The Human Language Project: Building a Universal Corpus of the World's Languages

Steven Abney
University of Michigan
abney@umich.edu

Steven Bird
University of Melbourne and
University of Pennsylvania
sbird@unimelb.edu.au

Abstract

We present a grand challenge to build a corpus that will include all of the world's languages, in a consistent structure that permits large-scale cross-linguistic processing, enabling the study of universal linguistics. The focal data types, bilingual texts and lexicons, relate each language to one of a set of reference languages. We propose that the ability to train systems to translate into and out of a given language be the yardstick for determining when we have successfully captured a language. We call on the computational linguistics community to begin work on this Universal Corpus, pursuing the many strands of activity described here, as their contribution to the global effort to document the world's linguistic heritage before more languages fall silent.

1 Introduction

The grand aim of linguistics is the construction of a universal theory of human language. To a computational linguist, it seems obvious that the first step is to collect significant amounts of primary data for a large variety of languages. Ideally, we would like a complete digitization of every human language: a Universal Corpus.

If we are ever to construct such a corpus, it must be now. With the current rate of language loss, we have only a small window of opportunity before the data is gone forever. Linguistics may be unique among the sciences in the crisis it faces. The next generation will forgive us for the most egregious shortcomings in theory construction and technology development, but they will not forgive us if we fail to preserve vanishing primary language data in a form that enables future research.

The scope of the task is enormous. At present, we have non-negligible quantities of machine-readable data for only about 20–30 of the world's 6,900 languages (Maxwell and Hughes, 2006). Linguistics as a field is awake to the crisis. There has been a tremendous upsurge of interest in documentary linguistics, the field concerned with the the "creation, annotation, preservation, and dissemination of transparent records of a language" (Woodbury, 2010). However, documentary linguistics alone is not equal to the task. For example, no million-word machine-readable corpus exists for any endangered language, even though such a quantity would be necessary for wide-ranging investigation of the language once no speakers are available. The chances of constructing large-scale resources will be greatly improved if computational linguists contribute their expertise.

This collaboration between linguists and computational linguists will extend beyond the construction of the Universal Corpus to its exploitation for both theoretical and technological ends. We envisage a new paradigm of universal linguistics, in which grammars of individual languages are built from the ground up, combining expert manual effort with the power tools of probabilistic language models and grammatical inference. A universal grammar captures redundancies which exist across languages, constituting a "universal linguistic prior," and enabling us to identify the distinctive properties of specific languages and families. The linguistic prior and regularities due to common descent enable a new economy of scale for technology development: cross-linguistic triangulation can improve performance while reducing per-language data requirements.

Our aim in the present paper is to move beyond generalities to a concrete plan of attack, and to challenge the field to a communal effort to create a Universal Corpus of the world's languages, in consistent machine-readable format, permitting large-scale cross-linguistic processing.

2 Human Language Project

2.1 Aims and scope

Although language endangerment provides urgency, the corpus is not intended primarily as a Noah's Ark for languages. The aims go beyond the current crisis: we wish to support cross-linguistic research and technology development at the largest scale. There are existing collections that contain multiple languages, but it is rare to have consistent formats and annotation across languages, and few such datasets contain more than a dozen or so languages.

If we think of a multi-lingual corpus as consisting of an array of items, with columns representing languages and rows representing resource types, the usual focus is on "vertical" processing. Our particular concern, by contrast, is "horizontal" processing that cuts indiscriminately across languages. Hence we require an unusual degree of consistency across languages.

The kind of processing we wish to enable is much like the large-scale systematic research that motivated 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 ... they can approach questions systematically and on a grand scale. They can study ... how tens of thousands of genes and proteins work together in interconnected networks to orchestrate the chemistry of life. (Human Genome Project, 2007)

We wish to make it possible to investigate human language equally systematically and on an equally grand scale: a Human Linguome Project, as it were, though we have chosen the "Human Language Project" as a more inviting title for the undertaking. The product is a Universal Corpus,[1] in two senses of *universal:* in the sense of including (ultimately) all the world's languages, and in the sense of enabling software and processing methods that are language-universal.

