The Human Language Project

Steven Abney

December 16, 2007

Contents

1 Overview

1.1 Motivation

By far the most important and urgent task facing linguistics is language documentation, that is, the collection and archiving of primary linguistic data. Rarely, perhaps never, has any science faced such a rapid and widespread loss of its primary data as linguistics now faces. Languages are being lost at an alarming rate, and the loss is only expected to accelerate. The next generation of linguists will forgive us for the most egregious shortcomings in theory construction, but they will not forgive us if we fail to record vanishing primary data in such a way as to serve the needs of linguistic research in perpetuity.

The scope of the task is enormous. The amount of funding available for language documentation has certainly increased in recent years, but even so, it is inadequate for the task. The only hope of success is through a community process to establish and maintain a repository containing, ideally, a complete digitization of every human language. Such a repository represents, not a Universal Grammar, but a Universal Corpus.

Linguistics as a field is awake to the problem of language loss, but unfortunately the currently dominant school of linguistic theory undervalues language documentation, and current practices in language documentation undervalue primary data.

The currently dominant research methodology is radically "top-down": theory construction is largely driven by subjective notions like "elegance" or "minimality," and language data is consulted only where it directly bears on questions arising in the course of theory construction. Collecting large quantities of data that do not directly address the handful of currently interesting theoretical questions is considered a waste of time and resources. Language documentation is viewed at best as a harmless diversion for people who are not interested in theory.

Enhancing the prestige of language documentation would not be sufficient to solve the problem, however. Current practice in language documentation focuses almost exclusively on the production of print grammars and lexica, with perhaps a small sample of illustrative running text as an appendix. Primary data is collected mostly by elicitation in the process of constructing a lexicon or grammar. After the lexicon or grammar has been constructed, some of the primary data can be found as examples embedded in lexical entries or grammatical descriptions, but the majority languishes on paper slips in file cabinets for some number of years, and is eventually destroyed.

For future generations of linguists, secondary data in the form of lexica,

morphological paradigms, and grammatical observations are all very valuable, but arguably the most important resource is primary data: running text, in sizeable quantities. (We interpret *text* broadly, as subsuming not only printed texts but also recorded speech.)

It is difficult to say very precisely what an adequate digitization of a language would look like, but a natural if rough criterion is the ability to "reconstitute" the language in something very close to its original form, using only the digitization. There is at least one example of an extinct language that was resurrected, namely, Hebrew; and there are several examples of extinct languages where enough material remains that a few scholars are able to obtain near-native competence in the language. In all cases, the critical resource is primary data: running text.

The field of natural language processing (NLP) has long recognized the importance of large collections of primary data, a.k.a. **corpora**, and has acquired a great deal of experience in the construction and exploitation of corpora and other electronic resources, including both primary data (text and speech) and secondary data (electronic lexica, treebanks). Unfortunately, the focus in NLP is on resources for which one can mostly easily obtain funding, which means resources that enable technology of commercial or national-security interest. These priorities are not entirely aligned with long-term scientific interests. In particular, endangered languages are almost by definition uninteresting – a language with few speakers has no commercial value and little or no intelligence value.

Our proposal can be viewed as combining the questions posed by linguistics with the methods of NLP. We take from linguistics the high value placed on long-term scientific goals, and the sense of urgency regarding language loss; and from NLP, we take the high value placed on primary data and the know-how for creating and using it.

We believe that introducing computational methods to linguistics is not only a practical response to the loss of data, but also a way of improving the scope and effectiveness of linguistic inquiry. Linguistics has not really understood how to apply the scientific method to language. The scientific method does *not* consist in searching for isolated examples that can be used in an argument to distinguish between two equally *a priori* theories. Success in science is measured by a theory's ability to make accurate predictions about an entire class of phenomena, as estimated by the accuracy of predictions on a random sample of new observations from the class.

"Treebank parsing" is an excellent example. Based on a collection of previous observations, one constructs an hypothesis (a parser) that assigns syntax trees to sentences. The parser makes nontrivial predictions – for any given sentence, it predicts which syntax tree a human will assign to that

sentence. Its accuracy is gauged by drawing a test sample of sentences, and comparing its predictions to actual human syntax-tree assignments. Importantly, the test sentences are not selected by the designer of the parser. They constitute a representative sample of the population of naturally-occurring sentences of the language of interest, and provide an unbiased estimate of the accuracy of the parser's predictions over the entire population.

Parsing is traditionally seen as a question for NLP, not for linguistics, but that rests largely on misapprehensions about what questions a syntactic model should answer, and how parsers work.

Linguistics recognizes human judgments about grammaticality and ungrammaticality of sentences as a central topic for syntax. But even more fundamental are judgments concerning the correct syntactic structure for a given sentence. Not only do grammaticality judgments presume an assignment of structure, but the structures themselves are of much more central interest to syntax than the mere classification of sentences as grammatical or not. Remarkably, though, most current syntactic accounts make no predictions about which syntax tree is the correct tree for a given sentence. Syntactic accounts focus on admitting some trees and rejecting others. But any syntactic account that is general enough to admit the correct tree for a sufficiently large range of sentences, will necessarily admit *many* trees for every sentence – even though most of those sentences are judged unambiguous by humans. This fact is not appreciated by most syntacticians. It forces one to adopt a syntactic model (a grammar) that does not simply distinguish grammatical from ungrammatical, but assigns degrees of goodness to trees. There may be many trees that the grammar admits for a given sentence, but, for most sentences, there is a unique best tree according to the grammar; the grammar predicts that its best tree is the one that a human will judge to be correct.

Parsers are usefully factored into two parts: a grammar of the kind just described, and a search method. The job of the search method is to compute which tree is best according to the grammar. Judging the accuracy of predictions is done by comparing the best tree found to the correct tree according to human judgment. Two kinds of errors can arise: the search may fail, so that the best tree found is not in fact the best tree according to the grammar, or the grammar may make the wrong prediction, meaning that the best tree according to the grammar is not the correct tree. There are no published results concerning the relative proportion of search errors versus grammar errors, but exhaustive search is tractable, so the fundamental bottleneck is the grammar. Only the search question is a computational issue; the design of the grammar is a purely linguistic question. If search errors are comparatively rare in practice, then "treebank parsing" is really a test

of the grammars used. Much remains to be done: the current best parsers only get about 90% of the syntax-tree nodes correct. Since typical sentence lengths are 25–40 words, nearly every sentence has at least one error in it.

A grammar of the sort that we have just been discussing is a model of a single language. To even begin to do *universal* linguistics, we require systematic data collection over a large range of languages. The situation in linguistics is similar to that facing molecular biology ten years ago: a major hindrance to asking questions about the regulation and interaction of genes was the lack of a basic database that included all human genes. This quote from "The science behind the Human Genome Project" [2] is apropos:

> One of the greatest impacts of having the sequence may well be in enabling an entirely new approach to biological research. In the past, researchers studied one or a few genes at a time. With whole-genome sequences and new high-throughput techniques, they can approach questions systematically and on a grand scale. They can study all the genes in a genome, for example, or all the transcripts in a particular tissue or organ or tumor, or how tens of thousands of genes and proteins work together in interconnected networks to orchestrate the chemistry of life.

We require a basic database that includes all human languages, in a form that enables automated processing. The basic questions of linguistics – such as how language is learned, how it is processed, and how it changes through time – can only be answered by making intelligent use of computation and large-scale data resources, and a large multi-lingual corpus is the obvious first step. As Bloomfield wrote, "a truly universal linguistics, when it comes, will be inductive rather than deductive" [1]. Natural language processing has advanced to a point where a universal linguistics based on systematic data processing is technically feasible.

I have been careful to write "natural language processing" rather than "computational linguistics." The two names are usually treated as synonymous, but I would like to make a distinction between them. NLP is a subfield of artificial intelligence, hence a branch of computer science. I am proposing not only the creation of a corpus, but the creation of a new kind of linguistics, a *computational* linguistics, that combines the strengths of NLP and linguistics. Unlike NLP, this computational linguistics is linguistics, not computer science. It has at least three branches: (1) computational methods for linguistics, that is, the use of computational and "data-heavy" methods to pursue the scientific study of language, (2) computational models of human language processing, and (3) the development of human language technologies. The field of NLP has paid significant attention only to

the third of these topics, technology. The first two topics are the proper domain of linguistics and psycholinguistics, respectively, though they have been largely neglected to now. The proposed corpus and the investigations that it enables are examples of the first topic, computational methods for linguistics.

