SeedLing: Building and using a seed corpus for the Human Language Project

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

Abstract

A broad-coverage corpus such as the Human Language Project envisioned by Abney and Bird (2010) would be a powerful resource for the study of endangered languages. Existing corpora are limited in the range of languages covered, in standardisation, or in machine-readability. In this paper we present SeedLing, a seed corpus for the Human Language Project. We first survey existing efforts to compile cross-linguistic resources, then describe our own approach. To build the foundation text for a universal corpus, we crawl and clean texts from several web sources that contain data from a large number of languages, and convert them into a standardised form consistent with the guidelines of Abney and Bird (2011). The resulting corpus is more easily-accessible and machine-readable than any of the underlying data sources, and, with data from 1451 languages covering 105 language families, represents a significant base corpus for researchers to draw on and add to in the future. To demonstrate the utility of SeedLing for cross-lingual computational research, we use our data in the test application of detecting similar languages.

1 Introduction

At the time of writing, 7105 living languages are documented in Ethnologue,[1] but Simons and Lewis (2011) calculated that 37% of extant languages were at various stages of losing transmisson to new generations. 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.

Several years ago, Abney and Bird (2010; 2011) posed the challenge of building a Universal Corpus, naming it the Human Language Project. Such a corpus would include data from all the world's languages, in a consistent structure, facilitating large-scale cross-linguistic processing. The challenge was issued to the computational linguistics community, from the perspective that the language processing, machine learning, and data manipulation and management tools well-known in computational linguistics must be brought to bear on the problems of documentary linguistics, if we are to make any serious progress toward building such a resource. The Universal Corpus as envisioned would facilitate broadly cross-lingual natural language processing (NLP), in particular driving innovation in research addressing NLP for low-resource languages, which in turn supports the language documentation process.

We have accepted this challenge and have begun converting existing resources into a format consistent with Abney and Bird's specifications. We aim for a collection of resources that includes data: (a) from as many languages as possible, and (b) in a format both in accordance with best practice archiving recommendations and also readily accessible for computational methods. Of course, there are many relevant efforts toward producing cross-linguistic resources, which we survey in section 2. To the best of our knowledge, though, no existing effort meets these two desiderata to the extent of our corpus, which we name SeedLing: a seed corpus for the Human Language Project.

To produce SeedLing, we have drawn on four web sources, described in section 3.2. To bring the four resources into a single common format and data structure (section 3.1), each required different degrees and types of cleaning and standardisation. We describe the steps required in section

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

4, presenting each resource as a separate mini-case study. We hope that the lessons we learned in assembling our seed corpus can guide future resource conversion efforts. To that end, many of the resources described in section 2 are candidates for inclusion in the next stage of building a Universal Corpus.

We believe the resulting corpus, which at present covers 1451 languages from 105 language families, is the first of its kind: large enough and consistent enough to allow broadly multilingual language processing. To test this claim, we use SeedLing in a sample application (section 5): the task of language clustering. With no additional pre-processing, we extract surface-level features (frequencies of character n-grams and words) to estimate the similarity of two languages. Unlike most previous approaches to the task, we make no use of resources curated for linguistic typology (e.g. values of typological features as in WALS (Dryer and Haspelmath, 2013), Swadesh word lists). Despite our approach being highly dependent on orthography, our clustering performance matches the results obtained by Georgi et al. (2010) using typolological features, which demonstrates SeedLing's utility in cross-linguistic research.

2 Related Work

In this section, we review existing efforts to compile multilingual machine-readable resources. Although some commercial resources are available, we restrict attention to freely accessible data.[2]

Traditional archives. Many archives exist to store the wealth of traditional resources produced by the documentary linguistics community. Such documents are increasingly being digitised, or produced in a digital form, and there are a number of archives which now offer free online access.

Some archives aim for a universal scope, such as The Language Archive (maintained by the Max Planck Institute of Psycholinguistics), Collection Pangloss (maintained by LACITO), and The Endangered Languages Archive (maintained by SOAS). Most archives are regional, including AILLA, ANLA, PARADISEC, and many others.

However, there are two main problems common to all of the above data sources. Firstly, the data is not always machine readable. Even where the data is available digitally, these often take the form of scanned images or audio files. While both can provide invaluable information, they are extremely difficult to process with a computer, requiring an impractical level of image or video pre-processing before linguistic analysis can begin. Even textual data, which avoids these issues, may not be available in a machine-readable form, being stored as pdfs or other opaque formats. Secondly, when data is machine readable, the format can vary wildly. This makes automated processing difficult, especially if one is not aware of the details of each project. Even when metadata standards and encodings agree, there can be idiosyncractic markup or non-linguistic information, such as labels for speakers in the transcript of a conversation.

We can see that there is still much work to be done by individual researchers in digitising and standardising linguistic data, and it is outside of the scope of this paper to attempt this for the above archives. Guidelines for producing new materials are available from the E-MELD project (Electronic Metastructure for Endangered Languages Data), which specifically aimed to deal with the expanding number of standards for linguistic data. It gives best practice recommendations, illustrated with eleven case studies, and provides input tools which link to the GOLD ontology language, and the OLAC metadata set. Further recommendations are given by Bird and Simons (2003), who describe seven dimensions along which the portability of linguistic data can vary. Various tools are available from The Language Archive at the Max Planck Institute for Psycholinguistics.

Many of these archives are part of the Open Language Archive Community (OLAC), a subcommunity of the Open Archives Initiative. OLAC maintains a metadata standard, based on the 15-element Dublin Core, which allows a user to search through all participating archives in a unified fashion. However, centralising access to disparate resources, while of course extremely helpful, does not solve the problem of inconsistent standards. Indeed, it can be difficult even to answer simple questions like "how many languages are represented?"

In short, while traditional archives are invaluable for many purposes, for large-scale machine processing, they leave much to be desired.

[2] All figures given below were correct at the time of writing, but it must be borne in mind that most of these resources are constantly growing.

Generic corpus collections. Some corpus collections exist which do not focus on endangered languages, but which nonetheless cover an increasing number of languages.

MetaShare (Multilingual Europe Technology Alliance) provides data in a little over 100 languages. While language codes are used, they have not been standardised, so that multiple codes are used for the same language. Linguistic Data Consortium (LDC) and the European Language Resources Association (ELRA) both offer data in multiple languages. However, while large in size, they cover only a limited number of languages. Furthermore, the corpora they contain are stored separately, making it difficult to access data according to language.

Parallel corpora. The Machine Translation community has assembled a number of parallel corpora, which are crucial for statistical machine translation. The OPUS corpus (Tiedemann, 2012) subsumes a number of other well-known parallel corpora, such as Europarl, and covers documents from 350 languages, with various language pairs.

Web corpora. There has been increasing interest in deriving corpora from the web, due to the promise of large amounts of data. The majority of web corpora are however aimed at either one or a small number of languages, which is perhaps to be expected, given that the majority of online text is written in a handful of high-resource languages. Nonetheless, there have been a few efforts to apply the same methods to a wider range of languages.

HC Corpora currently provides download of corpora in 68 different language varieties, which vary in size from 2M to 150M words. The corpora are thus of a respectable size, but only 1% of the world's languages are represented. A further difficulty is that languages are named, without the corresponding ISO language codes.

The Leipzig Corpora Collection (LCC)[3] (Biemann et al., 2007) provides download of corpora in 117 languages, and dictionaries in a number of others, bringing the total number of represented languages up to 230. The corpora are large, readily available, in plain-text, and labelled with ISO language codes.

The Crúbadán Project aims to crawl the web for text in low-resource languages, and data is currently available for 1872 languages. This represents a significant portion of the world's languages; unfortunately, due to copyright restrictions, only lists of n-grams and their frequencies are publically available, not the texts themselves. While the breadth of languages covered makes this a useful resource for cross-linguistic research, the lack of actual texts means that only a limited range of applications are possible with this data.

Cross-linguistic projects. Responding to the call to document and preserve the world's languages, highly cross-linguistic projects have sprung up, striving towards the aim of universality. Of particular note are the Endangered Languages Project, and the Rosetta Project. These projects are to be praised for their commitment to universality, but in their current forms it is difficult to use their data to perform large-scale NLP.

3 The Data

3.1 Universal Corpus and Data Structure

Building on their previous paper, Abney and Bird (2011) describe the data structure they envisage for the universal corpus in more detail, and we aim to adopt this structure where possible. Two types of text are distinguished:

Aligned texts consist of parallel documents, aligned at the document, sentence, or word level. Note that monolingual documents are viewed as aligned texts only tied to a single language.

Analysed texts, in addition to the raw text, contain more detailed annotations including parts of speech, morphological information, and syntactic relations. This is stored as a table, where rows represent words, and columns represent: document ID, language code, sentence ID, word ID, wordform, lemma, morphological information, part of speech, gloss, head/governor, and relation/role.

Out of our data sources, three can be straightforwardly represented in their aligned text structure. However, ODIN contains richer annotations, which are in fact difficult to fit into their proposal, and which we discuss in section 3.2 below.

3.2 Data Sources

Although data size matters in general NLP, *universality* is the top priority for a universal corpus. We 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): the Online

[3] http://corpora.uni-leipzig.de

	Langs.	Families	Tokens	Size
ODIN	1,270	100	351,161	39 MB
Omniglot	129	20	31,318	677 KB
UDHR	352	46	640,588	5.2 MB
Wikipedia	271	21		37 GB
Combined	1,451	105		

Table 1: Corpus Coverage

Database of Interlinear Text (ODIN), the Omniglot website, the Universal Declaration of Human Rights (UHDR), and Wikipedia.

Our resulting corpus runs the full gamut of text types outlined by Abney and Bird, ranging from single-language text (Wikipedia) to parallel text (UDHR and Omniglot) to IGTs (ODIN). Table 1 gives some coverage statistics, and we describe each source in the following subsections. For 332 languages, the corpus contains data from more than one source.

Universal Declaration of Human Rights. The Universal Declaration of Human Rights (UDHR) is a document released by the United Nations in 1948, and represents the first global expression of human rights. It consists of 30 articles, amounting to about four pages of text. This is a useful document for NLP, since it has been translated into a wide variety of languages, providing a highly parallel text.

Wikipedia. Wikipedia is a collaboratively-edited encyclopedia, appealing to use for NLP because of its large size and easy availability. At the time of writing, it contained 30.8 million articles in 286 languages, which provides a sizeable amount of monolingual text in a fairly wide range of languages. Text dumps are made regularly available, and can be downloaded from `http://dumps.wikimedia.org`.

Omniglot. The Omniglot website[4] is an online encyclopedia of writing systems and languages. We extract information from pages on *'Useful foreign phrases'* and the *'Tower of Babel'* story, both of which give us parallel data in a reasonably large number of languages.

ODIN. ODIN (The Online Database of Interlinear Text) is a repository of interlinear glossed texts (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. An IGT canonically consists of three lines: (i) the source, a sentence in a target language, (ii) the gloss, an analysis of each source element, and (iii) the translation, done at the sentence level. The gloss line can additionally include a number of linguistic terms, which means that the gloss is written in metalanguage rather than natural language. In ODIN, translations are into English, and glosses are written in an English-based metalanguage. An accepted set of guidelines are given by the Leipzig Glossing Rules,[5] where morphemes within words are separated by hyphens (or equal signs, for clitics), and the same number of hyphens should appear in each word of the source and gloss.