However, we do *not* aim for a collection that is universal in the sense of encompassing all language documentation efforts. Our goal is the construction of a specific resource, albeit a very large resource. We contrast the proposed effort with general efforts to develop open resources, standards, and best practices. We do *not* aim to be all-inclusive. The project does require large-scale collaboration, and a task definition that is simple and compelling enough to achieve buy-in from a large number of data providers. But we do not need and do not attempt to create consensus across the entire community. (Although one can hope that what proves successful for a project of this scale will provide a good foundation for future standards.)

Moreover, we do not aim to collect data merely in the vague hope that it will prove useful. Although we strive for maximum generality, we also propose a specific driving "use case," namely, machine translation (MT), (Hutchins and Somers, 1992; Koehn, 2010). The corpus provides a testing ground for the development of MT system-construction methods that are dramatically "leaner" in their resource requirements, and which take advantage of cross-linguistic bootstrapping. The large engineering question is how one can turn the size of the task—constructing MT systems for all the world's languages simultaneously—to one's advantage, and thereby consume dramatically less data per language.

The choice of MT as the use case is also driven by scientific considerations. To explain, we require a bit of preamble.

We aim for a *digitization* of each human language. What exactly does it mean to digitize an entire language? It is natural to think in terms of replicating the body of resources available for well-documented languages, and the pre-eminent resource for any language is a treebank. Producing a treebank involves a staggering amount of manual effort. It is also notoriously difficult to obtain agreement about how parse trees should be defined in one language, much less in many languages simultaneously. The idea of producing treebanks for 6,900 languages is quixotic, to put it mildly. But is a treebank actually necessary?

Let us suppose that the purpose of a parse tree is to mediate interpretation. A treebank, arguably, represents a theoretical hypothesis about how interpretations could be constructed; the primary data is actually the interpretations themselves. This suggests that we annotate sentences with representations of meanings instead of syntactic structures. Now that seems to take us out of the frying pan into the fire. If obtaining consen-

[1] http://universalcorpus.org/

sus on parse trees is difficult, obtaining consensus on meaning representations is impossible. However, if the language under consideration is anything other than English, then a translation into English (or some other reference language) is for most purposes a perfectly adequate meaning representation. That is, we view machine translation as an approximation to language understanding.

Here is another way to put it. One measure of adequacy of a language digitization is the ability of a human—already fluent in a reference language—to acquire fluency in the digitized language using only archived material. Now it would be even better if we could use a language digitization to construct an *artificial speaker* of the language. Importantly, we do not need to solve the AI problem: the speaker need not decide *what* to say, only how to translate from meanings to sentences of the language, and from sentences back to meanings. Taking sentences in a reference language as the meaning representation, we arrive back at machine translation as the measure of success. *In short, we have successfully captured a language if we can translate into and out of the language.*

The key resource that should be built for each language, then, is a collection of primary texts with translations into a reference language. "Primary text" includes both written documents and transcriptions of recordings. Large volumes of primary texts will be useful even without translation for such tasks as language modeling and unsupervised learning of morphology. Thus, we anticipate that the corpus will have the usual "pyramidal" structure, starting from a base layer of unannotated text, some portion of which is translated into a reference language at the document level to make the next layer. Note that, for maximally authentic primary texts, we assume the direction of translation will normally be from primary text to reference language, not the other way around.

Another layer of the corpus consists of sentence and word alignments, required for training and evaluating machine translation systems, and for extracting bilingual lexicons. Curating such annotations is a more specialized task than translation, and so we expect it will only be done for a subset of the translated texts.

In the last and smallest layer, morphology is annotated. This supports the development of morphological analyzers, to preprocess primary texts to identify morpheme boundaries and recognize allomorphs, reducing the amount of data required for training an MT system. This most-refined target annotation corresponds to the *interlinear glossed texts* that are the de facto standard of annotation in the documentary linguistics community.

We postulate that interlinear glossed text is sufficiently fine-grained to serve our purposes. It invites efforts to enrich it by automatic means: for example, there has been work on parsing the English translations and using the word-by-word glosses to transfer the parse tree to the object language, effectively creating a treebank automatically (Xia and Lewis, 2007). At the same time, we believe that interlinear glossed text is sufficiently simple and well-understood to allow rapid construction of resources, and to make cross-linguistic consistency a realistic goal.