There is a natural migration of knowledge and techniques developed by mathematicians and, more recently, computer scientists, into the subject-matter areas. In physics, mathematics is so taken for granted that it would be redundant to say "mathematical physics." The term could only be used for a subdiscipline involving mathematics that is sufficiently esoteric that mathematicians are still actively involved in its development. In the future, I expect a similar integration of mathematical and computational knowledge into linguistics, so that what I today call "computational linguistics" will simply be "linguistics" in the future. The proposed corpus will both require and enable a central role for mathematics and computation in linguistics.

1.2 Goals

Despite the appeal of the "Universal Corpus" mentioned above, it is naive, and even counterproductive, to think in terms of creating *the* universal repository of linguistic data. We certainly should not duplicate efforts that are already underway. Our focus will be on what is missing from current efforts: (1) creating data in fully machine processable format, (2) with consistency of annotation across languages, (3) for a large set of languages selected with a priority on endangered languages and endangered documentation rather than commercial or intelligence interest. Current efforts in computational linguistics satisfy criterion (1), and can be used as a point of departure for formats and annotation, but they do not make (2) or (3) a priority. Current efforts in the language documentation community focus on creating page images of traditional print documentation, rather than fully machine processable data. For them, criterion (3) is a priority, but not (1) or (2).

Our aim is to create a corpus satisfying all three criteria, building as much as possible on existing resources. In the process, we expect to establish (or refine) data formats and create tools that will be of use to the larger language documentation and language preservation communities, and we hope to persuade linguistics as a field of the value of systematic, computationally-supported methods for addressing core questions of linguistics, in particular, questions of universal grammar.

1.3 Basic principles

The basic principles of the envisioned corpus are these:

- **Universality.** As many languages should be included as possible. An ambitious but not unreasonable first goal is to include a "minimal sample" from 1,000 languages. We should include at least a token amount of data from every recorded language.

- **Machine readability and consistency.** An important aim is to enable new types of linguistic inquiry, involving machine processing across many languages. Materials intended to be read by humans are rarely suitable for machine processing without considerable preprocessing.

- **Community ownership.** We cannot expect a central agency, or any small group of researchers, to assemble a resource of this scale. It will be necessary to get community buy-in if it is to come into existence. In keeping with community ownership, as well as data security through replication, the repository should not be in the sole possession of any one institution but should be mirrored (replicated) at multiple sites.

- **Accessibility.** Being owned by the community, the corpus should be freely available to the entire community. While preserving the intellectual property rights of contributors, and sufficient editorial control for quality assurance, it is important to place as few limits as possible on community members' ability to obtain, enhance, and redistribute the corpus. The corpus should be mirrored at multiple locations, and the entire corpus should be downloadable.

- **Objectivity.** The corpus aims to be objective and descriptive, not an encoding of any theoretical hypotheses. Annotation should be of the style of the better descriptive grammars or instructional materials, using plain, unbiased, and clearly-defined terms. To the extent practicable, the information in the corpus should be cast in a form that is useful to the entire community, and should not require commitments to any particular school of linguistics.

- **Reproducability.** Traditional linguistic analysis, in the form of grammars and lexica (for example), should certainly be included in the repository, but such products represent *secondary information*, not primary data. To the extent possible, all primary data on which grammars, lexica, or other secondary products are based, should be included in the repository. Ideally, one should be able to reproduce every step of any analysis, from primary data.

2 Current efforts

There are several relevant current efforts.

2.1 EMELD

The EMELD project (`emeld.org`) has just ended. Its ultimate aim was similar to ours:

> "without adequate collaboration among archivists, field linguists, and language engineers ... a common standard for the digitization of linguistic data may never be agreed upon; and the resulting variation in archiving practices and language representation would seriously inhibit data access, searching, and cross-linguistic comparison. ... If linguistic archives are to offer the widest possible access to the data and provide it in a maximally useful form, consensus must be reached about certain aspects of archive infrastructure." [`http://emeld.org/`]

It created best practices, several pilot samples of best-practice annotation, and the GOLD ontology that provides a candidate annotation vocabulary.

2.2 The Rosetta project

The Rosetta project (`rosettaproject.org`) aims to collect data from all human languages. They collect grammars, vocabularies, and texts. The focus is on human-readable rather than machine-readable documents: in almost all cases, the materials are scanned images, not machine-processable texts. They currently have materials from over 2,000 languages. They are funded by the National Science Foundation and private donations, and are hosted by the Stanford University libraries. They are a project of the Long Now foundation.

