

# Empowering Indigenous Communities through Citizen Linguistics, Language Resources and Human Language Technologies

Christopher Cieri, Mark Liberman

University of Pennsylvania, Linguistic Data Consortium

3600 Market Street, Philadelphia, PA 19104 USA

{ccieri, myll}@ldc.upenn.edu

## Abstract

This paper demonstrates the link between UNESCO goals, language technologies and the requirements of language resources. It describes the causes of the scarcity of language resources and proposes a novel method to increase the supply of linguistic data by empowering indigenous language communities to contribute directly.

**Keywords:** citizen linguistics, language resources, corpora, human language technologies

## Résumé

Questo documento dimostra il legame tra gli obiettivi dell'UNESCO, le tecnologie linguistiche e i requisiti delle risorse linguistiche. Descrive le cause della scarsità delle risorse linguistiche e propone un nuovo metodo per aumentare la fornitura di dati linguistici dando potere alle comunità linguistiche indigene di contribuire direttamente.

## 1. Introduction

A chain of dependencies links UNESCO goals regarding the Indigenous Communities directly to Language Technologies and then in turn to Language Resources highlighting the critical need for such resources and for innovations that supplement current methods in order to accelerate the production to the benefit of indigenous communities.

### 1.1 UNESCO Goals

UNESCO supports indigenous languages in order to preserve unique knowledge systems, to promote peace through international cooperation and sustainable development, ensure fundamental human rights, improve education and move toward an inclusive society that acknowledges the value of cultural diversity and heritage<sup>1</sup>.

The UNESCO International Year of Indigenous Languages 2019 operates in five key areas of which the following are relevant to Language Technologies and Language Resources: increasing understanding and international cooperation, supporting knowledge sharing relative to indigenous languages<sup>2</sup>, integrating indigenous languages into the digital society, empowering indigenous communities through capacity building, elaborating new knowledge to support growth and development.

The International Conference *Language Technologies for All (LT4All)* promotes the “*human rights and fundamental freedoms of all language users to access information and knowledge in languages that are best understood*” and “*linguistic diversity, truly multilingual internet and language technologies, with special focus on indigenous languages.*”<sup>2</sup>

### 1.2 Language Technologies

The term *Human Language Technology* in the current context refers to any technology that operates upon any human language whether spoken or written or signed.

These include Language Identification technologies that recognize what language is being spoken based upon a few seconds of speech. Variants of such technologies also recognize which dialect of a language is spoken and recent work focuses on trying to detect the linguistic origin of the speakers based upon their speech in their native or non-native languages. Speaker Recognition and allied technologies indicate whether two utterances were produced by the same speaker or sort the utterances of a long recording according to the speakers. Speech Recognition technologies automatically produce a transcript of spoken language; related technologies within Dialog Systems convert spoken commands into operations a system can perform. Speech Synthesis or Text to Speech technologies reverse the process and produce utterances of a text or an representation of knowledge retrieved by a system, for example the current time or weather. Information Retrieval and related technologies find written and spoken documents related to a query whether it is expressed as a sequence of search terms or an example document. Information Extraction, and allied technologies find the entities and events and relations among them in a text or spoken document for purposes of answering questions and building knowledge bases. Other Natural Language Processing technologies map the relationship between grammar and meaning in spoken or written language. Finally, Machine Translation, including Speech to Speech Translation, translates content from one language to another.

Together these Language Technologies address many of the goals UNESCO has identified with respect to its work with Indigenous Communities. Specifically they offer methods for constructing and easing access to knowledge in indigenous languages and making that information available to other communities. Similarly language technologies offer ways to provide the world’s knowledge to indigenous communities in their own languages. By enabling the development and free flow of information in the languages spoken by user communities, including

<sup>1</sup> <https://en.iyil2019.org/>

<sup>2</sup> <https://en.unesco.org/LT4ALL>

indigenous communities, language technologies support the UNESCO goals of protecting fundamental human rights, improving education, promoting sustainable development and international cooperation while acknowledging the value of diversity.

The dominant paradigm in Human Language Technology research and development over the past decades, and the one that has led to such marked advancement, is that of *machine learning*. Under this approach a class of general purpose algorithms develop their ability to process linguistic data by emulating specific human behaviors encoded in annotated data. Thus a system to translate from, for example, Xhosa to Sotho is built from general purpose algorithms and many examples of utterances translated in that way. A great advantage of this approach is that roughly the system can be trained to perform translations between a different pair of languages by providing it with the appropriate data. Another advantage is that system performance tends to increase with the quantity and quality of the “training data” provided and such data has other uses. A disadvantage is that such algorithms tend to require large amounts of training data.

