knowledge and to facilitate students adherence to the accepted standards.

6.3. Achieving Reliable Provenance Through Publication

A major question is how to achieve reliable data provenance in the linguistic community and promoting the sharing of data. Creating data sets is a time consuming, highly skilled task and the scientific community as a whole needs to acknowledge this and to provide support for those working on creating and maintaining data sets. Individuals contributing to and creating these data sets need to get institutional credit for data publication. For example, these should count for tenure reviews and other review processes and should be an integral part of grant applications, both in having data set publication be part of the grant work and in the granting agencies favoring researchers who publish curated data sets, just as they favor those with a proven track record of traditional academic publications.

Given the nature of the linguistics community and the working paradigms that they are used to, we suggest promoting curated data sets as publications. The technology is currently available to treat curated data as publication. There is extensive archival work on linguistic data sets, including the work done by organizations such as the Linguistic Data Consortium. There are also examples from other scientific fields where publication of data collection has become an established scientific practice. However, there needs to be extensive institutional and social engagement in order for curated data as publication to become the norm in the linguistics and language studies communities.

Researchers need to be encouraged to publish curated data sets. This can be done in part by requiring it at the institutional level or as part of a grant reward. It can also be aided by providing more infrastructure to publish data sets, including providing information on best practices and on how to access the necessary technologies. In addition, researchers need to cite the published data sets that they used in their research. Reviewers and publishers of articles and books should reject submissions if they do not cite the data sources they used in their research. These citations should be standardized by using the metadata and the publishers of the curated data sets should facilitate this by providing information as to how the data set should be cited.

The linguistic community will also need to provide support to ensure annotation and data quality control, just as it does with published academic papers. Not all data collection and annotation is done equally well and is of equal value. The community needs to have a way to recognize and
acknowledge this, similar to the relative value of different books publishers, journals, and conference proceedings.

Publication of data through a recognized publication channel with an ISSN would make it easier to cite the data uniformly and correctly, would clearly establish authorship, would allow for different editions, would enforce the use of standards and metadata and would a catalyst for giving academic credit to the makers. It would also promote reviews and rating systems.

In a different field, the data journal Earth System Science Data\footnote{http://www.earth-system-science-data.net/} promotes the rapid publication of research on original data sets. Their policy is outlined as follows:

Articles in the data section may pertain to the planning, instrumentation and execution of experiments or collection of data. Any interpretation of data is outside the scope of regular articles.

In the first stage, papers that pass a rapid access peer-review are immediately published on the Earth System Science Data Discussions (ESSDD) website. They are then subject to Interactive Public Discussion, during which the referees’ comments (anonymous or attributed), additional short comments by other members of the scientific community (attributed) and the authors’ replies are also published in ESSDD. In the second stage, the peer-review process is completed and, if accepted, the final revised papers are published in ESSD. To ensure publication precedence for authors, and to provide a lasting record of scientific discussion, ESSDD and ESSD are both ISSN-registered, permanently archived and fully citable.

It should be noted that articles in ESSD seem to contain the data mostly within the articles themselves. A journal is, however, not the same as a cyberinfrastructure. A data publication is meant to announce to the community that data has been available and how it was obtained and annotated, and also to give credit where it is due. The actual data can be accessed in a variety of ways, not necessarily through the same channel. Since publication is by nature public, there may be some issues with restricted data. Full metadata could be published, but the metadata could stipulate restrictions on the accessibility of the actual data (e.g. due to proprietary data or privacy issues).

One could imagine organizations acting as the publishers of articles with their data sets. Publishers of data will have the responsibility of checking at least the formal aspects of published data, such as proper use of metadata, adherence to standards etc. It might be possible that the same data is published at different hosts. We would need buy-in from institutions and linguists to support this. Publishers would themselves be rated and would need to actively advertise their data publications and make them attractive to researchers. Peer review of data publications should be stimulated. Perhaps language resources would be published unedited first, then reviews can be added on later, where annotation might count as a type of review; by having the language data out, even in unedited form, people would be encouraged to annotate it. Many academic credit systems (e.g. in Norway) require full peer reviewing as well as the use of recognized academic publishing channels.
6.4. Persistence and Fine-Grained Provenance Information

Given that linguists change institutions and that URLs shift over time, it is important that future researchers be able to access the same data that is being used today and to be certain that this is the same data as was used by other researchers. On the Internet, the assignment of provenance information to a piece of information can be assured through the use of a Persistent Identifier (PID). PIDs are globally unique identifiers that remain the same even if the URL of a resource changes. Central PID resolvers are used to administer the locations and additional metadata information of the resources. One PID can be used to refer to multiple identical copies of a resource in different locations. Examples of PID systems are the Handle system\(^2\) and the DOI system.\(^3\) PID systems only work though as long as the administration of the resource links is maintained, so the assignment of PIDs alone does not guarantee the long-term stability of the resource references. The linguistic community might consider setting up a registration authority for linguistic data.

Every entity involved in data set creation can be identified by a unique PID. These include entities such as people, organizations, and their roles, the data sets and documents themselves, and different views and mashups of the data. Added value by annotations etc. of existing material can be managed by cascading PIDs with different rights. A proliferation of provenance information may increase the size of the information by orders of magnitude but disk space is cheap and so we can do this very fine grained, e.g. sound clips indicated by start and stop time in a speech corpus. Assigning full provenance information can however be complicated, e.g. when material is translated or when value is added to copyrighted material (e.g. Wall Street Journal corpus) or when a speech corpus has a radio broadcast in the background.

Tomorrow’s cyberinfrastructures will not be limited to static storage of data, but will be dynamic systems that process and present data according to user’s needs. This context will present special challenges to handling provenance information. The following dynamic functionalities can be considered:

- customized presentation of data: filtering, reformatting, style sheets
- pipelined processes
- mashups: the integration of data from various sources (sometimes combining various modes)

In contrast to paper materials, which are static and pre-edited, a cyberarchive allows (and should allow) the user to participate in filtering and presenting information (“play editor yourself”). An example is the Wittgenstein archives at Bergen,\(^4\) which contains digitized manuscripts: the user can choose to include or exclude certain pieces of information and has options for visualization. This creates challenges for provenance. How would you cite a specific view among many other possible views of this material? Such reference may be possible by generating a unique URI and PID for the transformed and formatted web page. An “I want to cite this” button would make this process easy for the user. Dynamic customized views could help with scientific collaboration.

\(^2\) http://www.handle.net/

\(^3\) http://www.doi.org/

\(^4\) http://wab.aksis.uib.no/transform/wab.php?modus=opsjoner
since it is a good way to bring things together at the presentation level; the tools for this are very useful since most linguists do not have the user interface design skills to do this themselves.

The annotation layered on curated data is often done in conjunction with specific software. Having “software as a service” available to the linguistics community can aid in this process. It will be particularly valuable for institutions with less extensive computing infrastructure, allowing their researchers access to state of the art data set curation facilities. New frameworks such as UIMA\(^5\) can aid in the interactive pipelining of processes on data. This again is a challenge for provenance information. Ideally, a PID can be assigned to every step in the pipelining process; note that at every step, intermediate data could be worth storing as a new resource.

Furthermore, Rosetta, Freebase,\(^6\) the Internet Archive, etc. allow for mashups of data. Cyberinfrastructures should provide mashup functionality, i.e. the smart combination of data from various sources. Provenance is a challenging for mashup data since the mashup process is dynamic and every combination of specific versions of data produces a new mashup version. Also, some people might want to reference the particular rendering of a mashup. Referencing views could easily escalate when every user can have a personal view.

6.5. Reliable Identification, Authorization and Rights Management

As part of provenance, in recording the who, what, and when of metadata, it is necessary to have trusted identification of individuals, organizations, and services. This identification needs to persist over time so that decades after a data set is created, it is still possible to determine who created it and how it was created. Reliable identification of individuals is an important prerequisite for at least two purposes:

• Identifying authors and sources to give credit to data creation.
• Identifying users to provide authorization based on licenses and rights.