2.3 The Linguistic Data Consortium

The Linguistic Data Consortium (LDC) makes a huge amount of data available, but at a cost. The model is appropriate for large industrial efforts and well-funded researchers in computer science, but is beyond the means of most linguists.

At least one LDC-related project aims to collect data for a moderately large number of languages. Namely, there is a DARPA funded effort to create corpora and related materials for a few tens of languages, in service of quick development of machine translation systems. (DARPA is the Defense

Advanced Research Projects Agency.) The data will be made available through the LDC. The priority in language choice is the likelihood of the speakers of the language representing national security interests; there is no priority placed on endangered languages.

2.4 Traditional print archives

A number of traditional print archives are making efforts to digitize their collections. "Digitization" generally means simply transfer to electronic media, not the creation of data sets that support automated processing. There are often rather heavy restrictions on access to the data.

2.5 Community data collection efforts

There are community resource-creation efforts that can serve as models, or at least inspiration, for the present effort. Project Gutenberg is one example. It relies on volunteers to scan and edit out-of-copyright books. It has been very successful in creating a large repository of texts of generally good quality: it currently contains over 20,000 works in more than 50 languages. One can imagine making the texts of the corpus available through Project Gutenberg, and both the organization of volunteer work and the redistribution license might well be adapted for a linguistic corpus.

There are other electronic library efforts as well, such as the Internet Archive and the Open Content Alliance. Open software efforts are also relevant. Of particular interest are the GNU Free Software Directory and Source Forge.

2.6 Metadata collections

A *metadata repository* is essentially a card catalog that spans data collections. The Open Language Archives Community (OLAC) is a metadata repository for language resources. Although valuable for finding existing resources, they do not imply anything about the accessibility or form of the resource contents.

The library community is very interested in distributed virtual libraries and collaborative cataloging. Sites of note are LibraryThing and the Open Library project. LibraryThing is interesting as a model of the production of a good-quality database by volunteer effort.

2.7 What sets the proposed corpus apart

To the best of our understanding, none of the existing efforts just listed are on a trajectory to produce the corpus that we desire. Of the corpus

principles listed earlier, the ones that receive the least attention in current efforts, and hence define the unique contribution of the proposed corpus, are:

- Community ownership and accessibility. None of the existing efforts espouse community ownership or the idea that the entire corpus should be freely downloadable. The LDC corpora are downloadable, but at significant cost and under various restrictions. Mirroring and redistribution are not permitted. None of the other efforts even permit downloading.

- Machine readability and cross-language consistency. The LDC corpora come closest: they are machine readable, and have at least the consistency of using standard encodings. Most of the other efforts do not even produce machine-readable texts, but stop with page images.

  However, LDC corpora for individual languages are largely created separately, using different conventions for each language. Universality is not a high priority at the LDC; it is strongly driven by commercial and intelligence interests.

  Nonetheless, the distance from the LDC corpora to the cross-language consistency we envision is not great, and adapting the LDC data would be a huge leg up, if the redistribution issue could be addressed.

There is another important way that the envisioned corpus differs from the other efforts listed above. We do not propose to create a library or repository, but rather a single, coherent, albeit very large, data set. We do not propose to include *all* available data from every human language, only *enough* data from each language – a representative sample of each language. We also aim to annotate those samples in a way that is consistent across all languages, for the sake of language-universal linguistic research.

This makes our goals sufficiently different from those of the other efforts to ameliorate any sense of competition. We are not seeking to displace other efforts, but rather to acquire a sample of each language, and to add value through a universally consistent annotation. There is no reason why the envisioned corpus should not be distributed through existing repositories, such as the LDC and Rosetta, in addition to independent mirroring and redistribution. (There are in fact precedents for freely-available corpora being distributed through the LDC.)

3 How to build it

A corpus covering the majority of the world's languages is a major undertaking, and cannot be accomplished by any one person or even any one funded project. It would be quixotic to start from scratch: a general principle is to build on existing data and standards where possible, and to maximize cooperation with other language data-collection efforts. It must be a collaborative effort. Obtaining the cooperation necessary to create the corpus will only be possible if a large enough proportion of the community can be persuaded of its value, and if barriers to its creation can be eliminated. The steps that must be taken are not only technical, but also social.

