

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 Abney and Bird (2010). 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 Ethnologue¹, but Krauss (1992) 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 (Maxwell and Hughes, 2006) - 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 (2010; 2011) 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 (Tiedemann, 2012) cov-

¹<http://www.ethnologue.com>

ers over 90 languages, only 1.27% of the world’s languages; The Leipzig Corpora Collection² (Biemann et al., 2007) 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 (Tan and Bond, 2011) 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 interest in crawling websites to produce corpora. For example, Baroni and Bernardini (2004) introduced the BootCaT toolkit to bootstrap specialised corpora, and Sharoff (2006) 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. Ferraresi et al. (2008) produced ukWaC, a large-scale English corpus, and (Baroni et al., 2009) applied a similar method to produce corpora in German, Italian, and French. To build better quality web corpora, Versley and Panchenko (2012) 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) (Biemann et al., 2007), and COpora from the Web (COW) (Schäfer and Bildhauer, 2012). 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

Scannell (2007) 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, Brunello (2009) 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

Abney and Bird (2011) 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 lightweight, 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 Haspelmath (2011) argues that there is no cross-linguistically valid definition of "word", which would render such an endeavour impossible from the start.

²corpora.uni-leipzig.de

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 parallel 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-es` 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. Bird and Simons (2003) 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

³<http://www.eva.mpg.de/lingua/resources/glossing-rules.php>

⁴While 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.

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 straightforward to convert since it can be accessed according to these guidelines.

4 Data Sources

Although data size matters in general NLP (Banko and Brill, 2001) [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 (Lewis, 2006; Lewis and Xia, 2010). 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 database⁵,

⁵<http://odin.linguistlist.org/download>

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 to ISO 639-3⁶ were excluded in the uniWaC compilation. [Huh? Should this really happen? - the codes are given on their website and look valid to me...]

21 a. o lesu mai
2sg return here
'You return here.'

Figure 1: Fijian IGT from ODIN.

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

```

Figure 2: Fijian IGT in ODIN's XML format.

The IGTs from an academic publication 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 in ODIN's XML format.

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 approach 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) $^{\wedge} (? \backslash s ? \backslash w \{ 1, 5 \} \backslash s * [] : .] \backslash s *$
(ii) $[\backslash [\backslash (] . \{ 1, \} [\backslash \backslash)]$

- *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 website⁷ is an online encyclopedia of writing systems and languages. We chose to

⁶<http://www-01.sil.org/iso639-3>

⁷<http://www.omniglot.com>

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 urls for these resources are:

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

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.⁸

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 distribution⁹, 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 website¹⁰, 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

At the time of writing, Wikipedia contains 30.8 million articles in 287 languages, which provides a sizeable amount of clean monolingual text in a fairly wide range of languages. Text dumps are made regularly, and can be downloaded from www.dumps.wikimedia.org. One major issue with using the Wikipedia dump extracting text from a glut of Wikipedia markup. To convert compressed Wikipedia dumps to textfiles, we used the WikiExtractor¹¹ tool. After conversion into textfiles, we used the several regexes to delete residual Wikipedia markup and so-called “magic words”¹².

Wikipedia uses its own set of language codes, most of which are in ISO 639-1 or ISO 639-2/3.

⁸<http://www-01.sil.org/iso639-3/iso-639-3.tab>

⁹http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml

¹⁰<http://unicode.org/udhr/d>

¹¹http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

¹²http://en.wikipedia.org/wiki/Help:Magic_words

We converted all of them into ISO 639-3, following documentation from the Wikimedia Foundation.¹³

5 Copyright Issues

Corpora copyright remains a grey area and a hobgoblin for corpora compilation and distribution (Kilgarriff, 2002)[@Li, please add this reference to the bibliography file]. Often, free access to language data online does not imply the right to download, retain, remix and redistribute.

ODIN refers itself as a database of IGTs that are returned as a result of querying from the source documents. As such, only the links to the documents and the appropriate citations and not the copies of the source documents are maintained on the ODIN site. However, the ODIN developers have left the users with no specific license/copyrights restriction for readaptation or redistribution.

The Omniglot phrases are free to be used for non-commercial purposes as stated in the FAQ page¹⁴ but the rights of the Tower of Babel snippets from different bible sources remains iffy. For example, Omniglot cited www.bibles.org as one of its sources for the Tower of Babel chapter but the bibles.org site stated in fine print that ‘*Except expressly permitted herein, you may not modify, copy, distribute, decompile, disassemble, reverse engineer, create derivative works, or otherwise use or manipulate the Site or any of its content without our prior written consent*’¹⁵. [Ga: actually, about copying information, Omniglot says: ‘*If you would like to copy or re-use any of the language-related articles, please ask the authors for permission. Contact details of the authors can be found at the bottom of the articles in most cases, if they’re not there, please contact me, and I’ll try to find them for you.*’ - I suspect Adger has been given permission for each translation, and I think we would have to ask every single author if we want to re-distribute the data. For the tower of babel translations, this could end up a nightmare...]

Although users are free to use, copy, modify, merge, publish and distribute Unicode Data Files, [Ga: Do we have a reference for this? Particularly distribution] there is no explicit ownership of the UDHR textfiles from the Unicode website. However, the original source of the UDHR docu-

¹³http://meta.wikimedia.org/wiki/Special:language_codes

¹⁴<http://www.omniglot.com/faqs.htm>

¹⁵<http://bibles.org/pages/legal#copyright>

ments from the Office of the High Commissioner for Human Rights clearly states that ‘*None of the materials provided on this web site may be used, reproduced or transmitted, in whole or in part, in any form or by any means, electronic or mechanical, including photocopying, recording or the use of any information storage and retrieval system, except as provided for in the Terms and Conditions of Use of United Nations Web Sites, without permission in writing from the publisher.*’¹⁶. [Ga: Actually, it’s not quite so bad - if we look at the terms and conditions, it says: ‘*The United Nations grants permission to Users to visit the Site and to download and copy the information, documents and materials (collectively, “Materials”) from the Site for the User’s personal, non-commercial use, without any right to resell or redistribute them or to compile or create derivative works therefrom, subject to the terms and conditions outlined below, and also subject to more specific restrictions that may apply to specific Material within this Site.*’¹⁷ Specifically for the UDHR, it says: ‘*If UDHR translations or materials are reproduced, users should make reference to this website as a source by providing a link.*’¹⁸ So, it looks like we’re okay to use the UDHR as we currently are, but if we want to redistribute it, we need to get their permission in writing.]

The Wikipedia texts are licensed under Creative Commons Attribution Share-like¹⁹, which allows redistribution and adaptation. Hence the clean Wikipedia texts we have compiled can be freely redistributed without infringing any copyrights. We have made this resource publically available.²⁰

6 Representation and Universality

Number of languages, size of corpus, etc.

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...

¹⁶<http://www.un.org/en/aboutun/copyright/>

¹⁷<http://www.un.org/en/aboutun/terms/>

¹⁸<http://www.ohchr.org/EN/UDHR/Pages/Introduction.aspx>

¹⁹<http://creativecommons.org/licenses/by-sa/3.0/legalcode>

²⁰To maintain anonymity, we do not give details of how to access the data, but, if accepted for publication, instructions will be included in the print version of this paper.

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 Bird and Simons (2003). [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.

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*, pages 88–97. Association for Computational Linguistics.
- Steven Abney and Steven Bird. 2011. 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*, pages 120–127. Association for Computational Linguistics.
- Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 26–33. Association for Computational Linguistics.
- Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *LREC*.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.
- Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The leipzig corpora collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic 2007*.
- Steven Bird and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language*, pages 557–582.

- Marco Brunello. 2009. The creation of free linguistic corpora from the web. In *Web as Corpus Workshop (WAC5)*, page 9.
- Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google?*, pages 47–54.
- Martin Haspelmath. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1):31–80.
- Michael Krauss. 1992. The world’s languages in crisis. *Language*, 68(1):4–10.
- William D Lewis and Fei Xia. 2010. Developing ODIN: A multilingual repository of annotated language data for hundreds of the world’s languages. *Literary and Linguistic Computing*, 25(3):303–319.
- William D Lewis. 2006. ODIN: A model for adapting and enriching legacy infrastructure. In *e-Science and Grid Computing, 2006. e-Science’06. Second IEEE International Conference on*, pages 137–137. IEEE.
- 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. Association for Computational Linguistics.
- Kevin P Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.
- Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *LREC*, pages 486–493.
- Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. *WaCky*, pages 63–98.
- Liling Tan and Francis Bond. 2011. Building and annotating the linguistically diverse NTU-MC (NTU-Multilingual Corpus). In *PACLIC*, pages 362–371.
- Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*, pages 2214–2218.
- Yannick Versley and Yana Panchenko. 2012. Not just bigger: Towards better-quality web corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 44–52.