CLARIN\(^7\) has stated that some system with global e-identities can solve many problems associated with the current situation of people having different usernames at the various sites they use. A single logon should identify the user in an easy way. This is a largely solved problem on technical level, but various solutions are available and none are widely used by the community:

• OpenID\(^8\)
• Federations of local e-identity providers

Authorization is a tuple linking a set of rights, the identity of a piece of data and the identity of a user. Authorization can take various forms depending on restrictions, e.g. access can be given to anyone, can be based on email domains (i.e. affiliated institutions), can require the acceptance of

\(^5\) http://uima-framework.sourceforge.net/
\(^6\) http://www.freebase.com/
\(^7\) http://www.clarin.eu/
\(^8\) https://www.myopenid.com/
a user license, etc. dependent on what the author or stakeholder decides. There are many such restrictions, e.g., access may be given for non-commercial purposes only, or access to sacred songs can be restricted to the initiated, or subparts of data may be proprietary and requires additional agreements. There is a need to train people on what types of rights and access are appropriate. There may be privacy issues with sources such as hospital patient data, sign language data, etc. which may place restrictions on availability.

Some language data, especially spontaneous speech and sign language, cannot be distributed due to privacy issues, in particular in utterances referring to people. becomes very complex with international access, where different countries may have different rules for guarding privacy. This situation may require different country-specific licenses, so international cooperation may need legal advice from the start.

Some technology exists to anonymize source materials, e.g. by masking, manual and automatic, e.g. making non-words out of words, but keeping POS; media is harder to deal with than strict text, especially sign language where need much of the exact original data; any encoding of a facial expression is going to be on the edge. Perhaps 3D models or avatars could be useful, but often not automatic and sometimes not possible, e.g. in sign language.

At the start of the DOBES program, legal specialists advised keeping all data closed. Since that was not an option, a code of conduct was worked out which provides a workable solution. Yet, some data may need to remain closed to all but core researchers in a tightly controlled project. One could use information from original institution waivers to guide what permissions to assign to the data. The CLARIN project has a working group on IPR and licensing issues. The European parliament has expressed an interest in solving the complexity of the issues and may want to promote a revision of legislation, hopefully leading to wider availability of data for research purposes. Some lobbying towards legislators may be necessary.

Researchers are often unwilling to turn over their data for storage and distribution in repositories. One reason is that some people feel their data is not ready yet: once data is in a repository they feel it is cast in stone, which is a real problem for data which is never quite complete such as a dictionary of a living language. It is therefore crucial that repositories offer versioning and updating of the stored materials. Some researchers might prefer to control distribution themselves from their own homepage. A possible solution could be that links or web pages could be generated from the repository automatically. This could also be useful for university administration which has to document research production. A soft approach to repositories could help lead people to understand what a data repository is and what it can start to enable.

A special situation may occur when someone data is retracted, either for privacy reasons or because of errors or other reasons. One could mark data as deleted, invalid or superseded without actually destroying the data; one may also want to be able to temporarily restrict data for certain reasons.

\[9 \text{ http://ttl07.uib.no/papers/6.pdf}\]
\[10 \text{ http://www.mpi.nl/DOBES}\]
6.6. Suggested First Steps

Curated data as publication and the corresponding data provenance and reliability could involve a major, long-term infrastructure project for the linguistic community. We would like to suggest a few simple first steps that the community as a whole can take. Our basic suggestion is to provide both carrots and sticks to the community and to pursue proactive education in data set publication. Although some linguists will become major contributors to curated data sets while others will play a more minor role, all linguists should understand and appreciate their importance: no linguist left behind. It will therefore be useful for the linguistic community to engage in extensive dissemination and training efforts and to establish links with ongoing generic projects on metadata standards and preservation (e.g. PREMIS\textsuperscript{11}).

Encouraging students is not enough since the actions of successful researchers speak much more loudly than words. Therefore, if well-known, successful members of the linguistic community publish and share their data sets, this will set a powerful example for the next generation. Those who do publish their data should provide a simple “cite as” button with data to make it easy for those using the data to cite it in their works. This is particularly necessary since many linguists are unsure of the proper way in which to cite such data sets. Finally, it would be good to provide a service provision for data structure and integrity validation and for format conversion. Data structure and integrity validation can, at a minimum, check that the relevant metadata is in place before publication. As a more complex process, it can check over the data structure to be sure that it is correct and complete, e.g. a tab delineated file should have tabs between fields. A format conversion service would allow data set creators to publish their data sets in a variety of formats, both making it easier for researchers to use them and helping ensure the long term survival of the data as formats change.

There are also some institutional safeguards that can easily be put into place to ensure that data publication occurs and is done properly. First, publishers, editors, and reviewers can require provenance information for all data sources before agreeing to publish research work. Second, editors and funding agencies encourage data sets to be published. This is already starting but should be further encouraged. Also, the funding agencies should ensure that the data is in fact published if it was part of the grant and if possible should publicize that the data is now available. If grantees do not publish their data as promised, either the funding agency could withhold any future grants to the grantee and its institution or could not pay out the final part of the grant until the data is published, similar to how some dissertation fellowships reserve part of the grant until the signed dissertation is submitted.

Finally, the establishment of electronic data publishing journals in conjunction with a cyberinfrastructure should be considered, so as to provide a formal channel for establishing authorship of data sets and creating a scholarly reference in addition to a framework for peer review.

\textsuperscript{11} http://www.loc.gov/standards/premis/
7. **White Paper: Working Group 5**

*Scott Farrar (co-chair), D. Terence Langendoen (co-chair), Steven Moran, Cornelius Puschmann, Dwight van Tuyl*

7.1. **Abstract**

This document presents the findings of Cyberling 2009: Working Group 5 which was charged with reviewing how other fields have been successful (or not) at implementing a cyberinfrastructure. First, we present the rationale for having such a working group to begin with. We then turn to a brief introductory summary of the successes of key related fields. Next, we discuss how the following attributes and practices work to comprise the best recipe for a cyberinfrastructure: (1) having no central data store, (2) using appropriate conceptualizations, (3) providing flexible visualization tools, (4) being open and (5) providing easy ways to cite materials. We then turn to a description of collaborative modularity, our philosophy on the ideal organizational structure for the field. Finally, we present particular technologies and practices that we found to be compelling and instructive for the field of linguistics.

7.2. **The Charge and Rationale**

The field of linguistics appears to be well equipped as a model science for the 21st century. Linguists are by nature users of technology and progressive. Linguistics is one of the first disciplines to use the Web to further its activities and build its community (cf. Linguist List, founded in 1989, and present on the Web since the mid-1990s). We as a field are users of advanced instrumentation to collect and analyze our data, and have produced a variety of software to assist our everyday work, e.g. Shoebox, ELAN, and Praat. On the other hand, linguistics is challenged in specific ways, and these have caused a slow adoption of many recent technologies that could help advance our science. For instance, the field is by its nature theoretically speculative and eschews the confines of standards and normative pressures. As a result we find ourselves with a lack of infrastructure and standards. This is particularly evident when it comes to our views on data and data providers. Linguists on the whole have not actively tried to make their data widely available to others, either because they have not seen the need to, not wanted to, or not known how to. Those who have made their data available have generally made little effort to link their data with the data of others. There are exceptions to this situation of course. In terms of standards, there are the IPA, Unicode, and ISO/Ethnologue language codes. Linguists have a rich tradition of citing data ‘snippets’ such as example sentences, and are beginning to recognize the importance of publishing data, even without an accompanying analysis. Still, linguists do not generally agree on a core conceptualization of linguistics (e.g., the basic inventory of feature types, language varieties, etc.) which makes it difficult to link large amounts of diverse data.

At Cyberling 2009 held in Berkeley, CA, our working group was charged with reviewing how other fields had been successful (or not) at implementing a cyberinfrastructure and handling their data. We met with a plan to discuss the current state of affairs for linguistics, especially where we could compare linguistics with other fields. We decided to choose a few particular fields and to discuss how they have approached their own cyberinfrastructure needs. We organized our discussion by first enumerating the key facets of the approaches that contribute to success. We also discussed bad examples of cyberinfrastructure from other fields. A good portion of the
discussion centered around tools that have contributed to success. Such was the process that led to the current public statement (this document). We chose to explore the best facets of cyberinfrastructure from various fields (not necessarily choosing a specific field or tool to copy) including facets of the following categories:

- Data handling practices
- Collaborative/organizational structure
- Internet technology and tools

The remainder of this document explores each of these categories in hopes that the field as a whole can learn from the efforts of other fields.