3.1 Motivation for data contributors

To attract contributors of data, the project must offer clear benefits to them. The primary benefit offered by the envisioned corpus is a permanent scientific record of human languages in a form that supports universal linguistic research.

That benefit may be too abstract for linguists who work on only one or a few languages in depth. Similarly, communities of native speakers are interested in their own languages, and are much less interested in an abstract promise of universal linguistics. But cross-linguistic consistency of formats and annotation is beneficial even for those interested in single languages.

A speaker of a given language, or a linguist working on a particular language, obviously has the direct benefit of materials contributed by others working on their language. But in addition, and perhaps more importantly, the corpus provides economies of scale for the development of software and other resources for language research, instruction, and preservation. The standards required for consistency across the corpus enable the development of truly language-universal software. Even if one is interested in only one language, methods for processing linguistic data are applicable across all languages, and having one's own language in standard format makes it possible to use software developed by anyone in the larger linguistic community.

Such software currently does not exist. To give a single example, despite the existence of standards like Unicode, no current browsers under current operating systems have, in their default configuration, fonts and display methods that support the entirety of Unicode. Such fonts and display methods are required, however, for uniform access to a universal corpus. We return to this issue shortly.

Not only tools for processing existing data, but also tools for getting raw data into appropriate formats, need to be developed, and they are of interest

to everyone in the community. Of particular importance – and particular difficulty – is optical character recognition (OCR) to convert document images into machine-processable text. There are large-scale document scanning efforts currently underway, including at least the Google digitization efforts and the Open Content Alliance. OCR is widely considered a solved problem, but it is far from fully automatic. Considerable manual effort is needed, in post-editing as well as identifying and re-scanning poorly scanned documents. OCR for texts in languages other than English is *not* a solved problem.

To create corpora for many languages quickly, it obviously makes sense to negotiate with existing providers such as the LDC or Rosetta to contribute data. The value to them would be the promise of having the data converted to a universal format and annotated. For data sets that are already in machine-readable format (as the LDC corpora), the value added is consistency of annotation across languages, and, ultimately, universal coverage. For page images (as in the Rosetta collection or traditional print archives), the value added is conversion to machine-readable form.

In the long run, the main source of data for the corpus will be researchers working on particular languages. Individual researchers are likely to be a major source of data in the short run, as well. To attract individual contributors, one needs to create a critical mass of existing data. Computational linguists have seen the benefits of large machine-processable data sets sufficiently often that creating such data sets for community benefit is a matter of course. A similar culture needs to be established in linguistics.

3.2 Pilot data sets

A first order of business is to begin pilot efforts with linguists working on individual languages for which electronic resources do not already exist. Languages that diverge from electronically well-documented languages, and stretch the bounds of standard annotation schemes, are of particular interest.

In parallel, we need to begin pilot experiments in converting existing annotated data sets to a common format. A prime question, for example, is whether one can map the existing treebanks – the Penn treebank, the Susanne treebank, the Czech dependency treebank, the Chinese treebank – to a common format.

3.3 Tools

To create even a pilot for a few languages, some basic tools are needed. Specifically, we need a browser and editor for arbitrary texts in the Unicode character encoding. As mentioned above, such a browser and editor do

not already exist. Many browsers and editors exist that accept Unicode, but none of them work correctly unless one has installed the appropriate fonts and input methods for each language one wishes to deal with. In a language-universal corpus, this is an intolerable requirement.

We propose, and are currently constructing, a browser and editor with a functionality roughly comparable to TextEdit (on the Mac) or WordPad (under Microsoft), that uses redistributable fonts to cover as nearly all of Unicode as possible. It also provides input methods for a number of languages, as well as a simple input-method editor for allowing the end user to create specialized keyboards. It will be able to produce PDF with embedded fonts for consistent printing.

For portability across different hardware platforms and operating systems, the first implementation is in Java. Having additional native implementations for popular operating systems would be an added benefit. For Unix systems, a version of `more/less` that displays Unicode characters correctly without requiring special terminal modifications would also be useful.

A significantly more difficult issue is the development of tools to do OCR on documents in minority languages, in non-English fonts. It is not clear whether this is a viable path, or whether manually typing in texts will actually prove more cost-effective, given the bounded size of text per language, but the large number of languages targeted. An editorial process for manually converting page images, such as the procedures of Project Gutenberg, may prove more viable. OCR output can of course be used as a starting point where it is available.