### 1.3 The Role of Language Resources

Nearly all human language technologies, most modern research into language and great deal of pedagogical materials development rely upon the existence of *Language Resources* by which we mean here: organized collections of records of spoken, and in some cases written, language with annotations that are typically contributed by humans, often aided by technology, in order to support analysis. For purposes of documenting a language the most important of these are what are often called *raw data*: collections of speech, captured in audio or increasingly video recordings, and of texts if the language is written (Good 2011).

*Annotation*, by which we mean the application of human judgement to raw data, whether directly or mediated by computer, vastly increases its usability (Cieri 2015). The commonest annotation is transcription which generally employs the language’s native orthography if one exists in common usage. For purposes of language documentation, descriptive resources such as dictionaries and grammars supplement the raw data. Human language technology developers similarly require raw data but, with a few notable exceptions, tend to rely more heavily on annotations of raw data from which they can extract statistical information than on descriptive resources of the kind created by documentary linguists.

Human language technology performance can be sensitive to the situations under which training data is collected, for example the microphones used, the interlocutors or the genre, thus increasing requirements on quantity, diversity and quality control in language resource development.

## 2. Language Resource Scarcity

Having traced the chain of dependency from UNESCO goals to language technologies and from language technologies to language resources, we come to the central problem, that of language resource scarcity.

Despite the energetic efforts of a large number of:

- data centers such as Linguistic Data Consortium (LDC)<sup>3</sup>, European Language Resources Association (ELRA)<sup>4</sup>, Chinese LDC<sup>5</sup>, LDC for Indian Languages<sup>6</sup> and the South African Centre for Digital Language Resources (SADiLaR)<sup>7</sup>
- national or regional corpus efforts such as those for Austrian German<sup>8</sup>, British English<sup>9</sup>, Croatian, Czech<sup>10</sup>, German<sup>11</sup>, Hungarian<sup>12</sup>, Irish (Uí Dhonnchadha 2012), Maltese<sup>13</sup>, Dutch<sup>14</sup>, Polish<sup>15</sup>, Russian<sup>16</sup>, Slovakian<sup>17</sup>, South Tyrolean<sup>18</sup>, Swiss German<sup>19</sup>, US English<sup>20</sup>, and Welsh<sup>21</sup>
- multination projects to create and share language resources such as CLARIN<sup>22</sup> and META-SHARE<sup>23</sup>
- countless research laboratories that produce language corpora such as LIMSI<sup>24</sup>

it remains true that the number of publicly available language resources is only a tiny fraction of those needed to document and support the development of technologies for the world’s languages. Why should this be so?

First, the number of languages in the world is large, more than 7000 by some counts (Eberhard, Simons & Fennig 2019) and the number of resources needed to create a minimal set of technologies for any one language is also not small, perhaps two dozen (Krauwert 1998, Binnenpoorte, et al. 2002, Krauwert 2003). The result is that all of the world’s languages lack at least some of the language resources needed, even the languages of the wealthiest nations in the European Union (Rehm and Uszkoreit 2012). However it is also the case that new production does not proceed in a way that maximizes coverage of languages and resource types; rather considerable effort is devoted toward increasing the size of existing resources or producing new versions. (Cieri 2017). Even programs that focus on *under-resourced languages* tend to select from among these language with large

<sup>3</sup> <https://www ldc.upenn.edu>

<sup>4</sup> <http://www.elra.info>

<sup>5</sup> <http://www.chineseldc.org>

<sup>6</sup> <http://www.ldcil.org>

<sup>7</sup> <https://www.sadilar.org>

<sup>8</sup> <http://www.aac.ac.at>

<sup>9</sup> <http://www.natcorp.ox.ac.uk>

<sup>10</sup> <https://www.korpus.cz>

<sup>11</sup> <https://www1.ids-mannheim.de/s/corpus-linguistics/projects/corpus-development.html?L=1>

<sup>12</sup> <http://corpus.nytud.hu/mnsz>

<sup>13</sup> <http://mlrs.research.um.edu.mt>

<sup>14</sup> <http://lands.let.ru.nl/cgn>

<sup>15</sup> <http://nkjp.pl>

<sup>16</sup> <http://www.ruscorpora.ru>

<sup>17</sup> <https://korpus.sk>

<sup>18</sup> <http://www.korpus-suedtirol.it>

<sup>19</sup> <https://www.chtk.ch>

<sup>20</sup> <http://www.anc.org>

<sup>21</sup> <http://codah.swan.ac.uk/?p=334>

<sup>22</sup> <https://www.clarin.eu>

<sup>23</sup> <http://www.meta-share.org>

<sup>24</sup> <https://www.limsi.fr/fr/plateformes-et-ressources/corpus>

numbers of native speakers who control large portions of the world's wealth (Cieri 2016).