7.3. Brief Overview of Related Fields

We begin by giving a few examples of how other fields have handled their own cyberinfrastructure needs, showing what can be achieved when researchers, educators and technical support staff organize around the goal of enabling world-wide data collection, distribution and analysis. By “related fields” we simply mean those with data. Ordinarily, one would expect a group of linguists to look towards sister disciplines in the humanities or social sciences. For our charge, however, we chose to look much further afield, in particular, at the hard sciences and at current practices on the Web. We decided that these fields were more representative of the path that we feel leads to success and one which we believe linguistics should ultimately follow. Here is a summary listing of exemplary fields from our broad survey/discussion:

- The Integrated Public Use Microdata Series (IPUMS) https://international.ipums.org/international/, based at the University of Minnesota, provides users with free access to interoperable data from 130 censuses from a total of 44 countries, representing nearly 280 million anonymized person records. It enables researchers to analyze and visualize the world’s population in time and space in order to study social change and human impact on the environment.

- The Worldwide Protein Data Bank (wwPDB)¹ is an international group of organizations that serve as protein data deposition, processing and distribution centers. It maintains a single PDB archive of macromolecular structural data that is freely and publicly available to the worldwide community. The database contains approximately 50,000 proteins that researchers have discovered and shared openly with the global community. The associated infrastructure makes it possible to visualize and analyze these molecules, permitting further progress in many fields and applications such as engineered drug design.

- The Consortium for the Barcode of Life,² sponsored by the Smithsonian Institution and the Sloan Foundation, is an international effort to develop reliable means for the identification of biological species. Barcoding uses a short DNA sequence in an organism’s genome as a barcode equivalent to determine the species of a biological sample. Adoption of a barcode-

¹ http://www.wwpdb.org/
² http://barcoding.si.edu/
format standard allows for a sample, whether in a museum or field collection, to be instantly linked to related information resources worldwide.

- The National Virtual Observatory (NVO)\(^3\) is a partnership of US universities, observatories and federal agencies, to develop interactive visual portals to the sky, and is a founding member of the International Virtual Observatory Alliance (IVOA) http://www.ivoa.net/. These resources enable data from radio, optical and X-ray astronomical observatories and archives to be aggregated, compared and analyzed, providing opportunities for new discoveries about the nature and origins of the universe.

### 7.4. Data Handling Practices

In terms of how data are to be handled (processed, curated and searched) we find several aspects of other fields very compelling. The first is to have no central data store. By ‘central data store’, we refer to a single location (or site) at which all of the field’s data are housed and maintained. This means that data providers must adhere to only the central authority’s recommended practices and formats. It means that the data are centrally managed. Whereas such an approach has its merits (uniformity and quality control), research in linguistics and the data that are produced are too diverse for such a top-down approach. As articulated in the open paragraphs of this document, such an approach would likely require a common conceptualization for the field, something not yet in achievable.

However, without any conceptualization at all, it would not be possible to group any data together in the first place. That brings us to our second point: the necessity of having appropriate conceptualizations of the field. By ‘conceptualization’ we refer to terminologies, such as the elements used in the Leipzig Glossing Rules (http://www.eva.mpg.de/lingua/resources/glossing-rules.php) or the ISO/Ethnologue language codes (http://www.ethnologue.com), and to ontologies, as in the General Ontology for Linguistic Description (GOLD) (Farrar and Langendoen, 2003). We certainly do not advocate any one terminology or ontology, but do advocate the field’s awareness of such resources and their further development. Only recently have our journals, funding agencies and professional societies started to require the use of such resources. We see it as absolutely advantageous for all such organizations to recommend the use of standard terminologies and ontologies.

Terminologies and ontologies for data may not satisfy all users. One way to alleviate some of the resistance towards their use is to provide services on their benefits, e.g., flexible ways to visualize data. That is, even if one linguist uses a particular set of terminology, another linguist can view the same data using a different set. Visualization also concerns the way in which data are arranged for human consumption. Viewed in various ways, the same dataset can provide many types of information and potentially lead to a better understanding. For instance, consider Google Sky a service that combines astronomical data from different observatories. It provides a very user friendly, visually pleasing, and scientifically accurate way to access data from world-class observatories such as the Hubble Space Telescope, GALEX space-base Ultraviolet observatory, the Chandra space-based X-ray observatory, and others. The point is that the same data can be visualized according to a number of scientific dimensions, and such is a key selling point for using standard terminologies and ontologies.

\(^3\) http://www.us-vo.org/
Related is the concept of open data. For data, ‘openness’ means that anyone can access them. For many of the fields we surveyed (e.g., astronomy), there was no issue with opening up the data to the general scientific community or to the public. However, for some, the medical domain in particular, data are incredibly sensitive. These fields have, nevertheless, come up with very clever ways to publish their data. Often, this involves anonymization: the removal or masking of names, ages, addresses, etc. While anonymization may not be appropriate for all kinds of linguistics data (e.g., field data on endangered languages), it does provide one means by which data can be published.

Whether data are publicly accessible or not, they should be easy to cite and assign credit to. The field of geoscience, for instance, has created a state-of-the-art search facility for open datasets (see http://pangaea.de). Not only do users have access to the data, but datasets are given along with Digital Object Identifiers (DOIs) a scheme that allows for unambiguous citation.

In summary then, these aspects of linguistics data handling have the greatest potential for success:

- no central data store
- appropriate conceptualizations of the field
- flexible visualization of data
- open data (but sensitive to access restrictions)
- easy to cite (DOIs, URIs)

7.5. Collaborative/organizational structure

One of the most challenging aspects of creating a cyberinfrastructure is to accommodate collaboration while fostering individual innovation, all within some sort of organizational structure or structures. To this end we introduce more of an overall approach than an actual solution, one which we refer to as collaborative modularity. This approach concerns how projects are developed with an eye towards future collaboration. In the past most linguistics projects that could be regarded as digital were developed “in-house”. For example, consider the many on-line linguistic databases (e.g., WALS and various other typological efforts). Excellent resources as they are, very few can be searched within a common framework or be accessed with a common API.

A particularly instructive domain to illustrate benefits of the principle of collaborative modularity is astronomy. For this field there are three distinct models for a cyberinfrastructure. The first is NASA’s Extragalactic DB (http://nedwww.ipac.caltech.edu/) which is an older model that compiles numbers and names from the scientific literature. It uses a traditional database and presents data to the user in a traditional format. In most respects, this project is bottom-up and is well grounded in data that are actually written about. On the other hand, the project focuses on “legacy data” and, then, attempts an a posteriori merging of data to achieve meaningful comparisons. The second is the Sloan Digital Sky Survey,4 a more top-down approach which compiles data from a single observing instrument (more or less). What is interesting about this project is that it is managed, complete with a scientific council and spokesperson, governing body, and publication policy. The project by its very nature encourages interoperation with more

4 http://www.sdss.org/
than 150 scientists having contributed to this common cause. The pay-off for contributing data
includes wide dissemination of data and the benefit of the many tools usable with the available
datasets. Finally, there is the secular project Google Sky\(^5\) put together by Google, astronomers
and the University of Washington. It provides a very user friendly, visually pleasing, and
scientifically accurate way to access data from world-class observatories, including the Hubble
Space Telescope, GALEX space-base Ultraviolet observatory, the Chandra space-based X-ray
observatory, and others. It is an exemplar in display, but it is extraordinary because of the
amount of data that are made interoperable to achieve a singular effect.

The state of things for linguistics is probably something resembling the Extragalactic DB, with
sub-disciplines within linguistics having some organizational structures such as the Sloan project.
While each approach has its merits, our field could greatly benefit from a project such as the
Google Sky. In the end, tailoring data for a certain project is not the best model for collaborative
data. Instead, we suggest keeping the overall idea of data sharing in mind when projects are
planned. For the developer on the ground, this means using the latest technology and practices in
the most forward looking way possibly, even at the expense of time and other resources.

7.6. Internet Technology and Tools
This section covers specific Internet technology and software that we think will serve to enhance
how a cyberinfrastructure for linguistics is created and maintained. First we cover trends and
practices, ideas or technologies that are not embodied in any one specific tool. Second we turn to
specific pieces of software that are useful for either adoption or comparison.

7.6.1. Internet trends and practices
By Internet “trends and practices” we refer not to specific tools (cf. the following section) but to
general architectures and methods currently in use or becoming more popular. We consider these
as best practices for handling data on the Web, and they include: Linked Data, cloud computing,
and Web services.

We begin with Linked Data (http://linkeddata.org). Consider that the purpose of encoding data is
to make every assumption explicit especially regarding the meaning of annotations. Two
important criteria are required for an implementation: (1) uniqueness both of annotation elements
and data content and (2) the ability to place such information on the Web and link among
individual data. Uniqueness refers to the identifiability an individual datum at any level of
granularity. That is, a single phonetic feature instance should be as uniquely identifiable as an
entire text. We want the Web to be the medium of data storage and the ability to link among that
data. Uniqueness is achieved by using Uniform Resource Identifiers (URIs) for all data (Berners-
Lee et al., 1998). A URI is a reference scheme that can be used to refer to anything whatsoever,
from documents on the Web to actual physical entities (e.g., my horse) or even abstractions (e.g.,
world peace). As a naming scheme, URIs use the typical http protocol of the Web. For instance,
the following URIs might refer to linguistic constructs:

http://linguistics.org/PluralNumber

---

\(^5\) http://www.google.com/sky/
URIs are only the first requirement for the implementation. The second is a way to link the data. For this we use the Resource Description Framework (RDF) (Lassila and Swick, 1999) which is basically a graph model with a serialization (physical representation) amenable to the structure of the Web. Within an RDF graph, we require that graph nodes and arcs all have URIs. Thus, every node and arc is identifiable and definable. The ontological framework is built atop the RDF/URI system and statements are made using URIs as predicate and argument. The basic element of an RDF graph is the triple which is of the form subject-predicate-object. Each concept, role and individual has a URI and occupies a node (or arc) in the RDF graph. The entire enterprise of placing data on the Web using URIs and RDF is known as the Linked Data approach (Berners-Lee, 2006).

We now turn to another Web technology that has recently grown in significance, that of cloud computing. Cloud computing utilizes the resources provided by Internet servers in order to process, maintain, and share data in a non-localized manner. As the landscape of computing becomes more diverse, with each platform having its own idiosyncrasies of advantages and limitations, many application developers are turning to the Internet as a unified platform accessible from any web browser. The benefits of this transition from local to cloud computing are numerous; many exhibit key features of what a linguistics cyberinfrastructure requires in order to be successful.

- **Stable repository:** Data stored on a local machine require its owner to be diligent in their maintenance of a data back-up plan. Though in reality, many data owners do not have a data back-up plan to begin with. Storing data in the cloud frees data owners from this responsibility as they are able to rely on technicians skilled in preserving data.

- **Computational resources:** The problems involved with the diversity of hardware platforms can be alleviated by relying on the cloud or a server farm for its computational ability. This effectively frees the user’s platform, be it a personal computer or mobile device, from the burden of the required resources in order to utilize an application for storing, retrieving and analyzing data.

- **Shared Data:** Cloud computing, by its nature, is imbued with the ability to provide access to data that may have been produced from a variety of sources. In comparison to localized computing, cloud computing provides a significantly lower barrier for collaboration as analysis of data is not limited to one’s own local repository.

In order to develop software that would fully utilize the paradigm of cloud computing and a collaborative linguistics cyberinfrastructure, principles derived from the W3C’s definition of web services need to be accounted for in the design of our systems. What this means is that projects shouldn’t perceive themselves as segregated resources, but rather as modular
components which are able to communicate with other web services for editing, storing, analyzing and displaying data. A web service uses web related standards such as SOAP (as messaging protocol) and REST (an architectural style) to support interoperable machine-to-machine communication over a network. This machine-to-machine communication allows data from an originating repository to have a life of its own beyond the scope of an individual project’s charge. Though a project’s responsibilities may require the development of many kinds of services for editing, storing, analyzing and displaying data as a proof-of-concept blueprint, the communication interfaces between these components would be accessible so that the linguistic community may develop other web services that may evolve the utilization of its data. Each service may be exchanged for another in order to provide functionality not conceived in the original project’s design. Thus, we service translates to something value-added.

7.6.2. Sources of inspiration (pieces of software)

Here we summarize specific software that we found to be sources of inspiration for linguistics. The first is the Pangaea search tool for geoscience.\(^6\) It is both a repository and an exemplar of how datasets could be cited. Pangaea was already mentioned with respect to DOIs allowing for unambiguous citation. We could compare this project to the efforts of the Open Language Archives Community (OLAC) (Simons and Bird 2003).

Next is the FreeBase project,\(^7\) an open, free community model for literally any kind of data. The database resides “in the cloud” and currently contains more than 6 million topics with accompanying facts. Any individuals who have data can create additions to the DB using their free and easy-to-use API. Data providers are encouraged to link their own data to existing resources when possible. We could compare this project to the aims of e-Linguistics\(^8\) that of providing ways to gather and merge disparate descriptive data in linguistics.

Turning to a resource that excels at providing educational materials and as a portal for its field is Nanohub.\(^9\) Nanohub is a website for nanoscience and technology. The site is the state-of-the-art exemplar for how an individual field can muster its resources to present an inviting and educationally friendly portal to its science. The closest resource we have in linguistics is perhaps the Linguist List\(^10\) where researchers can share and discuss ideas (and data).

Next is OpenWetWare, an information and practices resource (in fact a wiki) for the dissemination and promotion of know-how and wisdom among researchers and groups who are working in biology and biological engineering. The wiki is impressive because of its experiment protocols and courses all found at a single location. In linguistics, we could compare the Glottopedia project\(^11\) to this effort. Finally for a tool that we found particularly compelling,

\(^6\) http://pangaea.de/
\(^7\) http://www.freebase.com/
\(^8\) http://purl.org/linguistics/e-linguistics
\(^9\) http://nanohub.org/
\(^10\) http://linguistlist.org/
\(^11\) http://www.glottopedia.de/
ManyEyes\textsuperscript{12} is used for the visualization of all kinds of data. First of all the site provides many open datasets. The main function of this service is that contributors can design and post novel visualization tools to see the same dataset, all according to an API.

7.7. Summary and Open Questions

We have presented the results of our Cyberling 2009 working group’s discussion on how a cyberinfrastructure can be achieved by looking at exemplars from other fields. We chose to explore the best facets of cyberinfrastructure from various fields and began with a few illustrations of what they have achieved. By “related fields” we simply mean those with data, and we found the hard sciences (biology, astronomy, etc.) and various Web movements (Linked Data, cloud computing and web services) to possess the most compelling facets. We organized these facets in terms of following categories:

- Data handling practices
- Collaborative/organizational structure
- Internet technology and tools

In terms of data handling we found that successful fields did not rely on a central data store, but instead adhered to a distributed model. Success was also seen when adequate conceptualizations (terminologies and ontologies) were employed. Much more than just “nice” for presentations, we found flexible visualization of data to be key. Data should be as open as possible, but service providers should be sensitive to access restrictions where required. Finally, all data should be easy to cite, using DOI or URI systems.

For collaboration/organizational structure, we expounded our main philosophy of collaborative modularity. This approach concerns how the project ought to be developed, namely, with an eye towards future collaboration. We contrasted a few past and current projects (those developed for “in-house” use) with more advanced projects that seek interoperation from the very beginning.

These common threads were found in a number of Internet technologies and tools from which we gained much inspiration. First, we gave an overview of the Linked Data movement, cloud computing and Web services. Finally, we presented key tools that we found to be sources of inspiration for linguistics: Pangaea, Freebase, ManyEyes, just to name a few.