3.4 Data formats

The tools described in the previous section are *not* exclusionary. Our goal is not to create a closed "linguist's workbench" or the like, using proprietary data formats. We intend to use open formats, and adhere as much as possible to established standards. Our intended contribution is not in the tools, but in the data. The tools are a means to an end, and other tools that handle the same data formats are more than welcome.

Minimally, we propose to use Unicode text encoding and XML markup. We will also support a tabular representation of structure (newline-separated records, and tab-separated fields within a record), which is more convenient for automated processing. XML will be supported as an interchange format, because it is an established standard.

We also prefer stand-off annotation, by which we mean annotation that is contained in a file separate from the one being annotated. This allows more flexibility in annotation, and conforms to the general principle of leaving the original unaltered, an application of the principle of replicability.

Also for the sake of replicability, it is important to include page images of the original documents whenever a machine-readable text is derived from a print document. A frustration with many currently existing corpora is that the texts are given without context, without even citation, and with no way of determining whether an obvious textual error was in the original or introduced in the course of conversion to machine-readable form.

Additional specifications for higher levels of structure are needed, though there is currently a great deal of variety. The Text Encoding Initiative (TEI) defines a variety of specializations of XML for representing particular kinds of documents. As for linguistic annotation, a variety of annotated corpora exist in different languages. A large number of corpora are annotated with parts of speech, though one must distinguish between those that are automatically annotated versus those for which annotations have been manually checked and corrected. There are also syntactically-annotated corpora (**treebanks**) for an increasing number of languages, including at least modern English (the Penn treebank and the Suzanne treebank), Middle English, Old English, Czech, and Chinese.

There is some agreement on conventions for parts of speech and syntactic categories (collectively, "tagsets"), though nothing like standardization. Fortunately, translating between different tagsets and structural conventions is not insuperable; anyone who has worked with the existing corpora has written scripts to do such translation.

We have little choice but to define our own conventions, but we will do so following two principles:

- maximizing consistency with current conventions, to the extent that they exist
- defining the meanings of symbols objectively, and explicitly testing (and improving) the agreement across different annotators in their application

3.5 Editorial process

The corpus will not come into existence without the contributions of many volunteers; hence, the barriers to contributing must be kept low. At the same time, there must be mechanisms in place to assure quality. There are plenty of examples of collaborative efforts on the internet that began well but devolved into chaos because they were *too* open. Not everyone who might like to contribute is competent to contribute, and not everyone is well-intentioned.

It is important to have good quality-control processes from the beginning, because in a collaborative effort of the kind envisioned, there is a

positive feedback loop: a high-quality resource attracts high-quality contributors, further improving quality, but a low-quality resource attracts the wrong sorts, and becomes worse.

Hence, an editorial process is required. It should be designed to keep barriers low for good contributors, but it should be effective at preserving a high level of quality. Exactly what form the editorial process should take is an open question, but existing projects such as Project Gutenberg, the World Wide Web Consortium, the Wikipedia, and online journals such as the *Journal of Machine Learning Research* provide models that can be considered.

3.6 Data as publication

Some barriers to contribution are social. Making contributions to a public data resource is not valued by the community the way a publication is. A field linguist with primary data may be reluctant to make the data publicly available because, first, the considerable effort involved in preparing the data is not recognized as a publication and hence does not contribute to tenure and promotion, and second, because of the concern that making the data available presents the risk of being "scooped": someone else may use the data to publish results before the original collector of the data has a chance to.

With regard to the first issue, journals like the above-mentioned *Journal of Machine Learning Research* are very relevant. In fields such as computational linguistics and machine learning, there is an increasing number of journals that publish articles describing resources that are available to the community, such as open source software. These journals aim to recognize the scientific importance of such resources, and encourage their creation and promulgation, by providing a publishing vehicle that is available in library-bound print form and maintains a high impact factor – hence addresses the concerns of college committees that make tenure decisions – while also maintaining a quick turnaround time and free online availability. It would be entirely reasonable to establish such a journal in connection with the envisioned corpus, or for such linguistic community resources more generally.

With regard to the second issue, it is ameloriated if one gets publication credit for the preparing and releasing of the data itself. In addition, having other people working on the same data can actually increase one's own productivity, by keeping up a stimulating stream of new ideas. Instead of thinking of other researchers as stealing one's own potential publications, one should think in terms of building on each other's work. Having a ready outlet for publications based on the data in the corpus may help nurture such an atmosphere.

3.7 Intellectual property issues

It is important to release all materials under a single license. As already emphasized, the envisioned corpus is not a library, but a single large data set. Models for such a license exist: the Project Gutenberg license is a prime example, and the Creative Commons license is another possibility. The license should allow appropriate mirroring and redistribution of data, but there is no need to go so far as to put the data in the public domain. Redistributability, the rights of the authors, and simplicity of administration all need to be balanced.

3.8 Long-term archiving

In addition to electronic distribution, the project should collaborate with experts on long-term archiving. Having at least some of the hosting sites be university libraries would enhance the utility and durability of the corpus. We note that the Rosetta Project is hosted by the Stanford University library.

4 What to include in the corpus

Ideally, we would like to have a digitization of each language. The basic principle defining an adequate digitization is that it contains sufficient information to learn the language. To some degree, one can think of the corpus as a linguistic Noah's Ark. But we do not wish to simply duplicate other such efforts; our focus is on machine processability and support for computational research on universal linguistics.

Essentially, the goal is to create a multi-lingual treebank that includes all the world's languages. Said that way, the goal sounds absurdly ambitious. After discussing the data to be collected, we return to the question of viability.

4.1 Primary data

Reflecting on the resources used in language instruction, we desire minimally:

(A1) Original documents:

- (a) Page images of original print documents.
- (b) Documents produced by word processor.
- (c) Audio recordings.

(A2) Unicode texts. Since our focus is on machine processability, we would like Unicode transcriptions of all the original documents. "Transcription" includes both the transcription of page images to Unicode text, possibly with the assistance of OCR, and the transcription of audio recordings to text. Priority is placed on natural text, but elicited sentences will often be more readily available and can be quite useful for illustrating phenomena that occur at low frequency.

(A3) A representation of the meaning of the texts, in one or more of the following forms:

  (a) A simple translation of the text.

  (b) Interlinear glossing. Actually, stand-off glossing is preferable; an interlinear format can be automatically produced on demand.

  (c) A word list with glosses in English or another reference language.

With sufficient text and glosses, one can learn to read and write a language. With sufficient audio recording, one can also learn to speak and understand the spoken language. A **minimal digitization** of a language is sufficient material that a learner who has mastered it would have "everyday competence" in the language. This is obviously a vague description.

4.2 Secondary data

We also aim to produce annotations for the texts:

(B1) Phonetic transcriptions for audio recordings, aligned downward to the audio, and upward to a conventional text transcription.

(B2) Morphological analysis and part of speech assignment.

(B3) Word sense disambiguation. This is equivalent to (A3b) above. The combination of (B2) morphological analysis, (A3b/B3) morph- and word-level glossing, and (A3a) running translation, represents most of the information in typical interlinear glossed text.

(B4) Syntactic structure annotation (treebank).

The production of secondary data is very labor-intensive, and it is *not* anticipated that all primary data will be annotated. Primary data is useful even without annotation. A person given sufficient basic grammatical information (morphological, syntactic, and semantic) can learn additional grammatical information from primary texts, and there are machine learning techniques under active development in computational linguistics for doing the same thing automatically.

4.3 Tools and processing resources

To enable automated processing of the language, certain tools are desirable. Each such tool combines a generic "engine" with language-specific resources. These tools can help produce the primary and secondary data; there is a feedback loop between producing the resources that drive the tools, and using the tools to produce primary and secondary corpus data.

(C1) Speech processing. The development of speech recognition and speech synthesis software is so resource- and labor-intensive that we will not make it a high priority. But there are some resources that are useful even in the absence of software, including:

  (a) phonetic and phonological inventories

  (b) a word list with pronunciations

  (c) a formal phonological grammar that can be used to map between concatenated dictionary pronunciations and the actual phone sequence for synthesis and recognition

(C2) Morphological inference. A morphological inference program takes text and produces a list of morphemes of the language. Very little if any language-specific information is required.

(C3) Morphological analyzer and part of speech tagger. Doing morphological analysis and assigning parts of speech requires:

  (a) morphological paradigms, and

  (b) part of speech dictionaries.