The data from ODIN poses the first obstacle to straightforwardly adopting Abney and Bird's data structure. The proposed data structure for the universal corpus is aligned at the word level, and includes a specific list of relevant features which should be used to annotate words. When we try to adapt IGTs into this format, we run into certain problems. Firstly, there is the problem that the most fundamental unit of analysis according to the Leipzig Glossing Rules is the morpheme, not the word. Ideally, we should encode this information explicitly in a universal corpus, assigning a unique identifier to each morpheme (instead of, or in addition to each word). Indeed, Haspelmath (2011) argues that there is no cross-linguistically valid definition of *word*, which undermines the central position of words in the proposed data structure.

Secondly, it is unclear how to represent the gloss. Since the gloss line is not written in a natural language, we cannot treat it as a simple translation. However, it is not straightforward to incorporate it into the proposed structure for analysed texts, either. One possible resolution is to move all elements of the gloss written in capital letters to the MORPH field (as functional elements are usually annotated in this way), and all remaining elements to the GLOSS field. However, this loses information, since we no longer know which morpheme has which meaning. To keep all information encoded in the IGT, we need to modify Abney and Bird (2011)'s proposal.

The simplest solution we can see is to allow morphemes to be a level of structure in the universal corpus, just as documents, sentences, and

[4]`http://www.omniglot.com`

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

		Total #	Omniglot	Wikipedia	UDHR	ODIN	Combined
0	International	6	100.0%	100.0%	100.0%	100.0%	100.0%
1	National	95	53.7%	73.7%	83.2%	83.2%	91.6%
2	Provincial	70	31.4%	48.6%	57.1%	71.4%	80.0%
3	Wider Comm.	166	3.6%	12.0%	20.5%	38.0%	44.6%
4	Educational	345	3.2%	8.1%	15.1%	33.0%	38.0%
5	Developing	1534	0.5%	2.2%	4.6%	23.2%	26.1%
6a	Vigorous	2502	0.0%	0.2%	0.4%	6.4%	6.7%
6b	Threatened	1025	0.6%	1.7%	2.9%	15.0%	17.1%
7	Shifting	456	0.2%	0.9%	1.8%	14.5%	16.0%
8a	Moribund	286	0.3%	1.0%	1.0%	22.4%	23.1%
8b	Nearly Extinct	432	0.2%	0.2%	0.9%	15.3%	16.0%
9	Dormant	188	0.5%	1.1%	0.0%	10.6%	11.2%

Figure 1: Heatmap of languages in SeedLing according to endangerment status

words already are. The overall architecture remains unchanged. We must then decide how to represent the glosses.

Even though glosses in ODIN are based on English, having been extracted from English-language documents, this is not true of IGTs in general. For example, it is common for documentary linguists working on indigenous languages of the Americas to provide glosses and translations based on Spanish. For this reason, we believe it would be wise to specify the language used to produce the gloss. Since it is not quite the language itself, but a metalanguage, one solution would be to use new language codes that make it clear both that a metalanguage is being used, and also what natural language it is based on. The five-letter code `gloss` cannot be confused with any code in any version of ISO 639 (with codes of length two to four). Following the convention that subvarieties of a language are indicated with suffixes, we can append the code of the natural language. For example, glosses into English and Spanish-based metalanguages would be given the codes `gloss-eng` and `gloss-spa`, respectively.

One benefit of this approach is that glossed texts are treated in exactly the same way as parallel texts. There is a unique identifier for each morpheme, and glosses are stored under this identifier and the corresponding gloss code. Furthermore, to motivate the important place of parallel texts in a universal corpus, Abney and Bird view translations into a high-resource reference language as a convenient surrogate of meaning. By the same reasoning, we can use glosses to provide a more detailed surrogate of meaning, only written in a metalanguage instead of a natural one.

3.3 Representation and Universality

According to Ethnologue, there are 7105 living languages, and 147 living language families. Across all our data sources, we manage to cover 1451 languages in 105 families, which represents 19.0% of the world's languages. To get a better idea of the kinds of languages represented, we give a breakdown according to their EGIDS scores (Expanded Graded Intergenerational Disruption Scale) (Lewis and Simons, 2010) in Figure 1. The values in each cell have been colored according to proportion of languages represented, with green indicating good coverage and red poor. It's interesting to note that vigorous languages (6a) are poorly represented across all data sources, and worse than more endangered categories. In terms of language documentation, vigorous languages are less urgent goals than those in categories 6b and up, but this highlights an unexpected gap in linguistic resources.

4 Data Clean-Up, Consistency, and Standardisation

Consistency in data structures and formatting is essential to facilitate use of data in computational linguistics research (Palmer et al., 2010). In the following subsections, we describe the processing required to convert the data into a standardised form. We then discuss standardisation of language codes and file formats.

4.1 Case Studies

UDHR. We used the plain-text UDHR files available from the Unicode website[6] which uses UTF-8 encoding for all languages. The first four lines of each file record metadata, and the rest is the translation of the UDHR. This dataset is extremely clean, and simply required segmentation into sentences.

Wikipedia. One major issue with using the Wikipedia dump is the problem of separating text from abundant source-specific markup. To convert compressed Wikipedia dumps to textfiles, we used the WikiExtractor[7] tool. After conversion into textfiles, we used several regular expressions to delete residual Wikipedia markup and so-called "*magic words*".[8]

Omniglot. The main issue with extracting the Omniglot data is that the pages are designed to be human-readable, not machine-readable. Cleaning this data required parsing the HTML source, and extracting the relevant content, which required different code for the two types of page we considered ('*Useful foreign phrases*' and '*Tower of Babel*'). Even after automatic extraction, some noise in the data remained, such as explanatory notes given in parentheses, which are written in English and not the target language. Even though the total amount of data here is small compared to our other sources, the amount of effort required to process it was not, because of these idiosyncracies. We expect that researchers seeking to convert data from human-readable to machine-readable formats will encounter similar problems, but unfortunately there is unlikely to be a one-size-fits-all solution to this problem.

ODIN. The ODIN data is easily accessible in XML format from the online database[9]. Data for each language is saved in a separate XML file and the IGTs are encoded in tags of the form `<igt><example>...</example></igt>`. For example, the IGT in Figure 2 is represented by the XML snippet in Figure 3.

The primary problem in extracting the data 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 4 includes two types of metadata: the language name, and a citation, both of which introduce noise. Finally, extraneous punctuation such as the quotation marks in the translation line need to be removed. We used regular expressions for cleaning lines within the IGTs.

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

Figure 2: Fijian IGT from ODIN

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

Figure 3: Fijian IGT in ODIN's XML format

4.2 Language Codes

As Xia et al. (2010) explain, language names do not always suffice to identify languages, since many names are ambiguous. For this reason, sets of language codes exist to more accurately identify languages. We use ISO 639-3[10] as our standard set of codes, since it aims for universal coverage, and has widespread acceptance in the community. The data from ODIN and the UDHR already used this standard.

To facilitate the standardization of language codes, we have written a python API that can be used to query information about a language or a code, fetching up-to-date information from SIL International (which maintains the ISO 639-3 code set), as well as from Ethnologue.

Wikipedia uses its own set of language codes, most of which are in ISO 639-1 or ISO 639-3. The older ISO 639-1 codes are easy to recognise, being two letters long instead of three, and can be straightforwardly converted. However, a small number of Wikipedia codes are not ISO codes at all - we converted these to ISO 639-3, following

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

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

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

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

[10] http://www-01.sil.org/iso639-3/default.asp

```
<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 4: Yimas IGT in ODIN's XML format

documentation from the Wikimedia Foundation.[11]

Omniglot does not give codes at all, but only the language name. To resolve this issue, we automatically converted language names to codes using information from the SIL website.

Some languages have more than one orthography. For example, Mandarin Chinese is written with either traditional or simplified characters; Serbian is written with either the Cyrillic or the Roman alphabet. For cross-linguistic NLP, it could be helpful to have standard codes to identify orthographies, but at present none exist.

4.3 File Formats

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) describe a set of best practices for maintaining portability of digital information, outlining seven dimensions along which this can vary. Following this advice, we have ensured that all our data is available as plaintext files, with UTF-8 encoding, labelled with the relevant ISO 639-3 code. Metadata is stored separately. This allows users to easily process the data using the programming language or software of their choice.

To allow access to the data following Abney and Bird's guidelines, as discussed in section 3, we have written an API, which we distribute along with the data. Abney and Bird remain agnostic to the specific file format used, but if an alternative format would be preferred, the data would be straightfoward to convert since it can be accessed according to these guidelines. As examples of functionality, our API allows a user to fetch all sentences in a given language, all sentences from a given source.

[11]http://meta.wikimedia.org/wiki/Special_language_codes

5 Detecting Similar Languages

To exemplify the use of SeedLing for computational research on low-resource languages, we experiment with automatic detection of similar languages. When working on endangered languages, documentary and computational linguists alike face a lack of resources. It is often helpful to exploit lexical, syntactic or morphological knowledge of related languages. For example, similar high-resource languages can be used in bootstrapping approaches, such as described by Yarowsky and Ngai (2001) or Xia and Lewis (2007).

Language classification can be carried out in various ways. Two common approaches are genealogical classification, mapping languages onto family trees according to their historical relatedness (Swadesh, 1952; Starostin, 2010); and typological classification, grouping languages according to linguistic features (Georgi et al., 2010; Daumé III, 2009). Both of these approaches require linguistic analysis. By contrast, we use surface features (character n-gram and word unigram frequencies) extracted from SeedLing, and apply an off-the-shelf hierarchical clustering algorithm.[12] Specifically, each language is represented as a vector of frequencies of character bigrams, character trigrams, and word unigrams. Each of these three components is normalised to unit length. Data was taken from ODIN, Omniglot, and the UDHR.

Experimental Setup. We first perform hierarchical clustering, which produces a tree structure: each leaf represents a language, and each node a cluster. We use linkage methods, which recursively build the tree starting from the leaves. Initially, each language is in a separate cluster, then we iteratively find the closest two clusters and merge them. Each time we do this, we take the two corresponding subtrees, and introduce a new node to join them.

We define the distance between two clusters by considering all possible pairs of languages, with one from each cluster, and taking the largest distance. We experimented with other ways to define the distance between clusters, but results were poor and we omit results for brevity.

To ease evaluation, we produce a partitional clustering, by stopping when we reach a certain number of clusters, set in advance.

[12]http://www.scipy.org

	Precision	Recall	F-score
Base. 1: random	0.184	0.092	0.068
Base. 2: together	0.061	1.000	0.112
Base. 3: separate	1.000	0.086	0.122
SeedLing	0.255	0.205	0.150

Table 2: Clustering compared with baselines

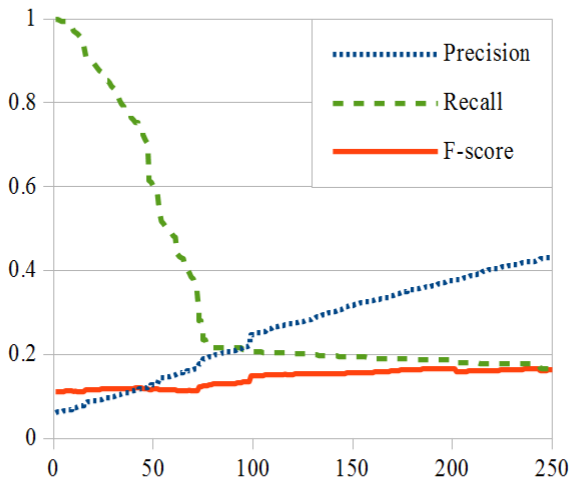

Figure 5: Performance against number of clusters

Evaluation. We compare our clustering to the language families in Ethnologue. However, there are many ways to evaluate clustering quality. Amigó et al. (2009) propose a set of criteria which a clustering evaluation metric should satisfy, and demonstrate that most popular metrics fail to satisfy at least one of these criteria. However, they prove that all criteria are satisfied by the BCubed metric, which we therefore adopt. To calculate the BCubed score, we take the induced cluster and gold standard class for each language, and calculate the F-score of the cluster compared to the class. These F-scores are then averaged across all languages.

In Table 2, we set the number of clusters to be 105, the number of language families in our data, and compare this with three baselines: a random baseline (averaged over 20 runs); putting all languages in a single cluster; and putting all languages in separate clusters. Our clustering outperforms all baselines. It is worth noting that precision is higher than recall, which is perhaps expected, given that related languages using wildly differing orthographies will appear distinct.

To allow a closer comparison with Georgi et al. (2010), we calculate pairwise scores - i.e. considering if pairs of languages are in the same cluster or the same class. For 105 clusters, we achieve a pairwise f-score of 0.147, while Georgi et al. report 0.140. The figures are not quite comparable since we are evaluating over a different set of languages; nonetheless, we only use surface features, while Georgi et al. use typological features from WALS. This suggests the possibility for cross-linguistic research to be conducted based on shallow features.

In Figure 5, we vary the number of clusters. The highest f-score is obtained for 199 clusters. There is a notable jump in performance between 98 and 99, just before the true number of families, 105.

Interpreting the clusters directly is difficult, because they are noisy. However, the distribution of cluster sizes mirrors the true distribution - for 105 clusters, we have 48 clusters of size 1 or 2, with the largest cluster of size 130; while in fact there are 51 families with only 1 or 2 languages in the data, with the largest of size 150.

6 Conclusion and Outlook

In this paper, we have described the creation of SeedLing, 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. 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). At present, SeedLing has data from 19% of the world's living languages, covering 72% of language families. We believe that a corpus with such diversity of languages, uniformity of format, cleanliness of data, and ease of access provides an excellent seed for a universal corpus. It is our hope that taking steps toward creating this resource will spur both further data contributions and interesting computational research with cross-linguistic or typological perspectives; we have here demonstrated SeedLing's utility for such research by using the data to perform language clustering, with promising results.

SeedLing (data, API and documentation) is currently available via a GitHub repository.[13] We have yet to fully address questions of long-term access, and we welcome ideas or collaborations along these lines.

[13] https://github.com/alvations/SeedLing

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.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486.

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.

Hal Daumé III. 2009. Non-parametric bayesian areal linguistics. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 593–601. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 385–393. Association for Computational Linguistics.

Martin Haspelmath. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1):31–80.

M Paul Lewis and Gary F Simons. 2010. Assessing endangerment: expanding fishman's GIDS. *Revue roumaine de linguistique*, 2:103–119.

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.

Alexis Palmer, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell, and Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation. *Linguistic Issues in Language Technology*, 3.

Gary F Simons and M Paul Lewis. 2011. The world's languages in crisis: A 20-year update. In *26th Linguistic Symposium: Language Death, Endangerment, Documentation, and Revitalization. University of Wisconsin, Milwaukee*, pages 20–22.

George Starostin. 2010. Preliminary lexicostatistics as a basis for language classification: a new approach. *Journal of Language Relationship*, 3:79–117.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north american indians and eskimos. *Proceedings of the American philosophical society*, pages 452–463.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *LREC*, pages 2214–2218.

Fei Xia and William D Lewis. 2007. Multilingual structural projection across interlinear text. In *HLT-NAACL*, pages 452–459.

Fei Xia, Carrie Lewis, and William D Lewis. 2010. The problems of language identification within hugely multilingual data sets. In *LREC*.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of NAACL-2001*, pages 200–207.