7.8. References Cited


\textsuperscript{12} http://manyeyes.alphaworks.ibm.com/manyeyes/

Mark Liberman (co-chair), David Lightfoot (co-chair), Anthony Aristar, Collin Baker, Helen Dry, Laura Welcher

8.1. Introduction

We’ve set up the problem in terms of a four-dimensional matrix. The first two dimensions are the most important ones: first, the projects and activities that need support; and second, the potential sources for support. The third dimension covers organizational structures through which support can flow; while the fourth describes general concepts and principles that may apply throughout.

Why such a complicated framework? The problem is diverse, diffuse, and dynamic. Different sorts of support are suited or available to different sorts of projects and activities; a single project is likely to have several different sources of support at once; and different kinds of support may be needed at different stages in a project or activity’s life cycle.

We begin by sketching each of our four dimensions; we then give illustrative examples for each category, and note particularly acute needs and particularly ripe opportunities. We close with some general recommendations for action.

Note that we’ve sometimes used the broader term “support” in place of the narrower “funding”, because support of other kinds (such as hosting and bandwidth, space and staff support, academic release time, etc.) can often substitute for cash.

8.1.1. What needs support?

Cyberlinguistic activities that need support include the creation, distribution and maintenance of linguistic data, tools, instructional materials, standards, APIs, and interfaces.

Examples to be considered include Audacity,1 AGTK,2 Champollion,3 CLAN,4 Emu,5 ESPS,6 Festival,7 GATE,8 HTK,9 Linguistica,10 LIWC,11 NLTK,12 OpenFST,13 Praat,14

---

1 http://audacity.sourceforge.net/
2 http://agtk.sourceforge.net/
3 http://sourceforge.net/projects/champollion/
4 http://childes.psy.cmu.edu/clan/
5 http://emu.sourceforge.net/
6 http://ldc.upenn.edu/software/esps60.6.linmac.src.tgz
7 http://www.cstr.ed.ac.uk/projects/festival/
8 http://gate.ac.uk/
9 http://htk.eng.cam.ac.uk/
10 http://linguistica.uchicago.edu/linguistica.html
11 http://www.liwc.net/
12 http://www.nltk.org/
SpeechStation, SenSyn, Transcriber, Trees, Wavesurfer, Xaira. Most of these were originally developed for the use of their authors, without specific funding, and then opened up to others by various means; but a few resulted from targeted efforts with funding allocated in advance.

We also observe that web-based tools (like the BYU Corpus Search system or various collaborative annotation systems), with or without open-source access to the server-side code, are likely to become more prominent.

The situation for datasets used in linguistic research is somewhat different. The first and most important distinction is between data that is published, in the sense that any researcher can get access by means of a convenient, open and durable process, and data that is privately held, and is either not available to outsiders at all, or is available only by some non-automatic special arrangement.

Some published data is in the public domain, but in most cases, its distribution and use is controlled by some sort of license. Most experts believe that licensing is a better option even when the goal is completely open access, in order (for example) to prevent plagiarism or commercial misuse. And sometimes dataset licenses must restrict distribution in various ways, even though the goals of scientific communication dictate that these restrictions be as limited as possible. Without such restrictions, many datasets could not be published, due to authors’ publishers’ and broadcasters’ copyright concerns, as well as issues of privacy and confidentiality. However, licenses should impose as few restrictions as possible on use in research and education; and costs should be kept as low as possible, consistent with maintaining the infrastructure needed for dataset creation, maintenance and distribution.

There are many models for creation, distribution and maintenance of datasets used in speech and language research. The entities involved include individual researchers or research groups, ad hoc partnerships, archives, consortiums, and governmental organizations.

Illustrative examples of linguistic data and/or associated organizations include ANAE, ANC, BAS, BNC, Brown Corpus, Childes, CLDC, Corpus.byu.edu, ELDA, Ethnologue,

13 http://www.openfst.org/
14 http://www.fon.hum.uva.nl/praat/
15 http://www.sens.com/speechstation/index.htm
16 http://www.sens.com/sensyn/
18 http://babel.ling.upenn.edu/~kroch/Trees.html
19 http://www.speech.kth.se/wavesurfer/
20 http://www.oucs.ox.ac.uk/rts/xaira/
21 http://corpus.byu.edu/
22 http://www.mouton-online.com/anae.php
23 http://www.americannationalcorpus.org/
24 http://www.phonetik.uni-muenchen.de/forschung/Bas/BasHomeeng.html
25 http://www.natcorp.ox.ac.uk/
Examples to be considered in the area of standards include GOLD\(^{(47)}\) (grant-based), OLAC\(^{(48)}\) (open community), ISO 639-3\(^{(49)}\) (SIL funding), Unicode\(^{(50)}\) (consortium membership model).

### 8.1.2. What are the sources of support?

Past and potential sources of support for Cyberlinguistic activities are government funding agencies (in the U.S. these include NSF, DARPA, DOD, NIST, NIH, DOE, NEH, NSAS, IMLS, and NLM); private foundations (e.g. VWS, Hewlett, Christensen, Packard, MacArthur, Moore, Sloan, Ford); corporations (including scientific project grants, in-kind donations, donations of linguistic resources and tools, and contracts for linguistic resources and tools to be published more widely); individual philanthropists; small-donor fund drives; direct commercial activities (selling data, tools, services, access, subscriptions, etc.); support from universities and colleges

---

27 http://childes.psy.cmu.edu/
28 http://www.chineseldc.org/EN/index.htm
29 http://view.byu.edu/
30 http://www.eldat.org/
31 http://www.ethnologue.com/
32 http://framenet.icsi.berkeley.edu/
33 http://icame.uib.no/
34 http://www.langrid.org/
35 http://ldc.upenn.edu/
36 http://www.ldcil.org/
37 http://librivox.org/
38 http://linguistlist.org/multitree/
39 http://ota.ahds.ac.uk/
40 http://www.oyez.org/
41 http://www.cis.upenn.edu/~treebank/
42 http://www.gutenberg.org/wiki/Main_Page
43 http://www.rosettaproject.org/
44 http://talkbank.org/
45 http://wordnet.princeton.edu/
46 http://wals.info/
47 http://dev.linguistics-ontology.org/
48 http://www.language-archives.org/
49 http://www.sil.org/iso639-3/
50 http://unicode.org/
We observe that most worthwhile activities in this area have started as bottom-up volunteer efforts, where someone scratches an itch and a larger community joins in. This pattern is likely to continue in the future, though finding ways to sustain and enlarge successful instances is a crucial issue.

8.1.3. What are the organizational structures?

Through which sorts of organizational structures can Cyberlinguistic activities be initiated and maintained? There are academic entities such as departments and centers; government entities such as the Smithsonian, the National Archives, NLM; free-standing or affiliated labs, institutes, and other organizations (e.g. ICSI, Haskins, MPI, SIL, Long Now); corporate entities; consortiums; and various confederations and communities such as OLAC or the groups nucleated around Praat or NLTK.

We should keep this diversity of organizational types in mind, since the portfolio of available funding and other forms of support is different in each case.

8.1.4. What are the principles?

Which principles cut across the enormous diversity of problems and solutions discussed here?

First, interdisciplinary cooperation is good, from the perspective of funding as well as for intellectual reasons. In some cases, as at NSF now, funds are specifically targeted at interdisciplinary work; but in general, attacking shared problems with colleagues in related fields can broaden the base of potential support as well as increasing the audience for results.

For similar reasons, international cooperation is good. Nearly all of the successful Cyberlinguistics efforts are now international. In addition to broadening the base of involved researchers, this creates additional diversity of funding sources.

Whatever the scope and scale, active community participation is key. An active, successful community of researchers, doing interesting things, can usually find funding. The Open Source / Open Access movement provides useful models for community building, some of them in the area of Cyberlinguistics; the DARPA “common task” method demonstrates in a very different way the potential value of shared data and shared problems in forming an active research community and in demonstrating progress. Examples to learn from include Open Translation Tools 2009 and the Open Translation Tools manual created in the subsequent book sprint (funded by Open Society Institute and the Ford Foundation); the many courses created around the world in association with the Natural Language Tool Kit (NLTK) project; the OLAC efforts; the large community of Praat users.

---

51 http://www.aspirationtech.org/events/opentranslation/2009
52 http://en.flossmanuals.net/opentranslationtools
53 http://www.nltk.org/courses
54 http://www.language-archives.org/
55 http://www.fon.hum.uva.nl/praat/
The open-source model is the best solution for most software tools aimed specifically at linguistic research, both because of this model’s intrinsic benefits, and also because the market for such tools is generally not large enough to support a significant and ongoing commercial development effort. We note, however, that there are a few examples of successful tools that have been well maintained over a period of years by developers who make binary versions available at a low cost, and use the proceeds to pay for maintenance and even further development (e.g. LIWC, \textsuperscript{56} Trees\textsuperscript{57}).

Wherever possible, tools, data, and other Cyberlinguistic contributions should be citable and cited. They should have authors, dates, perhaps ISBN numbers or other formal indicia. We should create citation formats and insist that authors use them. This leaves a citation-index trail that can play a very important role in generating and maintaining funding, as well as in promoting and rewarding participation by researchers.

Creators of Cyberlinguistic infrastructure should think about life-cycle problems. Support for long-term maintenance is as important as support for creation, and is often harder to find.

Finally, the public interest (and the interest of the public) is very important. Cyberlinguistic infrastructure has potential educational and clinical applications, as well as contributing to the public discussion of linguistic issues. In addition to turning these potentials into reality, we should work to ensure that the public is aware of our accomplishments and opportunities. This can have a large impact on levels of support.

\textbf{8.2. Recommendations for Action}

\textbf{8.2.1. Top-down initiatives:}

We should draft two letters, to be sent by the president of the LSA on behalf of the field.

1. A letter to Joan Maling, head of the Linguistics Program at the National Science Foundation. This letter would suggest additional outreach to linguists about opportunities to access additional NSF funding sources, specifically in the development of cyberinfrastructure, large-scale multi-disciplinary collaborations, and complexity science. In addition, the linguistics program should explore ways for our field to fit into framework of the new MOU between NSF and USAid, which provides additional funds for NSF from the State Dept. for scientific work that connected with foreign-aid programs. The development of resources and materials for indigenous language education, and for linguistic and cultural preservation, is one possible area.

2. An open letter from the president of the LSA, suggesting a major national initiative to rescue the cultural legacy of archival speech and language material, making it accessible today and for future generations. These archives are now in the form of analog tapes, paper notes or transcripts, etc.; and include oral histories, interviews, speeches, debates, courtroom recordings, field notes, and so on.

\textsuperscript{56} http://www.liwc.net/

\textsuperscript{57} http://babel.ling.upenn.edu/~kroch/Trees.html
The problem is a time-critical one. Some of this material will soon be lost forever, due to physical decay processes. In addition, the people who can provide crucial commentary and explanation are gradually dying off.

The opportunity is great: if this material were digitized, indexed and made accessible, its value would be enormous.

This problem is relevant to anthropologists, historians, sociologists, and many others in the humanities and social sciences; and the solution will be of value and interest to broad segments of the public.

Should go to NEH, NEA, Smithsonian, National Archives, AHA, AAA, APS, Oral History Association, etc.

3. Jeff Good and the LSA’s technical advisory committee should be urged to gather and publicize links to materials for graduate student education, especially in areas relating to “cyberlinguistics,” to include relevant mathematical and statistical foundations, as well as computational linguistics, phonetics, etc.

8.2.2. **Bottom-up initiatives:**

All linguists should be urged to build interdisciplinary and international bridges, over which funds can flow.

We observe three areas where the needs and opportunities are great, (lack of) funding is a roadblock, and bottom-up initiatives are needed to find good targets for funding.

1. There is a looming problem: if (when?) large numbers of linguists start taking our advice to publish their data, there will not be enough mechanisms for doing accepting, vetting, archiving, indexing and distributing the submissions. (In contrast, however, we don’t support more black-hole archives, where data checks in but never checks out.)

2. There are 200-300 biggish languages that are seriously under-documented -- in many cases, almost undocumented. If we had an adequate model for what such documentation should look like, and how it could be efficiently and quickly created and published, this would serve many educational and technological as well as linguistic needs. Existing initiatives in the U.S. and Europe are close to creating such a model — but there’s more to do.

3. The field needs better tools for corpus-based, distributed lexicography.

Brian MacWhinney (chair), Emily Bender, Nicoletta Calzolari, Nancy Ide, David Robinson

9.1. Foreword

The collaboration structure group is charged with considering methods for enhancing collaboration and communication on three levels. Level 1 involves forming communication pathways that can link individual researchers to the overall agenda of developing a shared cyberinfrastructure. Level 2 involves the collaborations that are needed between system developers and tool developers to assure maximum interoperability and open access between data formats and programs. Level 3 involves support for collaborations between linguists and researchers in other sciences, grounded on the use of a shared, open access cyberinfrastructure. For each of these levels, we need to design lightweight methods for ensuring ongoing collaboration and coordination.

The specific agenda items for this group include:

- Data level interoperability: roundtrips between formats, transductions of formats, funding for the process of developing compatibility
- Tool level interoperability and methods for collaboration in tool development.
- Issues arising from a commitment to open access.
- How to maximize data-sharing: the role of NSF, NIH, and LSA in terms of promoting greater commitment to data-sharing.
- Characterization of linguistic digital data types and methods for linking to non-digital data.
- An agenda for developing linkages to other sciences - the bigger picture of Extended Linguistics.
- Lightweight administration within a framework of complex organizations: NIH, NSF, LSA, CLARIN, etc.

Following are further analyses of these seven agenda items:

9.1.1. Data level interoperability

- Level 1 involves compatibility in annotation format, such as a formats provided by frameworks such as Annotation Graphs (AG) or the Linguistic Annotation Framework’s Graph Annotation Format (GrAF).
- Level 2 involves compatibility in terms of data categories (content), wherein categories are the same conceptually and can be mapped to one another. This can be facilitated by ontological resources like GOLD as well as the ISOcat
- Level 3 compatibility involves use of a common set of notational conventions to express a fully declared range of content categories.

9.1.2. Tool level interoperability and methods for collaboration in tool development.

- LRT, TalkBank, E-Meld standards are largely similar.
• Media standardization issues for streaming serving and programs. YouTube, Google, Mozilla, and others are developing standards and systems that we could adopt.

• Roundtrips between tool formats: CHAT, Anvil, EAF, AG, EXMaRLDA, Wavesurfer, TEI, SALT. Some of these tools are already interoperable, but these pathways need to be made clearer to users.

• AG Tools approach. This approach allows programmers to develop new tools from the AG Toolkit. However, this is only linked at Level 1. Can there be a similar approach at Levels 2 and 3?

9.1.3. Issues arising from a commitment to open access
What data must be kept away from the public and what data can be made freely available? How can linguists work together to increase access to larger amounts of linguistically important data?

• Sharing and IRB principles at talkbank.org/share.

• Legacy data vs. forward-looking protocols (E-Meld, AphasiaBank, as examples)

• Community, population constraints

9.1.4. Methods for promoting a higher level of data contribution and individual researcher “buy in”

• Inducements: publication, easy tool linkage

• Community: role of LSA

• Obligations and standards: role of NIH, NSF, DARPA, IE

• Leading role of the European Community in setting standards for data-sharing for grant recipients

9.1.5. Characterization of linguistic digital data types and methods for linking to non-digital data
Here, it is important to distinguish the emphasis on corpora and linked media from the many other types of digital data that are of interest to linguists. In the area of Linguistic Exploration, the fundamental objects may be word lists, sentence lists, or dictionaries. In Linguistic Anthropology digitized records of objects are important. This extends eventually over to Archaeology and even information on human genetics etc. In the Learning Sciences, there is an emphasis on linking classroom video to individual student portfolios that may include letters, tests, art work and so on. For digital libraries, it is important to make clear where the hard copies actually reside. For many of these objects, identification can be made through the assignment of digital object identifiers (DOIs). However, this is a largely unexplored territory for most linguists.

9.1.6. An agenda for developing linkages to other sciences
This is the bigger picture of Extended Linguistics. Here, the MacWhinney-Groves NSF report should be particularly helpful.
9.1.7. **Lightweight administration within a framework of complex organizations: NIH, NSF, LSA, CLARIN, etc.**

There is a perception that some work on the development of shared cyberinfrastructure has been top-heavy on committee work and reports without producing a significant amount of shared interoperable resources. Is there a way to build organizational structures that produce open-access products? Who should determine patterns of collaboration or should these patterns “emerge” through specific less-organized exchanges. But then how these interactions be guided toward cooperation and interoperability? Perhaps an emphasis on standards for collaboration might be possible.

9.1.8. **Recommended Readings**

- SIGAnn website¹
- TalkBank website² (see ground rules for sharing, software, browsable database)
- NSF SBE Cyberinfrastructure Report³
- SILT Proposal⁴
- FLaReNet website⁵
- ISO committee for Language Resource Management webpage⁶
- Linguistic Annotation Framework and Graph Annotation Format descriptions (attachments)

9.1.9. **Additional Links**

- ISO TC37 SC4 - Language Resource Management : Homepage⁷
  - ISO Linguistic Annotation Framework (LAF) Overview⁸ and Draft Standard⁹ (under revision); XML Serialization for LAF : ISO Graph Annotation Format (GrAF)¹⁰
- ISO Morphosyntactic Annotation Format (MAF)¹¹
- ISO Syntactic Annotation Format (SynAF)¹²

---

² [http://talkbank.org/](http://talkbank.org/)
³ [http://vis.sdsc.edu/sbe/reports/SBE-CISE-FINAL.pdf](http://vis.sdsc.edu/sbe/reports/SBE-CISE-FINAL.pdf)
⁴ [http://attachments.wetpaintserv.us/PkvenAHTiGiYvp4j%2B55GNg%3D%3D378238](http://attachments.wetpaintserv.us/PkvenAHTiGiYvp4j%2B55GNg%3D%3D378238)
⁵ [http://www.flarenet.eu/](http://www.flarenet.eu/)
⁹ [http://www.tc37sc4.org/new_doc/iso_tc37_sc4_N463_rev00_wg1 wd_LAF.pdf](http://www.tc37sc4.org/new_doc/iso_tc37_sc4_N463_rev00_wg1 wd_LAF.pdf)
9.2. Introduction

In considering collaboration, we distinguish first between joint research efforts and large-scale coordination. The former may cross many boundaries: of disciplines, institutions, and nations. It is also by and large already successful, and not something that we thought this group needed to focus directly on (though of course a cyberinfrastructure for linguistics would, as an important contribution, provide further support for this kind of collaboration). Rather, we defined our task as the fostering of large-scale coordination such as is required to get the field of linguistics (and sister disciplines in the language sciences) to organize around the creation of a cyberinfrastructure and its population with data. We saw three main facets to the large-scale coordination of effort: cooperation, community building, and communication. The following figure summarizes the relationships we see among these concepts.

---

14 http://www.timeml.org/
15 http://www.lexicalmarkupframework.org/
16 http://www.isocat.org/
17 http://linguistics-ontology.org/
18 http://cyberling.elanguage.net/page/Getting+Involved+in+ISO+Standards+Development
Large-scale coordination of effort

Under the heading of large-scale coordination of effort, we include such things as the reuse of data, especially across (otherwise) unrelated research groups; mechanisms for publishing/sharing data; evaluation/quality control of resources (broadly defined to include data sets, tools, standards, etc); establishment of and agreement on standards; and coherence of principles, goals, and architectures across groups involved in the creation of a cyberinfrastructure. In all of these concerns, we place a high priority on avoiding duplication of effort. In order to foster large-scale coordination of effort, we see three main classes of tools: formal modes of cooperation, explicit work towards community building, and vehicles of communication.

Coordination Models

We identified a few examples of models of coordination:

- ISO: exploit existing infrastructure for coordination
- TalkBank: coordinate among researcher groups such as AphasiaBank, CHILDES, PhonBank, ClassBank, etc.
- TEI: build infrastructure from the ground up
- CLARIN: exploit an EC framework (ESFRI) financing the feasibility part of establishing European infrastructures, for CLARIN in the Humanities, to be then continued with national funding from the various EU countries
- FLaReNet/SILT: parallel international funded efforts to establish international networks, with bottom up coordination by the projects’ coordinators

Clearly, the various coordination models are linked to different funding models, and can (to differing degrees) be deliberately fostered by funding agencies.
9.5. The Role of the LSA
An informal working group has been created within the LSA, consisting of LSA staff, leadership and technology consultants. The working group will make detailed recommendations to the LSA’s Executive Committee concerning specific actions the LSA can undertake, both immediately and in the future, to facilitate the development of a cyberinfrastructure for linguistics, disseminate information related to this endeavor, and promulgate a “culture change” with regard to sharing of data, tools, and results. These actions would build on the initial steps already taken by the LSA in this regard, such as its digital publishing platform, eLanguage.

9.6. Community Building
We began with the observation that successful technology is always supported by (while also supporting) a community of users. We define a community for these purposes as a group of people working on similar issues, using the same tools/platforms/resources, who talk to each other and who share principles and practices in their research efforts. Different communities may be display these various properties to differing extents. To give some examples, SIGs (special interest groups within larger scholarly bodies) illustrate communities primarily defined by working on similar issues. Computational linguistic communities that have grown up around the creation of use of tools and resources include the developer and user groups of WordNet, FrameNet, NLTK, GATE and UIMA. There are also communities of linguists who are producing shared databases, but not new computational tools or formats. These include groups such as CHILDES, PhonBank, AphasiaBank, LIDES, etc. Vehicles for communication (discussed further below) can also create communities of their own. Examples here include the readership of LINGUIST List and the readership of the Corpora list. Note also that communities can vary greatly in size (readership of LINGUIST List being at one extreme) and of course overlap with one another, as individuals belong to multiple different communities.

Since communities are important for both communication (see below) and the success of software projects, we considered ways to foster the development of communities, while recognizing that such things cannot be precisely engineered (nor their success precisely predicted). Means of community building include:

- Funding programs sponsoring multiple groups working on the same/similar problems (e.g., language documentation sponsored by ELDP, DARPA programs)
- Evaluation campaigns (lots of examples from complying here: Semeval, MUC, TREC, CLEF, CoNLL shared tasks)
- On-line fora (from big like LINGUIST List to small like user groups for particular tools)
- SIGs (of particular relevance here is ACL’s SIGANN)
- Workshops (e.g., LAW, E-MELD workshops, Cyberling 2009) and conferences (LREC is of particular note here in having created a community)
- Tutorials
  - At summer schools (LSA institute, EuroLAN, Johns Hopkins, ESSLLI)
  - On the web
  - As part of funded projects (e.g., SILT)
- Journals (LRE, Language, eLanguage, Computational Linguistics...
9.7. Communication

We identified communication as a problem that cross-cuts many aspects of large-scale coordination of effort. In particular, we need to communicate about standards (availability and development), tool and resource availability, needs assessment, and principles & practices. People working on tools and standards across linguistics and the language sciences more broadly need to be aware of each other and each other’s efforts and need to be able to communicate with potential users for needs assessment. (Mark Liberman commented that every successful piece of software starts with someone scratching an itch: they have a problem, build a solution, and share that solution. But not everyone who has a problem that can be solved with software has the means/skills to build that software themselves.) People potentially using tools and standards need to be able to find them. People who should be using tools and standards but don’t yet know about them, need to be reached.

