Getting the Ball Rolling: Building the Seed Corpus for the Universal Corpus of the World's Language

Abstract

We undertake the grand challenge posed by Abney and Bird (2010) to build the seed text for the Universal Corpus that will include as many of the world's languages as possible, in a consistent structure that permits crosslingual processing and analysis. This paper describes the compilation of a set of texts to form a foundation for the Universal Corpus using an adapted filestore implementation as suggested by Abney and Bird (2011). The seed corpus covers up to 1,275 languages from four different data sources. We urge the computational linguistics community to collaborate in the effort to (i) increase the size by collecting multilingual data from other sources and merging them with the foundation text we have compiled and (ii) improve usability of the corpus by building tools for annotation, search, archiving, presentation, etc.

1 Introduction

The grand aim of linguistics is the construction of a universal theory of human language. To accomplish the goal, the compilation of a Universal Corpus with significant data for a large variety of languages is necessary. Ideally, the Universal Corpus would be a complete digitization of every human language (Abney and Bird, 2010).

Currently multilingual corpora have limited coverage in number of languages and the number of language families the corpora represents; for instance, the OPUS corpus covers over 90 languages (Tiedemann, 2012), i.e. 0.013% of the total number of languages in the world. Even corpora that boast of linguistic diversity lack in language families' coverage; e.g. the linguistically diverse NTU-Multilingual Corpus covers only 7 out of 136 language families (Tan and Bond, 2011).

If we are ever to construct the Universal Corpus, the time is now; over 3,000 out of 6,900 languages are endangered (i.e. have fewer than 1,000 speakers) and 100 out of 420 language families are extinct (i.e. a loss of 24% in linguistic diversity) (Campbell et al. 2013). As computational linguists we do our part in the Universal Corpus initiative by collating, cleaning and processing readily available data into the machine-readable texts.

In the following sections of the paper we describe the (i) ***limitation of existing effort***, (ii) the ***data collection*** process, (iii) ***data format*** of the seed corpus, (iv) ***copyright*** issues, (v) ***compilation challenges***, (vi) ***data delivery and contribution***, (vii) ***future work***, (viii) ***further considerations*** for the Universal Corpus and (ix) ***conclusion***. We envision our seed corpus as a starting point, to be merged with larger collaborative projects when more authoritative data agencies and language archives incorporate the data into a larger Universal Corpus project.

2 Limitations of Existing Effort

The original desideratum of the Universal Corpus proposal is to support automatic processing across a large range of languages (Abney and Bird, 2010).

Traditional language archives that document endangered languages persist and continue to increase in coverage and data size. Also, with the recently initiated *Alliance for Linguistic Diversity* [1] , the infrastructure to upload and download data (audio, video or text) into the cloud is made easier (Endangered Languages, 2012). Though the coverage of languages increases, much data from these archives remains inconsistent (with regards to encodings and file formats). This impedes the development of crosslingual inference methods (for morphology, parsers or even machine translation) across diverse languages; a lack of ***machine readability***.

In the effort to emulate the *Human Genome Project* and to generate more revenue for the language service industry, TAUS established the *Human Language Project*[2] (HLP) as an open platform for sharing and developing language resources and language processing tools with web service APIs. In almost all aspects, it aligns with the original Universal Corpus proposal except its ***availability***; one has to purchase credits to buy the data in the shared pool. The term *Human Language Project* and *Universal Corpus* were used interchangeably in Abney and Bird (2010), to avoid confusion, we refer to our seed corpus as seed corpus / Universal Corpus.

Concerning the matter of developing standard RDF-based models or ISO standard corpus formats (e.g. *Lexical Markup Framework*[3]), we reiterate Abney and Bird's (2010) preference for lightweight formats that are easily machine readable given a simple script / API.

3 Data Collection

We adopt an ***omnivorous*** and ***opportunistic*** approach to the collection of data by (i) leveraging on readily available and open access multilingual repositories and (ii) accepting only data format that are in plaintext or with easily strippable markups. Our approach adheres to the key principles of ***universality*** (i.e. to cover as many languages as possible), ***availability*** (i.e. to ensure data is sharable and adaptable) and ***machine readability*** (i.e. to provide data with consistent formatting and encoding).

Although the ultimate aim is to digitize all possible data sources, regardless of language varieties, availability or machine readability, we subscribe to the general rule of parsimony in our collection for the initial mass of data, viz. '*if it's free and easily usable, use it*'.

Our initial compilation of the Universal Corpus comprises a substantial portion of written data, sparse audio recordings (elicited), various annotations (sentence / phrase / word / morpheme segmentation and their respective glosses). The data sources includes *ODIN: Online Database of Interlinear Glossed Text* (ODIN) (Lewis and Xia, 2010), *Omniglot* phrase lists and translations of the Tower of Babel (Ager, 2012), the *Universal Declaration of Human Rights* (UDHR), and *Wikipedia* dumps.

Data Source	**Medium**	**Data content**
ODIN	parallel IGT texts	sentences, words, morphemes, and glosses
Ominglot	parallel texts + audios	sentences, phrases, words, and glosses
UDHR	parallel texts	sentences, words
Wikipedia	comparable texts	sentences, words

Table 1: Data Source, Medium and Content

Different data sources provide texts at various granularities and hence the content that can be extracted varies. Table 1 presents a summary of the subcorpora with the type of texts they provide (i.e. *medium*) and their contents.

Data Source	**Annotations**
ODIN	morpheme/word alignments
Ominglot	sentence/phrasal alignments, audio to text mappings
UDHR	sentence alignments
Wikipedia	comparable article alignments

Table 2: Possible Annotations from the Data

[1] *www.endangeredlanguages.com*
[2] *www.tausdata.org*
[3] *www.lexicalmarkupframework.org*

Table 2 presents the possible annotations from the different data sources. The ODIN subcorpus contains the IGT texts which provide lexical/morphemic segmentations and alignments between the source word/morpheme to their gloss. The Omniglot data offers parallel text at sentence/phrasal level, occasionally with audio files; without additional processing, the sentences/phrases can be annotated with sentential/phrasal alignments and mappings to their respective audio files.

The UDHR provides parallel texts of a single document while the Wikipedia dumps provide comparable texts that can be treated as comparable documents. It is possible to automatically align the parallel UDHR data and the comparable Wikipedia texts at sentential level with algorithms (e.g. Gale-Church algorithm, 1993) or crosslingual textual similarity algorithms (e.g. Steinberger, 2012; Xia et al. 2011). However it is out of our scope for the seed compilation of the Universal Corpus.

Data Source	**Size**	**Universality (#Languages)**
ODIN	39 MB	1,275
Omniglot (aud)	24 MB	150
Omniglot (txt)	677 KB	178
UDHR	5.2 MB	388
Wikipedia	37 GB	288

Table 3: Data Size and Universality

Data Source	***MB/L***
ODIN	0.03
Omniglot (aud)	0.16
Omniglot (txt)	0.04
UDHR	0.13
Wikipedia	94

Table 4: Data Size and Universality

Traditionally the size of a corpus is presented in number of tokens/words but tokenization is not trivial for all languages. Tokenization for languages with specified word boundary (e.g. whitespaces for most Indo-European languages) is easily achievable but not otherwise. Although high resource languages have specialized tokenizers, low resource languages does not. To proceed with word alignments or further crosslingual analysis, one aspect of the next phase of the Universal Corpus must be to build tokenizers for low resource languages with no specified word boundary.

We attempt a novel way to quantify a corpus when universality is the primary interest, bytes per language (***B/L***). This measure could be view as a rough gauge of how much each data can contribute to the universal theory of language. However this measure is to be taken lightly because the amount of data for higher resource languages within each data source is often disproportional (much higher) as compared to low resource languages. Also, video/audio data would always achieve a higher MB/L. Table 4 shows the *B/L* of the different data sources, ***MB/L*** represents megabytes per language, the MB/L for Wikipedia excludes the 10GB dump for English and it's divided by 287 languages instead of 288.

4 Data Format

Abney and Bird (2010) proposed the *Simple Storage Model*[4] that stores the Universal Corpus texts as plaintext files and the metadata of the each text is stored together in the same file. Their storage model is lightweight and achieves the purpose of minimal processing to retrieve the data. However when the data sources provides different granularity of annotations, there is no one fix way to parse the individual files. Also, the different mediums of the texts (i.e. parallel/comparable) avert cookie-cutter solution as to how the data from different sources should be stored in plaintext.

Alternatively, Abney and Bird (2011) suggested a normalized database implementation to store the Universal Corpus which encompasses the desired range of linguistic objects, alignments and annotations. Further work on the seed corpus will include parallel implementations with both databases and the simple textfiles.

[4] aka the *filestore* model in Abney and Bird (2011)

Presently, we start with the *filestore* implementation and adapted it by incorporating tab delimited texts to allow simpler parsing when annotations/alignments are added to the corpus as compared to using multiple files to store the text alignments. Our adaptation is a matter of preference to keep one text per line and one column per annotation and one file for multiple texts rather than the original *filestore* suggestion of one file per text. The former will produce lesser number of files of larger file sizes, the latter will produce larger number of files of smaller sizes.

Our compilation of the Universal Corpus is made up of multiple directories, each directory represents a data source and it contains

- a compressed ***tarball/zipfile with all the textfiles*** (including transcription of audios)
- a ***textfile that contains the metadata*** of the files in the tarball/zipfile (`.meta`)
- a ***copyrights documentation*** specific to the data source (`.copyrights`)
- a separate **tarball/zipfile with all the audio** files
- if audio exists, ***a textfile contains the map of the audio filenames to their transcriptions*** (`.aud2txt`)
- a python script that serves as **an API to access the files in the subcorpora** (audios and texts)

Segregating data sources in different directories and encapsulating each data source into compressed files, it allows ease of file transfers and users can work with batches of data depending on their data source preference.

```
>>> import universalcorpus as uc
>>> uc.subcorpora()
['odin','omniglot,'udhr','wiki']
>>> uc.odin.texts()
['odin-aae.txt',…,'odin-zul.txt']
>>> uc.wiki.texts()
['wiki001-afk.txt',… ,'wiki012-zul']
```

Figure 1: Example of Universal Corpus API

The python API allows users to retrieve the relevant data without the fuss of writing scripts to parse the corpus format. If and when resource permits, the Universal Corpus' API would be written in user's preferred language; the choice to use python is to allow computational linguists who are familiar with NLTK (written in python) to integrate the corpus into their systems. Without any programming knowledge, users can still use the Universal Corpus as a plaintext files for manual analysis.

The language of a file is denoted by a dash '`-`' and its *ISO 639*[5] codes after the filename; e.g. the file `wiki001-bug.txt` refers to a Wikipedia textfile with Buginese texts. Every textfile is encoded in UCS Transformation Format—8-bit (`utf8`). Abney and Bird (2011) differentiated between sentence-aligned and document-aligned text, our seed compilation saved the texts files with respect to the different content of each data source:

ODIN *(sentence-aligned text)*: each tab-delimited line consists of the source language text, the morphological gloss, the English sentence gloss and citation to the original source

Omniglot *(sentence-aligned text)*: each tab-delimited line consists of the source language text, the English gloss and the filename of the relevant audio file (if exist)

UDHR *(sentence-aligned text):* each line indicates a new paragraph

Wikipedia *(document-aligned text):* each line indicates an article

5 Copyrights

Fair use of data from the seed corpus is crucial to future efforts of new data and the usage of the corpus to produce analysis/tools.

The IGTs from ODIN are subjected to fair used in line with linguistic custom and Section 107 of Copyright Law (Lewis and Xia, 2010; Lewis et al. 2006). The Omniglot data is

[5] *www.sil.org/iso639-3*

copyrighted to Simon Ager but the data is free to use for personal, educational or non-commercial purposes. The *Office of the High Commissioner for Human Rights* owns the copyright to the UDHR data. Wikipedia dumps are copyrighted under the Creative Common Attribution-ShareAlike (CC BY-SA).

The different subcorpora come with assorted copyrights, but they share the common aspect of being free and usable. As such our seed compilation of the Universal Corpus attributes the copyrights of different subcorpora to the original rights of the data source and users of the seed corpus are required to adhere to the rights of the subcorpora[6].

6 Compilation Challenges

Encoding and data cleaning is an issue for any corpus compilation (Escartín, 2013), our seed corpus is no different.

Standardizing the encoding of different data sources is no easy feat. The encoding of the ODIN, Omniglot and Wikipedia data is by default `utf8`. The NLTK's distribution[7] of the UDHR files were originally in Portable Document Format (.pdf) and were converted into plaintext files with various encodings. Even though the encodings of the files are stated in the filenames, several files are not decode-able with their designated encoding. Processing these files is difficult and re-encoding them into `utf8` is excruciating. To correctly decode the files, we had to use `libmagic` (Rullgård and Zoulas, 2009) to detect the right encoding for the files and then discard the files which cannot be decoded by the original designated encoding or the `libmagic`'s detected encoding. Nine files were discarded in the process.

Although the ODIN, Omniglot and UDHR data needs no cleaning, the Wikipedia dump is filled with markup. Providentially, there are various tools to clean Wikipedia markups[8]; our compilation used *wikiextractor* (Medialab, 2013).

7 Data Delivery and Contribution

Compared to larger data providers, we have limited resources to host a data access and contribution platform. Presently, we own a 100GB cloud storage Symform [9] where contributors are able to dump their raw language data (video, audio, textfiles, pdfs, etc.) and collaborators are free to collate unprocessed data and build new subcorpora for the seed corpus.

To encourage wider usage of our seed compilation, a canapé version [10] of the seed corpus without the Wikipedia dump and audio files is hosted on http://qr.net/muk8

Figure 2: QR for Canapé Seed Corpus

8 Future Work

The schedule release of the first edition of the seed corpus is 10 Jan 2014. Beyond the first version, collaborators of seed corpus will be looking into other data sources to increase the coverage and building tokenizers for languages without a tokenizer and without specified word boundary. A secondary task would be to devise a metric to quantify how much each data source is able to contribute in providing sufficient data for a universal theory of language. Additionally, the collaborators can try to implement the database

[6] We are in the midst of asking permission for hosting the data from the owner of each data source.

[7] *http://nltk.googlecode.com/svn/trunk/nltk_data/packages/corpora/udhr.zip*

[8] If one is coding in python, the python package index lists a variety of tools to process Wikipedia data, see *https://pypi.python.org/pypi?:action=search&term=wiki&submit=search*

[9] The cloud is hosted by www.symform.com, please contact the authors for access to the clouds.

[10] The *.meta*, *.aud2text* and APIs have yet been added upon paper submission, proper release scheduled on 10 Jan 2014.

model and add automatic annotations / alignments for the data.

9 Future Considerations

Abney and Bird (2010) highlighted two main issues with regards to building the seed corpus for the Universal Corpus ambition, (i) ***Licenses*** and (ii) ***Expenses/Effort***. Following which, Abney and Bird (2011) include considerations for (iii) ***Versioning*** and (iii) ***Publication*** issues.

Licenses: The licensing issue remains a problem for our seed corpus. Although we attempted to avoid it by (a) selecting only open, free and usable data, and (b) asking for permission to re-distribute, we resort to the not assigning any rights to the seed corpus and attributing each subcorpus to its source. Thus the seed corpus will serve as an archive and facilitate comprehensive citation.

Expenses/Effort: We do not trivialize the work of converting files to text format, standardizing encodings and even cleaning markup. But we show that with a little effort from a small team of volunteers, we are able to build a seed corpus, the challenge remains seeking attention of larger data providers and more volunteers in joining the call-to-arms to build the Universal Corpus.

Versioning: The simple and free cloud storage we are using does not provide versioning services and we see the need to ensure that further work on the seed corpus do not collide especially when annotations are added. As suggested, the notion of '***edition***' is important when increasing data size of the seed corpus or adding annotations for the seed corpora (Abney and Bird, 2011); different *versions* of the corpus can be treated as immutable but when merging into an '*edition'*, old versions of the corpus can be flagged as deprecated and eventually deleted.

Publication: The publication of the different editions remains downloadable online and an editorial process is required to avoid spam (Abney and Bird, 2011). In addition, a work distribution process needs to be included for volunteers to fit the newly contributed data into the Universal Corpus filestore /database.

10 Conclusion

Following the principles set by Abney and Bird (2010, 2011), we have compiled of a set of texts to form the seed corpus for the Universal Corpus that covers up to 1,275 languages from four different data sources. We implemented what used to be theoretical and there is still much work to be done towards the blueprint of designed by Abney and Bird (2011). We highlighted the challenges met during the compilation and set up a corpus delivery system through free cloud storage and encourage data contributors to help increase the size of the seed corpus and also collaborators to turn raw language data of various formats into machine readable formats. Finally we discuss future work to be done on the seed corpus and also further considerations in building the Universal Corpus.

We urge the computational linguistics community to start processing data from linguistic archives and increase the size of the Universal Corpus. And to *get the ball rolling*, we hope that the seed corpus provides a platform for more seed corpora to merge into the first edition of the Universal Corpus.

Acknowledgements

We thank *Anonymous1* and *Anonymous2* for their guidance in the corpus compilation process and feedback to an earlier version of this paper. (*Anonymous** due to blind review)

References

Steven Abney and Steven Bird. 2010. The Human Language Project: Building a Universal Corpus of the World's Languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden.

Steven Abney and Steven Bird. 2010. Towards a data model for the Universal Corpus. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*. Portland, Oregon.

Simon Ager. 2011. Omniglot - writing systems and languages of the world. Retrieved from www.omniglot.com.

Lyle Campbell, Nala Huiying Lee, Eve Okura, Sean Simpson, Kaori Ueki. 2013. New Knowledge: Findings from the Catalogue of Endangered Languages ("ELCat"). In *Proceedings of 3rd International Conference on Language Documentation & Conservation*. Hawaii, USA.

Endangered Languages. 2012. *The Linguist List at Eastern Michigan University and The University of Hawaii at Manoa*. Retrieved from http://www.endangeredlanguages.com

Parra Escartin, Carla. 2013. Encoding a parallel corpus: The TRIS corpus experience. In *The many facets of corpus linguistics in Bergen – in honour of Knut Hofland BeLLS Vol.3, Nr.1* (online version). Bergen: Bergen Language and Linguistics Studies (BeLLS). pp. 61-80.

William Lewis and Fei Xia, 2010. Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages. In *Journal of Literary and Linguistic Computing (LLC)*, 25(3):303-319.

Måns Rullgård and Christos Zoulas. 2009. *Magic number recognition library* - `libmagic(3)` [Software]. Available from http://linux.die.net/man/3/libmagic.

Medialab. 2013. *Wikipedia Extractor* - `WikiExtractor.py` [Software]. Available from http://medialab.di.unipi.it/wiki/Wikipedia_Extractor.

Liling Tan and Francis Bond. 2011. Building and Annotating the Linguistically Diverse NTU-MC (NTU-Multilingual Corpus). In *Proceedings of The 25th Pacific Asia Conference on Language, Information, and Computation*. Singapore, Singapore.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Istanbul, Turkey.