Each of these layers—primary text, translations, alignments, and morphological glosses—seems to be an unavoidable piece of the overall solution. The fact that these layers will exist in diminishing quantity is also unavoidable. However, there is an important consequence: the primary texts will be permanently subject to new translation initiatives, which themselves will be subject to new alignment and glossing initiatives, in which each step is an instance of semisupervised learning (Abney, 2007). As time passes, our ability to enhance the quantity and quality of the annotations will only increase, thanks to effective combinations of automatic, professional, and crowd-sourced effort.

2.2 Principles

The basic principles upon which the envisioned corpus is based are the following:

Universality. Covering as many languages as possible is the first priority. Progress will be gauged against concrete goals for numbers of languages, data per language, and coverage of language families (Whalen and Simons, 2009).

Machine readability and consistency. "Covering" languages means enabling machine processing seamlessly across languages. This will support new types of linguistic inquiry and the development and testing of inference methods (for morphology, parsers, machine translation) across large numbers of typologically diverse languages.

Community effort. We cannot expect a single organization to assemble a resource on this scale. It will be necessary to get community buy-in, and

many motivated volunteers. The repository will not be the sole possession of any one institution.

Availability. The content of the corpus will be available under one or more permissive licenses, such as the Creative Commons Attribution License (CC-BY), placing as few limits as possible on community members' ability to obtain and enhance the corpus, and redistribute derivative data.

Utility. The corpus aims to be maximally useful, and minimally parochial. Annotation will be as lightweight as possible; richer annotations will will emerge bottom-up as they prove their utility at the large scale.

Centrality of primary data. Primary texts and recordings are paramount. Secondary resources such as grammars and lexicons are important, but no substitute for primary data. It is desirable that secondary resources be integrated with—if not derived from—primary data in the corpus.

2.3 What to include

What should be included in the corpus? To some extent, data collection will be opportunistic, but it is appropriate to have a well-defined target in mind. We consider the following essential.

Metadata. One means of resource identification is to survey existing documentation for the language, including bibliographic references and locations of web resources. Provenance and proper citation of sources should be included for all data.

For written text. (1) Primary documents in original printed form, e.g. scanned page images or PDF. (2) Transcription. Not only optical character recognition output, but also the output of tools that extract text from PDF, will generally require manual editing.

For spoken text. (1) Audio recordings. Both elicited and spontaneous speech should be included. It is highly desirous to have some connected speech for every language. (2) Slow speech "audio transcriptions." Carefully respeaking a spoken text can be much more efficient than written transcription, and may one day yield to speech recognition methods. (3) Written transcriptions. We do not impose any requirements on the form of transcription, though orthographic transcription is generally much faster to produce than phonetic transcription, and may even be more useful as words are represented by normalized forms.

For both written and spoken text. (1) Translations of primary documents into a reference language (possibly including commentary). (2) Sentence-level segmentation and translation. (3) Word-level segmentation and glossing. (4) Morpheme-level segmentation and glossing.

All documents will be included in primary form, but the percentage of documents with manual annotation, or manually corrected annotation, decreases at increasingly fine-grained levels of annotation. Where manual fine-grained annotation is unavailable, automatic methods for creating it (at a lower quality) are desirable. Defining such methods for a large range of resource-poor languages is an interesting computational challenge.

Secondary resources. Although it is possible to base descriptive analyses exclusively on a text corpus (Himmelmann, 2006, p. 22), the following secondary resources should be secured if they are available: (1) A lexicon with glosses in a reference language. Ideally, everything should be attested in the texts, but as a practical matter, there will be words for which we have only a lexical entry and no instances of use. (2) Paradigms and phonology, for the construction of a morphological analyzer. Ideally, they should be inducible from the texts, but published grammatical information may go beyond what is attested in the text.

2.4 Inadequacy of existing efforts

Our key desideratum is support for automatic processing across a large range of languages. No data collection effort currently exists or is proposed, to our knowledge, that addresses this desideratum. Traditional language archives such as the Audio Archive of Linguistic Fieldwork (UC Berkeley), Documentation of Endangered Languages (Max Planck Institute, Nijmegen), the Endangered Languages Archive (SOAS, University of London), and the Pacific And Regional Archive for Digital Sources in Endangered Cultures (Australia) offer broad coverage of languages, but the majority of their offerings are restricted in availability and do not support machine processing. Conversely, large-scale data collection efforts by the Linguistic Data Consortium and the European Language Resources Association cover less than one percent of the world's languages, with no evident plans for major expansion of coverage. Other efforts concern the definition and aggregation of language resource metadata, including OLAC, IMDI, and

CLARIN (Simons and Bird, 2003; Broeder and Wittenburg, 2006; Váradi et al., 2008), but this is not the same as collecting and disseminating data.

Initiatives to develop standard formats for linguistic annotations are orthogonal to our goals. The success of the project will depend on contributed data from many sources, in many different formats. Converting all data formats to an official standard, such as the RDF-based models being developed by ISO Technical Committee 37 Sub-committee 4 Working Group 2, is simply impractical. These formats have onerous syntactic and semantic requirements that demand substantial further processing together with expert judgment, and threaten to crush the large-scale collaborative data collection effort we envisage, before it even gets off the ground. Instead, we opt for a very lightweight format, sketched in the next section, to minimize the effort of conversion and enable an immediate start. This does not limit the options of community members who desire richer formats, since they are free to invest the effort in enriching the existing data. Such enrichment efforts may gain broad support if they deliver a tangible benefit for cross-language processing.

3 A Simple Storage Model

Here we sketch a simple approach to storage of texts (including transcribed speech), bitexts, interlinear glossed text, and lexicons. We have been deliberately schematic since the goal is just to give grounds for confidence that there exists a general, scalable solution.

For readability, our illustrations will include space-separated sequences of tokens. However, behind the scenes these could be represented as a sequence of pairs of start and end offsets into a primary text or speech signal, or as a sequence of integers that reference an array of strings. Thus, when we write (1a), bear in mind it may be implemented as (1b) or (1c).

(1) a. This is a point of order .
    b. (0,4), (5,7), (8,9), (10,15), (16,18), ...
    c. 9347, 3053, 0038, 3342, 3468, ...

In what follows, we focus on the minimal requirements for storing and disseminating aligned text, not the requirements for efficient in-memory data structures. Moreover, we are agnostic about whether the normalized, tokenized format is stored entire or computed on demand.

We take an aligned text to be composed of a series of aligned sentences, each consisting of a small set of attributes and values, e.g.:

```
ID: europarl/swedish/ep-00-01-17/18
LANGS: swd eng
SENT: det gäller en ordningsfråga
TRANS: this is a point of order
ALIGN: 1-1 2-2 3-3 4-4 4-5 4-6
PROVENANCE: pharaoh-v1.2, ...
REV: 8947 2010-05-02 10:35:06 leobfld12
RIGHTS: Copyright (C) 2010 Uni...; CC-BY
```

The value of `ID` identifies the document and sentence, and any collection to which the document belongs. Individual components of the identifier can be referenced or retrieved. The `LANGS` attribute identifies the source and reference language using ISO 639 codes.[2] The `SENT` attribute contains space-delimited tokens comprising a sentence. Optional attributes `TRANS` and `ALIGN` hold the translation and alignment, if these are available; they are omitted in monolingual text. A provenance attribute records any automatic or manual processes which apply to the record, and a revision attribute contains the version number, timestamp, and username associated with the most recent modification of the record, and a rights attribute contains copyright and license information.

When morphological annotation is available, it is represented by two additional attributes, `LEX` and `AFF`. Here is a monolingual example:

```
ID: example/001
LANGS: eng
SENT: the dogs are barking
LEX: the dog be bark
AFF: - PL PL ING
```

Note that combining all attributes of these two examples—that is, combining word-by-word translation with morphological analysis—yields interlinear glossed text.

A bilingual lexicon is an indispensable resource, whether provided as such, induced from a collection of aligned text, or created by merging contributed and induced lexicons. A bilingual lexicon can be viewed as an inventory of cross-language correspondences between words or groups of words. These correspondences are just aligned text fragments, albeit much smaller than a sentence. Thus, we take a bilingual lexicon to be a kind of text in which each record contains a single lexeme and its translation, represented using the `LEX` and `TRANS` attributes we have already introduced, e.g.:

[2] `http://www.sil.org/iso639-3/`

```
ID: swedishlex/v3.2/0419
LANGS: swd eng
LEX: ordningsfråga
TRANS: point of order
```

In sum, the Universal Corpus is represented as a massive store of records, each representing a single sentence or lexical entry, using a limited set of attributes. The store is indexed for efficient access, and supports access to slices identified by language, content, provenance, rights, and so forth. Many component collections would be "unioned" into this single, large Corpus, with only the record identifiers capturing the distinction between the various data sources.

Special cases of aligned text and wordlists, spanning more than 1,000 languages, are Bible translations and Swadesh wordlists (Resnik et al., 1999; Swadesh, 1955). Here there are obvious use-cases for accessing a particular verse or word across all languages. However, it is not necessary to model n-way language alignments. Instead, such sources are implicitly aligned by virtue of their structure. Extracting all translations of a verse, or all cognates of a Swadesh wordlist item, is an index operation that returns monolingual records, e.g.:

```
ID: swadesh/47          ID: swadesh/47
LANGS: fra              LANGS: eng
LEX: chien              LEX: dog
```

4 Building the Corpus

Data collection on this scale is a daunting prospect, yet it is important to avoid the paralysis of over-planning. We can start immediately by leveraging existing infrastructure, and the voluntary effort of interested members of the language resources community. One possibility is to found a "Language Commons," an open access repository of language resources hosted in the Internet Archive, with a lightweight method for community members to contribute data sets.

A fully processed and indexed version of selected data can be made accessible via a web services interface to a major cloud storage facility, such as Amazon Web Services. A common query interface could be supported via APIs in multiple NLP toolkits such as NLTK and GATE (Bird et al., 2009; Cunningham et al., 2002), and also in generic frameworks such as UIMA and SOAP, leaving developers to work within their preferred environment.

4.1 Motivation for data providers

We hope that potential contributors of data will be motivated to participate primarily by agreement with the goals of the project. Even someone who has specialized in a particular language or language family maintains an interest, we expect, in the universal question—the exploration of Language writ large.

Data providers will find benefit in the availability of volunteers for crowd-sourcing, and tools for (semi-)automated quality control, refinement, and presentation of data. For example, a data holder should be able to contribute recordings and get help in transcribing them, through a combination of volunteer labor and automatic processing.

Documentary linguists and computational linguists have much to gain from collaboration. In return for the data that documentary linguistics can provide, computational linguistics has the potential to revolutionize the tools and practice of language documentation.

We also seek collaboration with communities of language speakers. The corpus provides an economy of scale for the development of literacy materials and tools for interactive language instruction, in support of language preservation and revitalization. For small languages, literacy in the mother tongue is often defended on the grounds that it provides the best route to literacy in the national language (Wagner, 1993, ch. 8). An essential ingredient of any local literacy program is to have a substantial quantity of available texts that represent familiar topics including cultural heritage, folklore, personal narratives, and current events. Transition to literacy in a language of wider communication is aided when transitional materials are available (Waters, 1998, pp. 61ff). Mutual benefits will also flow from the development of tools for low-cost publication and broadcast in the language, with copies of the published or broadcast material licensed to and archived in the corpus.

4.2 Roles

The enterprise requires collaboration of many individuals and groups, in a variety of roles.

Editors. A critical group are people with sufficient engagement to serve as editors for particular language families, who have access to data or are able to negotiate redistribution rights, and oversee the workflow of transcription, translation, and annotation.

CL Research. All manual annotation steps need to be automated. Each step presents a challenging semi-supervised learning and cross-linguistic bootstrapping problem. In addition, the overall measure of success—induction of machine translation systems from limited resources—pushes the state of the art (Kumar et al., 2007). Numerous other CL problems arise: active learning to improve the quality of alignments and bilingual lexicons; automatic language identification for low-density languages; and morphology learning.

Tool builders. We need tools for annotation, format conversion, spidering and language identification, search, archiving, and presentation. Innovative crowd-sourcing solutions are of particular interest, e.g. web-based functionality for transcribing audio and video of oral literature, or setting up a translation service based on aligned texts for a low-density language, and collecting the improved translations suggested by users.

Volunteer annotators. An important reason for keeping the data model as lightweight as possible is to enable contributions from volunteers with little or no linguistic training. Two models are the volunteers who scan documents and correct OCR output in Project Gutenberg, or the undergraduate volunteers who have constructed Greek and Latin treebanks within Project Perseus (Crane, 2010). Bilingual lexicons that have been extracted from aligned text collections might be corrected using crowd-sourcing, leading to improved translation models and improved alignments. We also see the Universal Corpus as an excellent opportunity for undergraduates to participate in research, and for native speakers to participate in the preservation of their language.

Documentary linguists. The collection protocol known as Basic Oral Language Documentation (BOLD) enables documentary linguists to collect 2–3 orders of magnitude more oral discourse than before (Bird, 2010). Linguists can equip local speakers to collect written texts, then to carefully "respeak" and orally translate the texts into a reference language. With suitable tools, incorporating active learning, local speakers could further curate bilingual texts and lexicons. An early need is pilot studies to determine costings for different categories of language.

Data agencies. The LDC and ELRA have a central role to play, given their track record in obtaining, curating, and publishing data with licenses that facilitate language technology development. We need to identify key resources where negotiation with the original data provider, and where payment of all preparation costs plus compensation for lost revenue, leads to new material for the Corpus. This is a new publication model and a new business model, but it can co-exist with the existing models.

Language archives. Language archives have a special role to play as holders of unique materials. They could contribute existing data in its native format, for other participants to process. They could give bilingual texts a distinct status within their collections, to facilitate discovery.

Funding agencies. To be successful, the Human Language Project would require substantial funds, possibly drawing on a constellation of public and private agencies in many countries. However, in the spirit of starting small, and starting now, agencies could require that sponsored projects which collect texts and build lexicons contribute them to the Language Commons. After all, the most effective time to do translation, alignment, and lexicon work is often at the point when primary data is first collected, and this extra work promises direct benefits to the individual project.

4.3 Early tasks

Seed corpus. The central challenge, we believe, is getting critical mass. Data attracts data, and if one can establish a sufficient seed, the effort will snowball. We can make some concrete proposals as to how to collect a seed. Language resources on the web are one source—the Crúbadán project has identified resources for 400 languages, for example (Scannell, 2008); the New Testament of the Bible exists in about 1200 languages and contains of the order of 100k words. We hope that existing efforts that are already well-disposed toward electronic distribution will participate. We particularly mention the Language and Culture Archive of the Summer Institute of Linguistics, and the Rosetta Project. The latter is already distributed through the Internet Archive and contains material for 2500 languages.

Resource discovery. Existing language resources need to be documented, a large un-

dertaking that depends on widely distributed knowledge. Existing published corpora from the LDC, ELRA and dozens of other sources—a total of 85,000 items—are already documented in the combined catalog of the Open Language Archives Community,[3] so there is no need to recreate this information. Other resources can be logged by community members using a public access wiki, with a metadata template to ensure key fields are elicited such as resource owner, license, ISO 639 language code(s), and data type. This information can itself be curated and stored in the form of an OLAC archive, to permit search over the union of the existing and newly documented items. Work along these lines has already been initiated by LDC and ELRA (Cieri et al., 2010).

Resource classification. Editors with knowledge of particular language families will categorize documented resources relative to the needs of the project, using controlled vocabularies. This involves examining a resource, determining the granularity and provenance of the segmentation and alignment, checking its ISO 639 classifications, assigning it to a logarithmic size category, documenting its format and layout, collecting sample files, and assigning a priority score.

Acquisition. Where necessary, permission will be sought to lodge the resource in the repository. Funding may be required to buy the rights to the resource from its owner, as compensation for lost revenue from future data sales. Funding may be required to translate the source into a reference language. The repository's ingestion process is followed, and the resource metadata is updated.

Text collection. Languages for which the available resources are inadequate are identified, and the needs are prioritized, based on linguistic and geographical diversity. Sponsorship is sought for collecting bilingual texts in high priority languages. Workflows are developed for languages based on a variety of factors, such as availability of educated people with native-level proficiency in their mother tongue and good knowledge of a reference language, internet access in the language area, availability of expatriate speakers in a first-world context, and so forth. A classification scheme is required to help predict which workflows will be most successful in a given situation.

[3] http://www.language-archives.org/

Audio protocol. The challenge posed by languages with no written literature should not be underestimated. A promising collection method is Basic Oral Language Documentation, which calls for inexpensive voice recorders and netbooks, project-specific software for transcription and sentence-aligned translation, network bandwidth for upload to the repository, and suitable training and support throughout the process.

Corpus readers. Software developers will inspect the file formats and identify high priority formats based on information about resource priorities and sizes. They will code a corpus reader, an open source reference implementation for converting between corpus formats and the storage model presented in section 3.

4.4 Further challenges

There are many additional difficulties that could be listed, though we expect they can be addressed over time, once a sufficient seed corpus is established. Two particular issues deserve further comment, however.

Licenses. Intellectual property issues surrounding linguistic corpora present a complex and evolving landscape (DiPersio, 2010). For users, it would be ideal for all materials to be available under a single license that permits derivative works, commercial use, and redistribution, such as the Creative Commons Attribution License (CC-BY). There would be no confusion about permissible uses of subsets and aggregates of the collected corpora, and it would be easy to view the Universal Corpus as a single corpus. But to attract as many data contributors as possible, we cannot make such a license a condition of contribution.

Instead, we propose to distinguish between: (1) a digital Archive of contributed corpora that are stored in their original format and made available under a range of licenses, offering preservation and dissemination services to the language resources community at large (i.e. the Language Commons); and (2) the Universal Corpus, which is embodied as programmatic access to an evolving subset of materials from the archive under one of a small set of permissive licenses, licenses whose unions and intersections are understood (e.g. CC-BY and its non-commercial counterpart CC-BY-NC). Apart from being a useful service in its own right, the Archive would provide a staging

ground for the Universal Corpus. Archived corpora having restrictive licenses could be evaluated for their potential as contributions to the Corpus, making it possible to prioritize the work of negotiating more liberal licenses.

There are reasons to distinguish Archive and Corpus even beyond the license issues. The Corpus, but not the Archive, is limited to the formats that support automatic cross-linguistic processing. Conversely, since the primary interface to the Corpus is programmatic, it may include materials that are hosted in many different archives; it only needs to know how to access and deliver them to the user. Incidentally, we consider it an implementation issue whether the Corpus is provided as a web service, a download service with user-side software, user-side software with data delivered on physical media, or a cloud application with user programs executed server-side.

Expenses of conversion and editing. We do not trivialize the work involved in converting documents to the formats of section 3, and in manually correcting the results of noisy automatic processes such as optical character recognition. Indeed, the amount of work involved is one motivation for the lengths to which we have gone to keep the data format simple. For example, we have deliberately avoided specifying any particular tokenization scheme. Variation will arise as a consequence, but we believe that it will be no worse than the variability in input that current machine translation training methods routinely deal with, and will not greatly injure the utility of the Corpus. The utter simplicity of the formats also widens the pool of potential volunteers for doing the manual work that is required. By avoiding linguistically delicate annotation, we can take advantage of motivated but untrained volunteers such as students and members of speaker communities.

5 Conclusion

Nearly twenty years ago, the linguistics community received a wake-up call, when Hale et al. (1992) predicted that 90% of the world's linguistic diversity would be lost or moribund by the year 2100, and warned that linguistics might "go down in history as the only science that presided obliviously over the disappearance of 90 per cent of the very field to which it is dedicated." Today, language documentation is a high priority in mainstream linguistics. However, the field of computational linguistics is yet to participate substantially.

The first half century of research in computational linguistics—from circa 1960 up to the present—has touched on less than 1% of the world's languages. For a field which is justly proud of its empirical methods, it is time to apply those methods to the remaining 99% of languages. We will never have the luxury of richly annotated data for these languages, so we are forced to ask ourselves: can we do more with less?

We believe the answer is "yes," and so we challenge the computational linguistics community to adopt a scalable computational approach to the problem. We need leaner methods for building machine translation systems; new algorithms for cross-linguistic bootstrapping via multiple paths; more effective techniques for leveraging human effort in labeling data; scalable ways to get bilingual text for unwritten languages; and large scale social engineering to make it all happen quickly.

To believe we can build this Universal Corpus is certainly audacious, but not to even try is arguably irresponsible. The initial step parallels earlier efforts to create large machine-readable text collections which began in the 1960s and reverberated through each subsequent decade. Collecting bilingual texts is an orthodox activity, and many alternative conceptions of a Human Language Project would likely include this as an early task.

The undertaking ranks with the largest data-collection efforts in science today. It is not achievable without considerable computational sophistication and the full engagement of the field of computational linguistics. Yet we require no fundamentally new technologies. We can build on our strengths in corpus-based methods, linguistic models, human- and machine-supplied annotations, and learning algorithms. By rising to this, the greatest language challenge of our time, we enable multi-lingual technology development at a new scale, and simultaneously lay the foundations for a new science of empirical universal linguistics.

Acknowledgments

We are grateful to Ed Bice, Doug Oard, Gary Simons, participants of the Language Commons working group meeting in Boston, students in the "Digitizing Languages" seminar (University of Michigan), and anonymous reviewers, for feedback on an earlier version of this paper.

References

Steven Abney. 2007. *Semisupervised Learning for Computational Linguistics*. Chapman & Hall/CRC.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media. http://nltk.org/book.

Steven Bird. 2010. A scalable method for preserving oral literature from small languages. In *Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries*, pages 5–14.

Daan Broeder and Peter Wittenburg. 2006. The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies*, 1:119–132.

Christopher Cieri, Khalid Choukri, Nicoletta Calzolari, D. Terence Langendoen, Johannes Leveling, Martha Palmer, Nancy Ide, and James Pustejovsky. 2010. A road map for interoperable language resource metadata. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*.

Gregory R. Crane. 2010. Perseus Digital Library: Research in 2008/09. http://www.perseus.tufts.edu/hopper/research/current. Accessed Feb. 2010.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: an architecture for development of robust HLT applications. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 168–175. Association for Computational Linguistics.

Denise DiPersio. 2010. Implications of a permissions culture on the development and distribution of language resources. In *FLaReNet Forum 2010*. Fostering Language Resources Network. http://www.flarenet.eu/.

Hale, M. Krauss, L. Watahomigie, A. Yamamoto, and C. Craig. 1992. Endangered languages. *Language*, 68(1):1–42.

Nikolaus P. Himmelmann. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel, editors, *Essentials of Language Documentation*, pages 1–30. Mouton de Gruyter.

Human Genome Project. 2007. The science behind the Human Genome Project. http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml. Accessed Dec. 2007.

W. John Hutchins and Harold L. Somers. 1992. *An Introduction to Machine Translation*. Academic Press.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Shankar Kumar, Franz J. Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 42–50, Prague, Czech Republic. Association for Computational Linguistics.

Mike Maxwell and Baden Hughes. 2006. Frontiers in linguistic annotation for lower-density languages. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 29–37, Sydney, Australia, July. Association for Computational Linguistics.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33:129–153.

Kevin Scannell. 2008. The Crúbadán Project: Corpus building for under-resourced languages. In *Cahiers du Cental 5: Proceedings of the 3rd Web as Corpus Workshop*.

Gary Simons and Steven Bird. 2003. The Open Language Archives Community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18:117–128.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21:121–137.

Tamás Váradi, Steven Krauwer, Peter Wittenburg, Martin Wynne, and Kimmo Koskenniemi. 2008. CLARIN: common language resources and technology infrastructure. In *Proceedings of the Sixth International Language Resources and Evaluation Conference*. European Language Resources Association.

Daniel A. Wagner. 1993. *Literacy, Culture, and Development: Becoming Literate in Morocco*. Cambridge University Press.

Glenys Waters. 1998. *Local Literacies: Theory and Practice*. Summer Institute of Linguistics, Dallas.

Douglas H. Whalen and Gary Simons. 2009. Endangered language families. In *Proceedings of the 1st International Conference on Language Documentation and Conservation*. University of Hawaii. http://hdl.handle.net/10125/5017.

Anthony C. Woodbury. 2010. Language documentation. In Peter K. Austin and Julia Sallabank, editors, *The Cambridge Handbook of Endangered Languages*. Cambridge University Press.

Fei Xia and William D. Lewis. 2007. Multilingual structural projection across interlinearized text. In *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics.