A Universal Corpus

Steven Abney

February 16, 2007

1 A modest proposal

- As a computational linguist, it seems obvious that the most urgent task facing linguistics is
  - Construction of **universal** corpus
  - containing **primary data**
  - with description in the form of **consistent annotation**
- Went to Digital Tools/EMELD workshops, surprised at lack of plans

1.1 What's so good about corpora?

- Heavy use of corpora basic to modern computational linguistics
  - Science is about prediction and control
  - Easy to come up with ideas that fit existing data, acid test is prediction of completely new data
  - Training and testing
  - Disambiguation
  - Learning
  - Language is a computational system
    - ∗ Individual: psychology
    - ∗ Aggregate: behavior of complex system
- Chomsky's original vision
  - Developing large-scale generative grammars of individual languages
  - Developing meta-grammar + learning methods to generalize to all languages
  - Amplification of effort by machine assistance
  - Automatic testing of grammar coverage

    - Automatic learning methods
    - Automatic transference to new languages

- Competence and performance
    - Grammar, dialogue context, world knowledge, etc. are all factors in prediction
    - Factoring of the problem is useful, but the factors need to fit together in a whole
    - Autonomy originally meant that syntax is not reducible to semantics, NOT that syntax is isolated from the rest of language, like a calcified foreign body embedded in the brain
    - Goal is understanding of complete system, and computational linguistics has matured to a point where the full picture can be tackled

- Systematic linguistics
    - Finding facts here and there that agree with one's predictions is no trick
    - The important thing is systematicity
    - The periodic table
    - The Linnaean taxonomy,
    - the Human Genome Project
    - systematic sky surveys
    - systematic automated sorting through collisions in a cyclotron

- How corpora are used in computational linguistics
    - Building parsers – current best parsers trained on the Penn Treebank
    - Building speech recognition systems
    - Building speech synthesis systems
    - Building machine translation systems
    - Building information extraction (language understanding) systems
    - Inducing morphology
    - Inducing phrase structure
    - Inducing subcategorization frames, selectional restrictions, semantic roles

- As in physics
    - Build experimental apparatus = experimental protocol (text collection, annotation protocol)
    - Experiment is a data source

  - Estimate model parameters (learn) from training data
  - Acid test is quality of prediction of new data from the source
  - Physics:
    * Experiment produces giant collection of collisions
    * program parses the "images" of the collisions, can account for all but a tiny fraction
    * That tiny fraction goes to humans to see if something new has been found
  - Comp Ling:
    * Annotation protocol produces collection of judgments regarding sentence structure, outside of control of experimenter
    * Current parsers only get about 90% of *phrases* correct – which means most sentence contain at least one error

- Language is messy
  - The clean, mathematical principles are in the learning algorithm

- Why not let them die?
  - Human psychology isn't changing, and when we unravel it, the languages are pure epiphenomenon.
  - That's like saying that once we have worked out a complete theory of physics, we can discard the universe. The essence of language is only partially in the psychology; it is also in the community – the same way that behavior is only partially in the genes, it is also in the environment. More than that, language is larger than an idiolect. It is an aggregate object, in the same way that a galaxy is an aggregate of stars or a body is an aggregate of cells. There is more to anatomy than just cell anatomy.

- From "The science behind the Human Genome Project":

  > 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.

- Incidentally, the human genome contains 3200 million base pairs, or 800 MB. A small (1MW) corpus of one language is 6.3 MB without annotation; the genome is the size of 127 languages.

1.2 The Aim

- What:
  - A **digitization** of (ideally) every human language
  - Corpus (primary data), grammar and lexicon (secondary data)
  - Every lexical entry and grammar statement linked (ideally) to occurrences in corpus, and vice versa
  - Speech and written texts
  - Community process, community resource
- Why?
  - Archiving languages that are disappearing, in a way that serves future research needs
  - The corpus does not aim to be the ultimate Noah's Ark; the more important thing is the appeal to field linguists to collect more primary data, not focus so heavily on secondary description (grammars and lexica)
  - Resource to language communities
  - Systematic linguistics

1.3 Principles

- **Universality.**
  - first goal: "minimum sample" from 1,000 languages.
  - at least a token amount of data from every recorded language.
- **Machine readability and consistency.**
  - enabling new types of linguistic inquiry
  - machine processing across many languages
  - Materials intended to be read by humans are not suitable
- **Community ownership.**
  - cannot expect a small group to assemble a resource of this scale
  - community buy-in is essential
  - repository is not possession of any one institution
    - mirrored at multiple sites.
    - also gives data security through replication
- **Accessibility.**
  - owned by the community, be freely available to the entire community

  - intellectual property rights
    * Wikipedia/Gutenberg consent – only unencumbered material is included
  - sufficient editorial control for quality assurance
    * wide-open Wikipedia section
    * vetted section
  - no limits on community members' ability to obtain, enhance, and redistribute the corpus.
  - mirrored at multiple locations
  - entire corpus should be downloadable.

