Walking the Talk: Building the Universal Corpus of the World’s Language

|  |  |  |
| --- | --- | --- |
| Susanne Fertmann  Universität des Saarlandes,  Saarbrücken, Germany  s9sufert@stud.uni-saarland.de | Guy Emerson  University of Cambridge,  Cambridge, UK  g.e.t.emerson@googlemail.com | Liling Tan  Universität des Saarlandes,  Saarbrücken, Germany  liling.tan@uni-saarland.de |

|  |  |
| --- | --- |
|  |  |

Abstract

We undertake the grand challenge posed by Abney and Bird (2010) to do our part in building a corpus that will include all of the world's languages, in a consistent structure that permits crosslingual processing and analysis. This paper describes the infrastructure setup and foundation text compilation for the Universal Corpus grail. 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) improving 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).

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. has fewer than 1,000 speakers) and 100 out of 420 language families are extinct (i.e. a loss of 24% in linguistic diversity) (Lyle 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 bitext storage model as described by Abney and Bird (2010).

In the following sections of the paper we describe our attempt to compile a foundation text corpus for the Universal Corpus. We envision our compilation to be merged with larger collaborative projects when more authoritative data agencies and language archives incorporate our data into a larger Universal Corpus project.

1. 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*** and ***machine readability***.

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 compilation of initial mass of data, viz. ‘*if it’s free and easily usable, use it*’.

The sources for our compilation of the Universal Corpus include the *Universal Declaration of Human Rights* (UDHR), *Wikipedia*, *Omniglot*’s phrase lists and translations of the Tower of Babel.

**Acknowledgments**

Do not number the acknowledgment section. Do

not include this section when submitting your paper for review.

References

Alfred. V. Aho and Jeffrey D. Ullman. 1972. The Theory of Parsing, Translation and Compiling, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. Publications Manual. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. Computing Reviews, 24(11):503-512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J.Stockmeyer. 1981. Alternation. Journal of the Association for Computing Machinery, 28(1):114-133.

Dan Gusfield. 1997. Algorithms on Strings, Trees and Sequences. Cambridge University Press, Cambridge, UK.