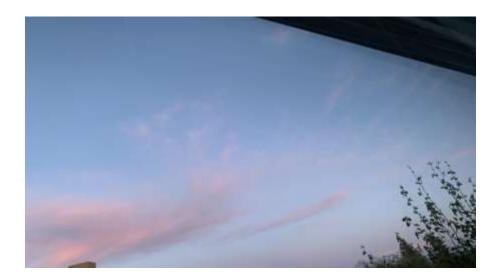
Language in India www.languageinindia.com ISSN 1930-2940 Vol. 20:8 August 2020

Generating a Parallel Corpus Stream for Odia: Mining Parallel Corpus from Odia Twitter

Anjan Kumar Panda, MSC IT, KSOU Mysore

Internet Application Specialist, Technology Manager Life Member, OSA. The Odisha Society of the Americas 5050, Hacienda Drive, Apt 2232, Dublin, CA, 94568 <u>panda.anjankumar@gmail.com</u> Contact: 1- 845-535-0961

Dr Arun Kumar Malik, PhD, Assistant Professor of Political Science Gujarat National Law University, Gandhinagar <u>amalik@gnlu.ac.in</u> Contact No. 8128650850



Note: This research article is part of a sequence of papers to enable the researchers in the field of computational linguistics in Odia.

Language in India www.languageinindia.com ISSN 1930-2940 20:8 August 2020 Anjan Kumar Panda, MSC IT, KSOU Mysore and Dr Arun Kumar Malik, PhD Generating a Parallel Corpus Stream for Odia: Mining Parallel Corpus from Odia Twitter 1

| Introduction | 2 |
|--|----|
| Methodology | 3 |
| Why Parallel corpus | 3 |
| How to make a pair and a corpus | 4 |
| The usefulness of these pairs | 5 |
| Why Parallel corpus Stream | 5 |
| Why social media is an effective medium for streaming parallel pairs | 6 |
| Why twitter is chosen for the stream | 6 |
| Why not other form of social media like facebook | 7 |
| Procedure : Steps to extract parallel pairs and suggested usage | 7 |
| Why this is effective | 8 |
| Problem formulation | 8 |
| The Corpus Stream Equation | 11 |
| Solution | 11 |
| Community Odia twitterati (CoOrT) | 11 |
| Eligible Pairs (Ep) | 12 |
| Algorithms | 12 |
| ToolSet | 12 |
| Process | 13 |
| Applications | 14 |
| Extensions | 14 |
| Conclusion | 15 |
| References | 16 |
| Glossary | 16 |
| Appendix | 17 |
| | |

Introduction

A corpus is a fundamental need for natural language process applications.

A parallel corpus is a foundational need for languages like Odia (Oriya - The Unicode Standard, Version 13.0."https://unicode.org/charts/PDF/U0B00.pdf. Accessed 8 Aug. 2020) which would enable explorations in natural language processing advancements into machine translation, (Machine translation - Wikipedia. "https://en.wikipedia.org/wiki/Machine translation. Accessed 8 Aug. 2020) computational language modelling, (Language model - Wikipedia."