(C4) A formal grammar. This is a syntactic model of the sort discussed in section 1.1, which can be used to automatically assign syntax trees to new sentences. Treebanks are typically produced by using an existing parser to do a rough analysis, which is then edited by hand. One can imagine annotating some sentences manually, inducing a grammar, using it to parse sentences automatically, editing the parsed sentences, and repeating the cycle.

There is research on the automatic transfer of tools from one language to another. There is also a great deal of active research on semi-supervised learning: machine learning in which one has a small amount of manually annotated data and a good deal of unannotated data.

4.4 Documentation

In addition to the machine-processable data listed above, it is important to have good documentation for human consumption. This documentation will include miscellaneous information that does not fit into the formal data schemata. In particular, it might include:

(D1) A traditional dictionary containing more complete definitions, photographs of cultural artifacts, and so on. Note that much of the information in a traditional dictionary has already been listed: (C1a) phonetic and phonological inventories, (C1b) word pronunciations, (C3a) morphological paradigms, (C3b) parts of speech, and (A3c) word glosses.

(D2) A phonetic and phonological description. Relevant items that have already been mentioned include (C1a) phonetic and phonological inventories and (C1c) formal phonological rules.

(D3) A descriptive grammar that arises essentially by organizing and commenting on the information in (B2) the morphological annotations and (B4) the treebank. Via the connection to the treebank, each unit of the grammar has pointers to all the places in the texts where it occurs.

4.5 Viability

Taking all the annotation together, we are essentially proposing to create a treebank containing all the world's languages. Treebanks have already been created for a very few languages; creating a treebank by conventional means requires an enormous investment of effort. Creating conventional treebanks for a large number of languages is not viable.

The main issue is the size of the treebanks. Conventional treebanks are so labor-intensive because they contain a million words or more of text. We will not attempt to do syntactic annotation on anything like that scale. We will not even aim for syntactic annotation of all text in the corpus. (For that matter, we do not expect that all available page images will be converted to manually edited Unicode, nor that all available Unicode text will be morphologically analyzed or glossed.)

Conventional treebanks are designed to enable brute force parser-construction methods. We will aim to enable only more refined computational methods that make heavy use of semisupervised learning. We require *some* annotated text, but we will aim to use it much more efficiently than is done in conventional treebanks.

We will also use active learning methods to pick and choose the kinds of annotation that give as much information as a conventional treebank at

much less labor cost. Supplementing running text with elicited sentences can be seen in this light, as attempting to maximize information benefit for manual labor cost. Similarly, the resources that traditional language documentation focuses on – morphological paradigms, dictionaries, descriptive grammars – can be seen as compact and efficiently-producable substitutes for annotation.

5 Plan

We will require an expert for each language in the corpus, hence a very large number of people. Before we contemplate such a large effort, it is important to work out and thoroughly test an exact specification on a handful of pilot languages. Hence there are two stages to the project: the pilot stage in which the corpus specification is worked out, and a production stage. There should be a process for revising the specification even during the production stage, just as other standards undergo continuous revision, but the rate of change in the specification should be dramatically lower after the pilot stage.

In the pilot stage, we require a handful of people working on individual languages but also pooling experience to refine specifications. The initial set of languages should be typologically diverse. We will require both computational linguists and language experts, working closely together. The aim is to produce flexible generic tools that adapt well to a wide variety of languages, and to produce specifications that work well in individual languages but also support cross-linguistic inquiry.

Tools that need to be created include:

- A universal editor/browser, as described in section 3.3.
- Manual annotation tools for the creation of the various pieces of the corpus listed in section 4.
- Automatic annotation tools, including:
  - morphological inference
  - morphological analysis
  - part of speech tagging, to the extent it is not already accomplished by morphological analysis
  - parsing
- Training components. The tools of the previous item consist of generic engines, and (conceptually separate) training programs are needed to construct the language-specific resources that drive them.

Tools that are of real research interest, but may or may not turn out to be practically viable, include:

- OCR for non-Roman scripts.
- Methods for transferring tools/resources from one language to another.
- The use of bootstrapping and semisupervised learning methods to increase efficiency of tool/resource development.

For each tool and each corpus component, detailed specifications need to be worked out. This includes:

- file formats
- tagsets
- syntactic annotation schemes

Finally, we also need to work out:

- redistribution license
- editorial process

Bibliography

References

[1] Bloomfield, Leonard. *Language.* Holt, New York. 1933.

[2] Human Genome Project. The science behind the Human Genome Project. `http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml`