Getting the Ball Rolling: Producing a Foundation Text for a Universal Corpus of the World's Languages

Liling Tan, Guy Emerson, Susanne Fertmann, Alexis Palmer and Michaela Regneri
Universität des Saarlandes
Campus, 66123 Saarbrücken, Germany
liling.tan@uni-saarland.de, emerson@coli.uni-saarland.de,
susfert@coli.uni-saarland.de

Abstract

The study of endangered languages is constrained by a lack of data. Existing corpora collections are limited in the range of languages covered, in standardisation, or in machine-readability. This makes the situation even worse for the computational linguist, especially one who would like to take a cross-linguistic or typological approach. We first survey existing efforts to compile cross-linguistic resources, then describe our own approach and give an example application - language clustering. To build the foundation text for a Universal Corpus, we crawled and cleaned texts from several web sources that contain data from a large number languages, and converted them into a standardised form consistent with the guidelines set out by **?**). The resulting corpus is more easily-accessible and machine-readable than any of the underlying data sources, and represents a significant base corpus for researchers to draw on and add to in the future.

1 Introduction

At the time of writing, 7105 living languages are documented in Ethnologue[1], but **?**) estimated 50% of these are not being learnt by new generations of speakers, and risk extinction by the end of the century. However, only a fraction of the world's languages are well documented, fewer have machine-readable resources, and fewer again have resources with linguistic annotations (**?**) - so the time to work on compiling these resources is now.

Although there have been some previous attempts to produce cross-linguistic resources, there are few which both cover a wide range of languages, and are also machine-readable. We survey existing efforts in section 2, and discuss their limitations in more detail.

Abney and Bird (**?**; **?**) posed the grand challenge of building a Universal Corpus, calling it the Human Language Project. Such a corpus would include all of the world's languages, in a consistent structure, facilitating large-scale cross-linguistic processing. They propose a specific data structure, which we describe and discuss in section 3. We have accepted their challenge, and have begun converting existing resources into a format consistent with their specifications for a universal corpus. We have drawn on four [or five?] web sources, cleaning and standardising them as described in section 4, to produce a seed corpus for the Human Language Project. In sections 6 and 5, we respectively give a summary of the data contained, and discuss copyright and distribution.

[Note: remove this section if we don't do the work!] We believe the resulting corpus is the first of its kind: large enough and consistent enough to allow language processing on a grand scale. In section 7, we give an example application of this corpus: language clustering. We use the frequencies of character n-grams and words to estimate the similarity of two languages. Despite this approach being highly dependent on orthography, we are able to reconstruct substantial parts of several language families, demonstrating the utility of this resource in cross-linguistic research. Finally, we discuss future work in section 8, and conclude in section 9.

2 Related Work

[There is a lot more stuff to be added here...] Currently, multilingual corpora efforts have limited coverage in the number of languages and the number of language families represented. For instance, the OPUS corpus (**?**) covers over 90 lan-

[1] http://www.ethnologue.com/

guages, only 1.27% of the world's languages; The Leipzig Corpora Collection[2] (**?**) contains corpora and dictionaries in 230 languages, 3.24% of the world's languages. Even corpora that boast of linguistic diversity include only a small number of language families, e.g. the linguistically diverse NTU-Multilingual Corpus (**?**) covers only 7 out of 136 language families. Producing a universal corpus requires the linguistic community to merge existing corpora and provide an ubiquitous access interface. To compile the foundation text for the Universal Corpus, we crawled and cleaned web data that contains multilingual texts and merged them with existing corpora collections to form the foundation text for the Universal Corpus.

2.1 Web as Corpus

In recent years, there has been increasing interesting in crawling websites to produce corpora. For example, **?**) introduced the BootCaT toolkit to bootstrap specialised corpora, and **?**) built corpora in various languages by issuing queries for random combinations of frequent words to create a balanced corpus not unlike the British National Corpus. **?**) produced ukWaC, a large-scale English corpus, and (**?**) applied a similar method to produce corpora in German, Italian, and French. To build better quality web corpora, **?**) introduced the notion of content-sensitive boilerplate detection for cleaning data.

However, despite the amount of text they include, such corpora are usually limited in the number of languages represented. Some projects have emerged which strive for language diversity, such as the Leipzig Corpora Collection (LCC) (**?**), and COrpora from the Web (COW) (**?**). The LCC currently offers free download of corpora in 117 languages, and dictionaries in many others, bringing the total number of languages up to 230. COW includes corpora in [XX] languages. These collections are certainly commendable, but they currently still fall short of universality - 230 languages represents only 3.3

?) describes the Crúbadán Project, an attempt to crawl the web for data in a much larger range of languages, including endangered ones. At the time of writing, the number of languages represented has grown to [XXX] However, due to copyright issues, not all of the data is freely available. Motivated by similar copyright concerns, **?**) stressed the notion of building free corpora from the web using documents released under Creative Commons licenses.

Unlike the above efforts, we have not performed seed-query web-searching or web-crawling to achieve a balanced resource. Instead, we have focused on a small number of sources which contain data in a wide variety of languages, as described in section 4. However, these sources represent a variety of data types which a linguist might encounter, ranging from single-language text with significant non-linguistic markup (Wikipedia) to structured data with detailed linguistic annotations (ODIN).

3 Data Structure

?) describe the data structure they envisage for the universal corpus in more detail, distinguishing between **aligned texts** and **analysed texts**: the former consists of multiple parallel texts, aligned at the document, sentence, or word level [add more detail]; the latter contains more detailed annotations including parts of speech, morphological information, and syntactic relations. In our work, we have strived to maintain a data structure consistent with their recommendations for aligned texts.

However, we disagree that analysed texts should be accommodated in this way. If such a universal corpus is to succeed at all, it must enjoy support from a substantial part, if not all, of the linguistics community, and doing this requires theory-neutrality. Although their data structure is fairly lightwight, it is not theory neutral, explicitly encoding support for dependency grammars but not constituency grammars, and positing a specific list of relevant features which should be used to annotate words. Other linguists might wish to use a different set of properties, which could threaten to fragment a universal corpus before it truly gets off the ground. More fundamentally, they assume that the corpus should be segmented into words, but **?**) argues that there is no cross-linguistically valid definition of "word", which would render such an endeavour impossible from the start.

It is not within the scope of this paper to resolve these theoretical concerns. Instead, we suggest that the data structure for a universal corpus should be even more lightweight than Abney and Bird suggest. We agree with their characterisation of aligned texts, but propose an alternative for analysed texts. To motivate the role of par-

[2] corpora.uni-leipzig.de

allel texts in a universal corpus, they propose using translations into a high-resource reference language as a convenient surrogate of meaning. By the same reasoning, we can use the glosses in an IGT to provide a more detailed surrogate of meaning. We propose that, just as we can align texts at the document, sentence, and word levels, we can also align texts at the morpheme level, as recommended by the Leipzig Glossing Rules,[3] and as practised by documentary linguists.

Then, only difference between analysed and aligned texts is that the text is aligned with descriptions in a metalanguage,[4] rather than a natural language. The benefits, however, are that we can use the same data structure to represent both aligned and analysed texts, and the representation is simple enough that linguistic controversies are difficult to raise. We admit that highly detailed annotations, such as trees and dependency graphs, are more awkward to encode, but such resources are unlikely to be available in more than a small handful of languages, and will remain outside the scope of a universal corpus, at least for the foreseeable future. To make it indicate what reference language is being used, we propose prefixing `gloss-` to the language code of the reference language - for example `gloss-eng` and `gloss-eng` if a text has been glossed into English or Spanish, respectively. Using a prefix rather than a suffix makes it clear that this is not a subvariety of the given language.

Finally, it is important to make sure that the data we have compiled will be available to future researchers, regardless of how the surrounding infrastructure changes. **?**) describes a set of best practices for maintaining portability of digital information, outlining ”seven dimensions” along which this can vary. [more detail?] Following this advice, we have ensured that all our data is available as plain text files, with utf-8 encoding, labelled with the relevant ISO 639-3 code. We have written an API to allow access to this data according to the guidelines of Abney and Bird, who remain agnostic as to the specific form of data storage. If, for reasons of space or speed, an alternative format would be preferred, the data would be straightfoward to convert since it can be accessed according these guidelines.

4 Data Sources

Although data size matters in general NLP (**?**) [is this the best reference?], *universality* is the utmost priority for a universal corpus. We chose to focus on the following data sources, because they include a large number of languages, include several parallel texts, and demonstrate a variety of data types which a linguist might encounter (structured, semi-structured, unstructured):

1. The Online Database of Interlinear Text (ODIN)
2. The Omniglot website
3. The Universal Declaration of Human Rights (UDHR)
4. Wikipedia
5. The Leipzig Corpora Collection (LCC) [do we include this?]

In the following subsections, we describe each source, and the processing required to convert the data into a standardised form. Our resulting corpus runs the full gamut of text types outlined by Abney and Bird, ranging from single-language text (Wikipedia and LCC) to parallel text (UDHR and Omniglot) to IGTs (ODIN).

4.1 ODIN

ODIN (The Online Database of Interlinear Text) is a repository of IGTs extracted from scholarly documents (**?**; **?**). Compared to other resources, it is notable for the breadth of languages included, and the level of linguistic annotation; however, the data requires further processing to bring it in line with the proposed format for a universal corpus.

The ODIN data is easily accessible in XML format from the online database[5], where data for each language is saved in a separate XML file and the Interlinear Glossed Texts (IGTs) are encoded in the `<igt><example>...</example></igt>` tags. Each XML file is saved under a filename that matches the respective language code. While cleaning the data, filenames that does not adhere

[3] http://www.eva.mpg.de/lingua/resources/glossing-rules.php

[4] While we would urge researchers to keep such annotations simple, as usually done in IGTs, there is nothing in principle to stop someone from using a complex metalanguage capable of encoding the full set of properties proposed by Abney and Bird, or even more complicated data structures.

[5] http://odin.linguistlist.org/download

to ISO 639-3[6] were excluded in the uniWaC compilation. [Huh? Should this really happen? - the codes are given on their website and look valid to me...]

a. o lesu mai
2sg return here

'*You return here.*'

Figure 1: IGT from ODIN.

```
<igt>
  <example>
    <line>21 a. o lesu mai</line>
    <line>2sg return here</line>
    <line>'You return here.'</line>
  </example>
</igit>
```

Figure 2: Fijian IGT from ODIN.

The IGTs from a web document usually follow the Leipzig Glossing Rules, where the first line is the source language text, the second line contains the word/morphemic equivalence and the third line is an English gloss. From the ODIN XML format, an IGT as of Figure 1 will be represented as in an XML snippet as in Figure 2. Eventually, we need to clean and extract (i) the source text '*o lesu mai*' without the preceding index '*a.*' and (ii) the target language gloss without the quotation marks '*You return here*'.

```
<igt>
  <example>
    <line>(69) na-Na-tmi-kwalca-t
    Yimas (Foley 1991)</line>
    <line>3sgA-1sgO-say-rise-PERF
    </line>
    <line>'She woke me up'
    (by verbal action)</line>
  </example>
</igit>
```

Figure 3: Yimas IGT from ODIN.

The primary problem in extracting the source text is a lack of consistency in the IGTs. In the above examples, the sentence is introduced by a letter or number, which needs to be removed; however, the form of such indexing elements varies. In addition, the source line in Figure 3 includes the language name, and a citation for the example, which introduces noise into the corpus.

Lewis and Xia (2010) noted that it is not uncommon for IGTs found in documents to deviate from the standard convention; when compiling the ODIN data, they attempted a regex-based approached and reported 59% F-score. Similar to their regex approaches for extracting IGT, we cleaned the source line with regexes and kept the ODIN data with two levels of 'cleanliness':

[This hasn't been implemented, but I think we could try this: Split source and gloss lines into tokens according to whitespace. Check the source has more tokens. Count the number of hyphens and equal signs in each token (since they denote morphemes). Match the gloss line with a substring of the source line, according to numbers of hyphens and equal signs. If there's exactly one match, keep it, and discard tokens at the ends of the line. If there's more than one match, look for other punctuation in the initial and final tokens.]

- *Cleaner*: Removed (i) all heading and trailing text embedded in square or rounded brackets and (ii) heading double character token ending with bracket or fullstop.

  (i) `^(?\s?\w{1,5}\s*[):.]\s*`

  (ii) `[\[\(].{1,}[\]\)]`

- *Cleanest*: Only source lines without punctuation.

The original version of the ODIN data contains XML files for 1275 languages, while the cleaner version of ODIN contains IGTs for 1042 languages and the cleanest version contains IGTs for 402 languages. The drop from 1275 to 1042 languages was largely because XXXX XML files from the original ODIN data had used language codes that were not in ISO 639-3 and for XXXX other files, the `<igt>...</igt>` tags were missing.

4.2 Omniglot

The Omniglot website[7] is an online encyclopedia of writing systems and languages. We chose to extract information from pages on *'Useful foreign phrases'* and the *'Tower of Babel'* story, both of which give us a parallel data in a reasonably large number of languages.

[The following discussion is far too technical.]

- `www.omniglot.com/language/phrases/*`

[6] http://www-01.sil.org/iso639-3

[7] http://www.omniglot.com/

- www.omniglot.com/babel/*

The *'Useful foreign phrases'* page contains parallel phrases in embedded within the <th><tr>...</tr><th> tags. The English phrase and foreign language phrase are encoded separately with <td>...</td> tags. Figure 4 shows an instance of an Omniglot *Useful foreign phrase* page. Since the HTML page was designed as WYSIWYG, several pages have non-textual elements within the <tr>...</tr> tags for aesthetics. For example, the non-breaking space, <td>&nbsp;</td> in Fig. 4; the extracted text was properly converted into plaintext format using the HTMLParser[8] module, during which the non textual elements would have been cleaned.

The same process of crawling URLs and cleaning was performed on the *'Tower of Babel'* pages and the only difference was that the texts were embedded in a <ol><li>...</li></ol> tags instead.

```
<th>
  ...
  <tr>
    <td>I don't understand</td>
    <td>No appo cumpresu nutta</td>
    <td>&nbsp;</td>
  </tr>
  <tr>
    <td>I don't understand</td>
    <td>Non d'isco</td>
    <td>&nbsp;</td>
  </tr>
  ...
<th>
```

Figure 4: A *Useful Foreign Phrase* page

One problem with this data is that only the language name is given, not the ISO 639-3 code. To resolve this issue, we automatically converted language names to codes using information from the SIL website.[9]

4.3 Universal Declaration of Human Rights

The Universal Declaration of Human Rights (UDHR) is a document released by the United Nations, which has been translated into a wide variety of languages. Although there was a pre-compiled version of the UDHR data from the Natural Language ToolKit (NLTK) corpora distribution[10], the distribution was laden with encoding problems during their conversion from pdf to plaintext format. Instead, we used the plaintext files available from the Unicode website[11], which are free of encoding issues. The first four lines of each file records metadata, and the rest is the translation of the UDHR. This dataset was extremely clean, and simply required segmentation into sentences.

4.4 Wikipedia

Tapping on the crowd-sourced Wikipedia articles for NLP forms a crucial part of developing data-driven NLP tools and application (citation needed). [It might be popular, but it's not crucial...] To automatically download the Wikipedia dumps, we used the wp-download[12] tool, adapted to an expanded set of languages.

One major issue with using the Wikipedia dumps is the sheer size and extracting text from a glut of Wikipedia markup. To convert compressed Wikipedia dumps to textfiles, we used the WikiExtractor[13] tool. After the conversion into textfiles, we used the following [missing information] regexes to delete residual Wikipedia markup and so-called "magic words"[14].

Wikipedia uses its own set of language codes, most of which are in ISO 639-1 or ISO 639-2/3. We automatically converted all of them into ISO 639-3.

4.5 Leipzig Corpora Collection

Do we want to mention this?

5 Copyright Issues

Creative Commons Licence, etc.

We have made all our resources publically available.[15]

6 Representation and Universality

Number of languages, size of corpus, etc.

[8] http://docs.python.org/2/library/htmlparser.html

[9] http://www-01.sil.org/iso639-3/iso-639-3.tab

[10] http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml

[11] http://unicode.org/udhr/d/

[12] https://github.com/babilen/wp-download

[13] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

[14] http://en.wikipedia.org/wiki/Help:Magic_words

[15] To maintain anonymity, we have removed details, but, if accepted for publication, instructions will be included in the print version of this paper.

7 Language Clustering

This would be an example application, if we have time...

8 Future Work

Specific suggestions on how to work on resources mentioned in section 2...

9 Conclusion

In this paper, we have described the creation of a foundation text for a universal corpus, following the guidelines of Abney and Bird (2010; 2011). To do this, we cleaned and standardised data from several multilingual data sources: ODIN, Omniglot, the UDHR, Wikipedia, and the LCC. The resulting corpus is more easily machine-readable than any of the underlying data sources, and has been stored according to the best practices suggested by **?**). [Give a summary statistic of size.] To demonstrate the utility of this resource, we used the data to perform language clustering, which gave promising results. We believe that this is the first corpus of its kind, which we hope will act as a seed corpus for the Human Language Project.