In short, our current approaches to creating language resources that enable language technology development will not adequately address the scarcity problem in the foreseeable future leaving us to face decades of the same kind of imbalance we currently seek to correct.

### 3. Innovative Solutions to Scarcity

One reason for the insufficiency of current approaches to creating resources to document the world's language is that it applies a finite and relatively small resource, funding, to a problem that, if not infinite, is at least multiple orders of magnitude larger.

An alternative is to identify renewable sources of the time and intellectual investment required. We take as our model a number of activities that show that the human drive for challenge, advancement, entertainment and the opportunity for people to contribute to their own betterment and that of their local communities and broader society are effectively boundless. This has been made clear repeatedly in the vast numbers of hours spent each day around the world in social media. More immediately relevant, tens of millions of language identification judgements were proffered by players of now defunct GreatLanguageGame (Skirgård, Roberts, & Yencken 2017) and hundreds of millions of contributions have been submitted by nearly two million contributors to the Zooniverse<sup>25</sup> citizen science portal. By providing similar incentives, we offer indigenous language communities a platform in which they can contribute directly to the documentation of their languages and the development of technologies that advance the UNESCO goals sketched above.

*LanguageARC* is a portal for the Citizen Science of Language, hereafter Citizen Linguistics.

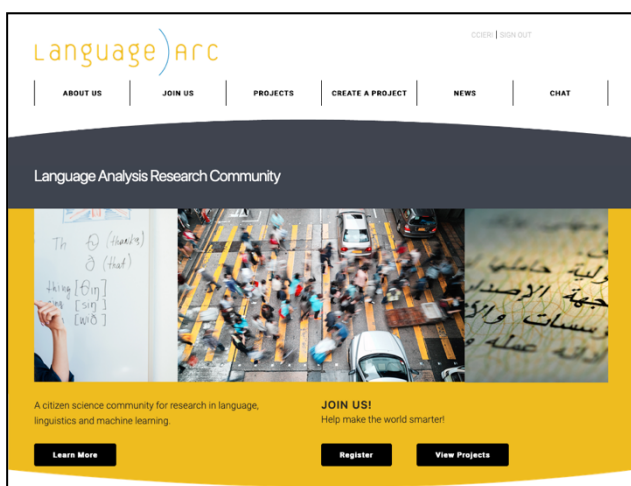


Figure 2: *LanguageARC Citizen Linguist portal*

*LanguageARC* presents Citizen Linguists with multiple *projects* to which they can contribute. Each project contains one or more *tasks*; each task involves a simple activity that may be applied to one or more *items*. For example a project

might seek to document the linguistic diversity of Italy including the regional and local dialects that are being displaced by the standard language and that have been suppressed by former governments. One task might ask contributors to name culturally relevant items from pictures while another might ask them to describe the people, things and activities they see in a sequence of silent videos. In these cases, the *items* are the pictures and video.

*LanguageARC* introduces each project via its title, call to action, image and pitch (as in elevator pitch). Each project may also include picture and descriptions of its research team and badges representing its partner organizations. To support community building with the project, each may also offer a range of discussion forums and a blog (currently external). Each task within a *LanguageARC* project may have its own title, call to action and image as well as a tutorial and reference guide and one additional discussion forum specific to the task.

*LanguageARC* tasks ask contributors to display provided texts or images or to play audio or video clips and to respond to instructions that are specific to the task or that vary with each item by speaking, enters a text response or selecting one or more items from a multiple choice list.

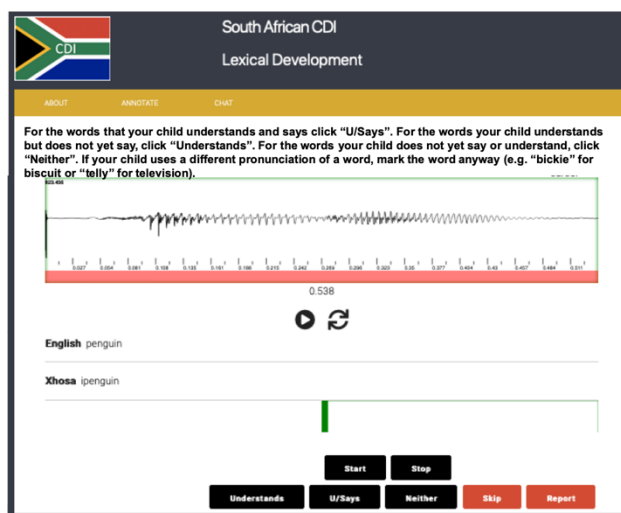


Figure 1: *Task prototype for collecting the Communicative Development Inventory in South Africa*

Figure 1 contains an image of a prototype, somewhat over-designed to show all possible approaches to collecting the Communicative Development Inventory (Hamilton, Plunkett, & Schafer 2000). for the languages of South Africa. In this case, mothers with young children indicate whether their child has active or passive knowledge of a number of common words for culturally relevant objects. The instructions are in English for the current readership. Beneath the instructions is an audio widget that plays the word in this case in the Xhosa language. Below that are text boxes showing the word in English and Xhosa (presumably only the latter would be used and only if the mother were literate in the language). Beneath that is a recording widget so that the mother can provide a spoken answer and at the bottom of the screen are multiple choice buttons in black and additional red buttons that allow the contributor to Skip

<sup>25</sup> <https://www.zooniverse.org>

or Report that something is wrong with the item (e.g. the audio did not play).

LanguageARC was built upon a toolkit that the Linguistic Data Consortium has used to create millions of annotations across more than 100 language resource projects over the past decade. The toolkit has been extended to make it open source, portable to new environments and capable even of being deployed to a laptop and taken into the field where internet access is not available. LanguageARC includes a project builder that allows users with no software development experience to create and deploy tasks in less than one hour each assuming the data and instructions are already available in an appropriate format

#### 4. Conclusion

UNESCO aim related to Indigenous Community rely necessarily upon Language Technologies which rely in turn upon Language Resources which are absent for most of the world languages. Current approaches will not solve the language resource scarcity problem in an acceptable timeframe. The use of novel incentives such as those offered by the LanguageARC citizen linguistics portal empowers indigenous communities participate directly in the creation of language resources that benefit themselves principally by enabling technology development but also by encouraging linguistic research and the creation of pedagogical materials.

#### 5. Acknowledgements

The authors acknowledge the support of the National Science Foundation via CISE Research Infrastructure (CRI) grants CRI CI-P 1629923 and CRI CI-NEW 1730377 as well as number partners who make this work possible.

#### 6. Bibliographical References

Binnenpoorte, Diana, Catia Cucchiarini, Elisabeth D'Halleweyn, Janienke Sturm and Folkert de Vriend (2002) Towards a roadmap for Human Language Technologies: Dutch-Flemish experience in Proceedings of the workshop "Towards a Roadmap for Multimodal Language Resources and Evaluation" at LREC 2002, Las Palmas, Canary Islands, June.

Cieri, C. (2017) Addressing the Language Resource Gap through Alternative Incentives, Workforces and Workflows, Keynote Speech at the 8th Language & Technology Conference, November 17-19, Poznań, Poland.

Cieri, Christopher, Mike Maxwell, Stephanie Strassel, Jennifer Tracey (2016) Selection Criteria for Low Resource Language Programs in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 23-28, Portorož, Slovenia.

Cieri, C. (2015) Data Bases and Statistical Systems: Linguistics In James Wright, ed. International Encyclopedia of Social & Behavioral Science 2nd Edition, Elsevier.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the World. Twenty-second edition. Dallas, Texas: SIL

International. Online version: <http://www.ethnologue.com>.

Good, J, (2011) Data and language documentation. In Peter Austin and Julia Sallabank (eds.), Handbook of Endangered Languages. Cambridge: Cambridge University Press. 212–234.

Hamilton, A., Plunkett, K., & Schafer, G., (2000). Infant vocabulary development assessed with a British Communicative Development Inventory: Lower scores in the UK than the USA. *Journal of Child Language*, 27, 689-705.

Uí Dhonnchadha, E., Frenda, A., Vaughan, B., (2012) Issues in Designing a Corpus of Spoken Irish, LREC: SALT MIL-AFLaT Workshop on "Language technology for normalisation of less-resourced languages, Istanbul, May 2012, edited by G. De Pauw, G-M de Schryver, M. Forcadea, K. Sarasola, F. Tyers, P. Waiganjo Wagach , 2012, pp1-6.

Krauwer, Steven (1998) ELSNET and ELRA: Common past, common future, ELRA Newsletter, Vol. 3:2, May.

Krauwer, Steven (2003) The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap, in International Workshop Speech and Computer (SPECOM-2003).

Rehm, Georg and Hans Uszkoreit, eds. (2012) META-NET White Paper Series: Europe's Languages in the Digital Age, URL: [www.meta-net.eu/whitepapers](http://www.meta-net.eu/whitepapers).

Skirgård, H., S.G. Roberts, L. Yencken (2017) Why are some languages confused for others? Investigating data from the Great Language Game, *PLOS ONE*, 12 (4) (2017), p. e0165934, 10.1371/journal.pone.0165934

Tadić, M. (2002). Building the Croatian national corpus. In Proceedings of LREC'2002 (pp. 441–446).