

Artificial Intelligence and Ethics (CT5142) Group Project

**Resource Type:** Blog

**Topic:** Artificial agents and XR - How should AI and artificial agents be used responsibly in society?

**Title:** Investigating the Ethics of Creating Corpora for Under-resourced Languages

**Group Members:** Jiarong Li, Yamini Girkar, William Ikenna-Nwosu

**Student IDs:** 20230033, , 22222642

**College:** College of Science and Engineering

**School:** School of Computer Science

**Course:** M.Sc. Artificial Intelligence

**Course Code:** 1MAI1

Part A - Final Resource

Part A - Reflection

### **Describe your target audience**

### **Describe where/how you would disseminate it to this target audience**

### **Explain why the chosen format is suitable to communicate with your target audience**

### **Explain what makes the chosen topic and specific example(s) used suitable for this particular audience; how do they connect to the audience’s existing knowledge, how do they help expand on this knowledge**

### **Explain the development process for the resource (e.g. how did you decide on the topic, how did your topic definition and approach develop over time, what challenges did you encounter, how did you resolve these challenges)**

### **What do you see as remaining limitations/weak points of your product? What might improve it further?**

### 

Part B - Literature Review

# **Investigating the Ethics of Creating Corpora for Under-resourced Languages: A Review of the Literature**

## Introduction

The term “under-resourced languages” introduced by Krauwer (2003) \ref{Krauwer, S., 2003} and Berment (2004) \ref{Berment, V., 2004} refers to a language with some of (if not all) the following aspects: lack of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise, lack of electronic resources for speech and language processing, such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, pronunciation dictionaries, vocabulary lists, etc.

## Challenges

## Case Studies

## Methodologies and Technologies

There are more than 6900 languages in the world but only a small fraction of them offer the resources to develop the applications of Human Language techniques because of the economics and political scene \ref{Laurent Besacier}. Lack of under-resourced language resources requires us to apply innovative data collection technologies \ ref{Gelas, H. 2011} or models \ref{Schultz, 2006}. Moreover, there is a gap between the native speakers and the technique experts and developers as it is not possible to find a native speaker with the technique skills to develop an under-resourced language system. Therefore, the general methodologies applied in this case is that we borrow the resource from the similar languages with the rich resource and port, adopt, create written and spoken resources for under-resourced language.

## Conclusions and need for Further Research

References

Krauwer, S., 2003. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In: Proceedings of the 2003 International Workshop Speech and Computer SPECOM-2003, Moscow, Russia, pp. 8–15.

Berment, V., 2004. Me ́thodes pour informatiser des langues et des groupes de langues peu dote ́es. Ph.D. Thesis, J. Fourier University – Grenoble I, May 2004.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Automatic speech recognition for under-resourced languages: A survey. Speech communication, 56:85–100, 2014.

Gelas, H., Teferra Abate, S., Besacier, L., Pellegrino, F., 2011. Quality assessment of crowdsourcing transcriptions for African languages. In: Interspeech 2011 Florence, Italy, 28–31 August 2011.

Schultz, T., 2006. Multilingual speech processing. In: Tanja Schultz, Katrin Kirchhoff (Eds.), Elsevier, Academic Press, ISBN 13: 978-0-12- 088501-5, 2006.