With these kinds of communication in mind, we developed a list of potential communication vehicles:

- Existing communities’ infrastructure (newsletters, meetings, websites), including both informal communities and scholarly organizations
- Teaching materials, especially made available over the web (syllabi, problem sets, web-based tutorials)
- Wikis/blogs (“bottom-up web-based communication”)
  - On-going maintenance of information collections
  - Reasons for people to come back to the on-line communication site
  - ad words: Set up context sensitive “ads” on the model of Google AdWords which could run on LINGUIST, lsadc.org, etc, where the things being advertised are relevant projects and standards (and no money is exchanged)
- Funded collaborations (e.g., SILT/FLaReNet)
- Workshops/tutorials
- Reviewing guidelines/review feedback
  - Pushing funding agencies to require plans (and follow through) for using standards and publishing data for proposals that use tools/create data
  - Pushing funding agencies to require proposals for new tools/standards to appropriately cite and situate themselves within the existing tools/standards ecology
  - Conference/journal reviewing check for appropriate citations of data, tools, resources
- Resource maps/eliciting metadata (cf. LREC 2010)
- Journals like Journal of Experimental Linguistics which publish code along with the resulting research.
- Idea from Steve Moran: A new journal (perhaps in the eLanguage set) on the model of Journal of Experimental Linguistics, which publishes data sets collected in the field. Maybe called “Journal of Linguistic Description”? 
9.8. Content
This section briefly outlines the content that we need to be communicating about within the large-scale coordination of effort required to bring about a cyberinfrastructure for linguistics (and the language sciences).

- Data level interoperability, on two levels: Interoperability of data and annotation format, including means of mapping between existing formats (cf. GrAF) and interoperability of annotation content, via ontologies (cf. GOLD) or other inventories of linguistic categories (cf. ISOcat).
- Tool level interoperability and methods for collaboration in tool development.
- Managing issues arising from commitment to Open Access, establishing and publicizing shared principles.
- Promoting individual researcher “buy-in”. A cyberinfrastructure is not useful until it is populated with data, but our field is in need of culture change in this respect. We believe this culture change can be achieved through a combination of making it easier for individual researchers to contribute data (through useful tools), educational campaigns on the part of the LSA and similar groups, funding agencies establishing policies requiring data sharing, and publication venues requiring testing against and citing existing available data.
- Connection to other fields: Linguistics is only one of the language sciences, and digitized data from many fields (education, political science, law, ...) can be valuable for linguists. As we develop our infrastructure, we need to be mindful of how it fits into this larger ecology, and where teaming up with other language sciences can bring economies of scale. Closer to home, the field of computational linguistics has a good deal of cyberinfrastructure (and communication around cyberinfrastructure) established. We envision creating a portal into cyberinfrastructure concerns for people who identify as linguists, which rather than attempting to encompass all language-related cyberinfrastructure itself links to existing efforts in allied disciplines.

9.9. Summary: 5 Cs of cyberinfrastructure
- Collaboration: Joint research alone isn’t enough to bring about a cyberinfrastructure, though it will play a key role. We need large-scale coordination of effort.
- Cooperation models: There are many ways to coordinate effort, and we will probably use all of them.
- Coordination: On the technical side, we need interoperability, which entails coordination on standards
- Communication: On the people side, we can’t achieve coordination without communication, bringing people in, making them aware of each other, and keeping them in touch.
- Community Building: Key to both successful communication and successful software.

9.10. Action items - Short term
- Draft recommendations to funding agencies regarding standards, data publication, etc.
• Draft recommendations to journal editors and conference organizers regarding citing tools/resources and publishing data
• Create teaching resources (through LSA?)
• Continue this conversation (all WGs, on the wiki for now)

9.11. Action items - Long term
• Ensure communication among projects/groups
• Push those developing standards to for specific areas (e.g., PHON group connected to TalkBank) to contribute to ISO TC37 SC4
• Work towards data/annotation harmonization
10. Appendix: List of Participants, Affiliations, and Relevant Projects


Aristar, Anthony. Department of Linguistics, Eastern Michigan University; LINGUIST List; LEGO; MultiTree.

Austin, Peter. Marit Rausing Chair in Field Linguistics at the School of Oriental and African Studies (SOAS) in London; Director of the Endangered Languages Academic Programme.

Baker, Collin. UC Berkeley; International Computer Science Institute (ICSI); FrameNET; American National Corpus (ANC).

Beckman, Mary. Department of Linguistics, The Ohio State University.

Beermann, Dorothee. Norwegian University of Science and Technology (NTNU); Typecraft.

Bender, Emily M. Department of Linguistics, University of Washington; The Grammar Matrix.

Bickel, Balthasar. Department of Linguistics, University of Leipzig; Autotyp.

Bollacker, Kurt. The Long Now Foundation; CiteSeer; The Internet Archive.

Byrne, Bill. Google.

Calzolari, Nicoletta. Director, Institute of Computational Linguistics, National Research Council, Pisa, Italy.

Churng, Sarah. Department of Linguistics, University of Washington.

Corbett, Greville. Department of Linguistic, Cultural, and International Studies, University of Surrey; Surrey Morphology Group.

Dry, Helen. Department of English Language & Literature, Eastern Michigan University; LINGUIST List.

Dwyer, Arienne. Department of Anthropology, University of Kansas; the Uyghur Dialect Archive.

Farrar, Scott. Department of Linguistics, University of Washington; e-Linguistics Toolkit; GOLD.

Fillmore, Chuck. Department of Linguistics, UC Berkeley (Emeritus); International Computer Science Institute (ICSI); FrameNET.

Forkel, Robert. Max Planck Digital Library; WALS Online.

Good, Jeff. University at Buffalo. LSA-TAC, Rosetta, Pangloss.

Haspelmath, Martin. Department of Linguistics, Max Planck Institute for Evolutionary Anthropology; WALS.

Ide, Nancy. Department of Computer Science, Vassar College; American National Corpus (ANC).
Jaeger, Florian. Departments of Computer Science and Brain and Cognitive Sciences, University of Rochester.

Kahn, Jeremy. Department of Linguistics, University of Washington; UW Signal, Speech and Language Interpretation Laboratory; SRI Speech Technology and Research Laboratory.

Kansa, Eric. School of Information, UC Berkeley; Information and Service Design Program; OpenContext.

King, Tracy. Powerset; LSA-TAC, Grammar Engineering Across Frameworks (GEAF).

Langendoen, Terry. Department of Linguistics, University of Arizona (Emeritus); E-MELD; GOLD; NSF.

Liberman, Mark. Departments of Linguistics and Computer and Information Sciences, University of Pennsylvania; Director, Linguistic Data Consortium.

Lightfoot, David. Georgetown University; NSF.

MacWhinney, Brian. Departments of Psychology, Computational Linguistics, and Modern Languages, Carnegie Mellon University; CHILDES Project; TalkBank.

McCloy, Daniel. Department of Linguistics, University of Washington.

Mihaylov, Pavel. Ontotext; Typecraft.

Moran, Steven. Department of Linguistics, University of Washington; Dogon Languages Project; PHOIBLE; E-MELD.

Nichols, Johanna. Department of Linguistics, UC Berkeley (Emeritus); Autotyp.

Palmer, Alexis. Saarland University (CoLi) and the University of Texas at Austin; M2CI Cluster of Excellence; EARL Project.

Puschmann, Cornelius. Heinrich-Heine University, Düsseldorf; eLanguage.

Robinson, David. Director of Membership and Meetings, Linguistic Society of America.

Robinson, Stuart. Powerset.

de Smedt, Koenraad. Department of Linguistic, Literary and Esthetic Studies, University of Bergen; Research Group on Language Models and Resources (LaMoRe); CLARIN; CLARA.

Sornlertlamvanich, Virach. National Electronics and Computer Technology Center (Thailand); Thai Computational Linguistics Laboratory (TCL); Asian Language Resource Group of the Asian Federation of NLP; Asia-Pacific Association for Machine Translation.

Trilsbeek, Paul. Language Archive at the Max Planck Institute for Psycholinguistics; DOBES; CLARIN.

van Tuyl, Dwight. Eastern Michigan University; LinguistList; GOLD; LEGO.

Wassink, Alicia. Department of Linguistics, University of Washington; Pacific Northwest Vowels Project.

Welcher, Laura. Director of the Rosetta Project, The Long Now Foundation.

Wright, Richard. Department of Linguistics, University of Washington; Vocal Joystick; PHOIBLE.