Lacuna Fund: Our Voice in Data

Resources for Proposals in NLP 2024

This document represents a collection of resources from the Technical Advisory Panel (TAP) as an addition to those referenced in the RFP document. These are intended to provide assistance in obtaining relevant background information, preparing a competitive proposal, and completing quality work.

These resources are not intended to be exhaustive nor authoritative. This document does not represent an endorsement of work by the Lacuna Fund Secretariat, the TAP, or individual members.

ACADEMIC PAPERS (ENGLISH)

- Essentials of Language Documentation. It is a compilation of articles edited by J. Gippert, N. Himmelmann and U. Mossel on various topics related to language documentation. These include the discipline's specific workflow, fundamental ethical aspects, and its relationship with other fields of linguistic work. This compilation can serve as a valuable starting point for those interested in the creation of corpora that will later be used in NLP.
- Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*. (*COLING 2020)*: It´s a review of colonizing discourses in speech and language technology, and suggests new ways of working with Indigenous communities, and seeks to open a discussion of a postcolonial approach to computational methods for supporting language vitality.
- The "Colonial Impulse" of Natural Language Processing: An Audit of Bengali Sentiment Analysis Tools and Their Identity-based Biases is a critique of sentiment analysis, associated with a specific Asian language.
- Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*. ACL Publications, 55-69 This paper reviews the research, the digital resources and the available NLP systems that focus on indigenous languages of the Americas.
- Datasheets for Datasets. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, this paper propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on, in order to facilitate better communication between dataset creators and dataset consumers, and encourage the machine learning community to prioritize transparency and accountability.
- Communities, ethics and rights in language documentation. This paper is about the general topic of communities, ethics and rights as they relate to language documentation, especially in the context of endangered languages.

ACADEMIC PAPERS (SPANISH)

- Lingüística de la documentación: textos fundacionales y proyecciones en América del Sur. Lucía Golluscio. It presents a compilation, in Spanish, of the main texts that provide the general

framework for defining the language documentation, as well as chapters aimed at presenting current experiences in South America.

FRAMEWORKS (ENGLISH)

- CARE Principles for Indigenous Governance– The Principles complement the FAIR principles encouraging open and other data movements to consider both people and purpose in their advocacy and pursuits.
- Check Before You Tech— A guide for communities choosing language apps and software. While meant for technology users, this resource can serve as a helpful guide for developers in order to lead with an ethical data approach to understand the questions communities are pondering when using language technology.
- International Decade of Indigenous Languages (2022-2032) - The goal of the International Decade is to guarantee the right of indigenous peoples to preserve, revitalize and promote their languages, and to integrate aspects of linguistic diversity and multilingualism in sustainable development efforts, with a particular focus on digital empowerment and language technologies.
- Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.
- Data-mining and Extraction: the gold rush of AI on Indigenous Languages. This paper starts a discussion on the topic of Data mining and Extraction of Indigenous Language data, describing recent events that took place within the Algonquian Dictionaries and Language Resources common infrastructure.
- Ethics in linguistics. A deep review of existing literature on ethics in linguistics, both as it relates to research and as it relates to broader practices, which we then situate within ongoing conversations across subfields.

FRAMEWORKS (SPANISH)

- Protocolo para el proceso de consulta y consentimiento libre,previo e informado con los pueblos indígenas que habitan en el Paraguay. Decree N 1039/2018. A appropriate approach to understand how Latin American and Caribbean states are regulating the relationship between scholars and communities.
- Lineamientos para el comportamiento ético en las Ciencias Sociales y Humanidades. Resolution N° 2857/2006 by the ethical committee of the CONICET in Argentina.

BOOKS

- Karën Fort on data annotation in NLP. This book presents a unique opportunity for constructing a consistent image of collaborative manual annotation for Natural Language Processing (NLP). NLP has witnessed two major evolutions in the past 25 years: firstly, the extraordinary success of machine learning, which is now, for better or for worse, overwhelmingly dominant in the field, and secondly, the multiplication of evaluation campaigns or shared tasks. Both involve manually annotated corpora, for the training and evaluation of the systems.
- Bases de la documentación lingüística. Besides being a manual of field techniques, this book offers a valuable set of reflections on linguistic fieldwork. It is an indispensable reference not only for those who work with indigenous languages and peoples, but also for those directly

involved in the collection and management of linguistic data in general and those who work with linguistic practices of a given community, indigenous or not.

DATABASES

- The South American Indigenous Language Structures (SAILS) is a large database of grammatical properties of languages gathered from descriptive materials (such as reference grammars) by a team directed by Pieter Muysken. SAILS Online was programmed by Harald Hammarström using the CLLD framework, with support from Robert Forkel.
- Sound of the Andes is a database containing lexical and phonological information on languages of the Quechua, Aymara, and Mapuche families. It is an interactive site that brings knowledge of these languages to the general public while presenting the information in a clear, transparent, and accessible manner for various computational analyses focused on comparing these languages.
- Glottolog 5.0 is a bibliographic database of the world's lesser-known languages made by Hammarström, Forkel, Haspelmath & Bank in the Max Planck Institute for Evolutionary Anthropology.

OTHER RESOURCES ON OPEN DATA

- Metatext is List of Translation Datasets for Machine Learning Projects including High-quality datasets are the key to good performance in natural language processing (NLP) projects. They have collected a list of NLP datasets for Translation task, to get started machine learning projects.
- Metatext large curated training base for Translation.
- Ancient Natural Language Processing aims to provide resources and tools for scholars, students, and enthusiasts who are interested in applying NLP techniques to ancient languages. Here can be found information about various projects that use NLP for ancient languages, such as machine translation, text analysis, and language learning. Includes online courses and tutorials that teach how to use NLP tools for ancient languages.
- 4th Workshop on Indigenous Languages in the Americas is a workshop colocated with the North American Chapter of the Association for Computational Linguistics (NAACL) 2024, with various resources and papers.