- **Objectivity.**
  - objective and descriptive, not an encoding of theoretical hypotheses
  - style of the better descriptive grammars or instructional materials
  - plain, unbiased, and clearly-defined terms
  - should not require commitments to any particular school of linguistics.
  - IPA for morphology, syntax

- **Replicability of analyses.**
  - grammars and lexica are **secondary data**, not primary data
  - primary data on which they are are based should be included
  - One should be able to reproduce every step of any analysis, from primary data.
  - Never throw away the originals (unlike e.g. Penn Treebank)

2 Reinventing the wheel?

- Isn't this already being done?

2.1 Existing archival efforts

- Build on existing efforts; community effort
- EMELD
  - project has just ended
  - Aim is very similar:

> “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.”

  - best practices
  - GOLD ontology – observational vocabulary
  - No aim to assemble corpus
  - GOLD *does* have a community process, EMELD does not
  - No community ownership
  - No downloadable corpus

- Rosetta project
  - Collecting data from as many languages as possible
  - 2376 languages
  - Funded by NSF and private donations
  - Hosted by Stanford University Libraries
  - Grammars, texts, vocabularies – typically are scanned print documents
  - Not machine readable
  - No provision for download
  - No mirroring
  - No community process
- DARPA funded effort to create translation corpora
  - “surprise language” task
  - Few tens of languages
  - in service of quick development of machine translation systems
  - data available through LDC
    - ∗ significant cost
  - Priority is national security interests, not endangerment
- SIL
- LDC

- Traditional print archives
  - Archive of the Indigenous Languages of Latin America (AILLA)
  - Alaska Native Languages Archive
  - “Digitizing” a collection generally means simply transfer to electronic media, not the creation of data sets that support automated processing.
  - Often heavy restrictions on access
- OLAC
  - Search and cataloging, not primary data
- ODIN
  - Search for interlinear text
  - “Virtual” corpus
  - Not downloadable
  - No editting
  - No consistency
- Terry: Why isn’t it already solved by Google? (Failure of American Language Corpus to get funded)
  - editorial process
  - preprocessing
  - PDF → machine readable
  - consistency
- Just another archive? No, differs from an archive in:
  - consistency across language
  - upper limit on needed sample from each language. If more material is available than needed, choose based on quality

2.2 Bird and Simons

- Bird and Simons “dimensions of portability”:
  - Content
    - breadth of coverage
    - quality
    - verifiability of description against primary data
  - Format
    - Openness

        - Encoding – use Unicode and IPA
        - Markup – use XML and TEI
        - Rendering – need portable fonts
    - Discovery – OLAC
    - Access
        - making primary data available
        - degree of difficulty of obtaining access
    - Citation
        - Citation standards
        - Persistence – the cited object doesn't move
        - Immutability – the cited object doesn't change
        - Granularity – citing items or parts of items
    - Preservation
    - Rights
        - Balancing benefit of use with protection of sensitive data

- Focuses on overall practice in language documentation and description
    - Documentation = collecting primary data
    - Description = secondary data, i.e., analysis (grammars and lexica)
    - Highly recommended reading
    - Current archives are designed for researchers interested in particular languages or at most particular language families

- Our aim is support of universal linguistics
    - Need a single repository so that one does not have to laboriously collect resources for all 6,000 languages
    - Need higher level of consistency so that one does not have to learn a new annotation scheme for each language
    - Smaller in scope: a digitization of each language, not an archiving of all materials available for each language

3 Why would anyone use it?

3.1 Who benefits?

- Community buy-in is essential
    - Community is aware of the importance of documenting endangered languages

  - Communicating what computational methods make possible, that would be prohibitively expensive or impossible using traditional print archives.
  - Both increasing pull and lowering barriers – Tools

- Linguists doing cross-linguistic research

- Communities of speakers of minority languages
  - Directly from materials in their language
  - Indirectly via economies of scale for the development of software etc for language instruction and preservation
  - The standards required for consistency across the corpus enable the development of truly language-universal software.
  - Such software currently does not exist
    - No current browsers have fonts and display methods that support all of Unicode. Required for uniform access to a universal corpus.

- Field linguists
  - Visibility
  - Archiving
  - Similar economies of scale in the development of software to support field work.
  - Conversion of legacy materials to standard machine-readable format
    - Recording data in Microsoft Word is like throwing it in the trash
    - The file format is proprietary, and Microsoft changes it with every new version of Word
    - Not a property of computer files, but specifically of bad formats
    - ASCII has been around since 1963. ASCII files written in 1963 are just as readable today as then. Unicode is an extension of ASCII.

3.2 Lowering barriers

- Make it easy for author to edit/update/delete data, to address embarrassment at poorly editted data

- Getting critical mass

- Promise of tools to ease automation. Add-on stand-off annotation (value added, nothing lost)

- Helen: tools for converting legacy data in exchange for including data in the corpus

- For native communities: don't have resources to produce tool. Tool produced by economy of scale for all communities
- Tenure
  - Getting contributions to corpus recognized as publication
  - ”LSA endorses use of electronic sources as ciations for tenure”

4 What to include in the corpus

- A “digitization” of each language.
  - Basic principle: sufficient information to learn the language
  - Proof of possibility: Hebrew, Akkadian
  - Excluding Christian literature, 10M words of classical Greek, 1M words of classical Latin survive. Estimated that largest Roman libraries contained 20M words, excluding duplicate works.
  - 10M words = 63M bytes * 6000 languages = 378 GB
  - Learning a language is a matter of degree. Minimally: “everyday competence”
  - Maximize fidelity of the image, the variety of future research that it supports
  - “Primer” version of the language – what you'd find in a first-year course in the language
- Minimally
  - Running text, either in orthography or phonemic transcription
  - Translation into English or other widely-known language
  - A dictionary giving the meaning of (ideally) every word in the texts
  - Audio recordings, at least some of which are transcribed
- Additional resources
  - Morphological paradigms
  - Interlinear glossing for at least some of the text example of IGT
    * Morphological analysis of words
    * Morpheme-by-morpheme categories and English glosses
    * Sentence-by-sentence English gloss
  - Syntactic structure annotation (treebank)
  - Grammatical description with pointers to examples in the texts
  - Formal grammars that support parsing and generation

  - Elicited sentences illustrating particular grammatical phenomena
  - Photographs of cultural artifacts to augment English glosses of cultural terms

- Highest priority on endangered languages

5 How to build it

5.1 Build on existing data and standards

- Where to get data from?
  - Language communities. Make it easy for them to put web-based language learning materials up; include those materials in the corpus.
  - Archives - share data with them, share tools with them
  - Tool efforts part of Digital Tool Summit
  - Coordinate with Rosetta
  - Individual field researchers - provide tools to make it easy to contribute
  - Wiktionaries
  - scanning/OCR – OCR is research issue
- Mining the web
  - It is possible to build minority-language corpora (Ghani et al)
  - ODIN
  - But permissions for redistribution are difficult to impossible
  - Distribute tools, not data
- Research: cross-linguistic transfer
  - Parsing glosses to bootstrap a parse tree (Will Lewis)
  - Yarowsky & students' work on transfering across languages
- Induction
  - E.g., Goldsmith, Finch
- Semi-supervised learning

5.2 Browser-editor-annotator

- Tool
  - to access corpus
  - low barriers to use
  - handles all scripts without installing anything extra
  - runs on any platform
  - runs in web browser or downloadable
- Project Perseus
  - Linked lexicon, corpus, morphological analysis
  - Good for language learning
- Digital Tools Summit
  - Not a corpus
  - Focus on tools for field work
  - Focus on human consumption of individual languages
- Tool should not be emphasis
  - Focus on data standard, as in web: single standard, competing browsers
  - Unlike e.g. Shoebox
  - Conversion among formats

5.3 Community process

- Community property
  - Freely available
  - Anyone can contribute, subject to editorial review
  - Editorial committee, not just one person
  - Anyone can mirror and redistribute, subject to license
  - Necessary to get permission from each contributor
- Model: Wikipedia
  - Anyone can edit anything
  - Anything you contribute becomes community property
  - Nothing is lost: history of earlier versions is kept
  - Vandalism occurs
  - Anyone can undo an edit

  - Sysops monitor changed pages, can block vandals
  - Each page has an associated talk page for resolving differences of opinion
  - Editorial policy (factuality, citations, no original research, etc)

- Model: W3C
  - Sets web standards, browsers implement them (as they please)
  - Members express interest, director creates activity proposal. If there is sufficient interest, a working group is created.
  - Working group creates recommendation, cycles of member and public review, revision
  - The Advisory Committee publishes final standards ("Recommendations")
  - Only members may participate
  - One must apply for membership, it will not necessarily be granted
  - Currently there are 434 members
  - Membership fee: $6350–$63,500, depending on annual gross revenue of organization

- Model: Project Gutenberg
  - Anyone can contribute: scanning, editting
  - Only public domain texts, or texts with permission of copyright holder
  - Entire body of texts can be downloaded
  - Anyone is allowed to set up a mirror site
  - Public domain texts are freely redistributable
  - Multiple languages
  - Some audio documents (read texts)

- Model: open software (e.g., SourceForge)
  - Only members may contribute
  - Public may apply for membership
  - Anyone may download and use
  - Generally two versions: stable, beta

- Quality assurance. You get what you pay for. Won't the people who contribute be the fruitcakes who are just looking for attention? Serious people are too busy.

- Balancing community ownership with quality

  - Some items are “frozen,” with standard authorship attribution
  - Some are collaborative with Wiki-like model
  - Keeping attributions of all contributions
  - But distinguish stable from beta versions

5.4 How can you hope to get agreement?

- Getting linguists to cooperate is like herding cats. How can you hope to get linguists to agree on “observational language”?
- Generalize from existing treebanks: Penn, Susanne, Czech, Chinese, Middle English, Old English.
- Layers (stand-off annotation, not deletion) and mappings.

5.5 Long-term archiving.

Partnership with library?

- Pilot datasets
- Start archive at UM Library? Collaborating with SI? Digital library?
- Estimate cost of collecting corpus
  - Annotation
- Get funding via learning support for communities? NEH, UNESCO?
- Create a web site with white paper
- GNU license?