Global Open Resources and Information for Language and Linguistic Analysis (GORILLA)

Damir Cavar, Małgorzata E. Ćavar, Lwin Moe
Indiana University
Bloomington, IN, USA
{dcavar,mcavar,lwinmoe}@indiana.edu

Abstract

The infrastructure Global Open Resources and Information for Language and Linguistic Analysis (GORILLA) was created as a resource that provides a bridge between disciplines such as documentary, theoretical, and corpus linguistics, speech and language technologies, and digital language archiving services. GORILLA is designed as an interface between digital language archive services and language data producers. It addresses various problems of common digital language archive infrastructures. At the same time it serves the speech and language technology communities by providing a platform to create and share speech and language data from low-resourced and endangered languages. It hosts an initial collection of language models for speech and natural language processing (NLP), and technologies or software tools for corpus creation and annotation. GORILLA is designed to address the Transcription Bottleneck in language documentation, and, at the same time to provide solutions to the general Language Resource Bottleneck in speech and language technologies. It does so by facilitating the cooperation between documentary and theoretical linguistics, and speech and language technologies research and development, in particular for low-resourced and endangered languages.

Keywords: Speech and Language Corpora, NLP, Low-resourced Languages, Transcription Bottleneck

1. Introduction

The Global Open Resources and Information for Language and Linguistic Analysis (GORILLA) project[1] builds a cyber-infrastructure and platform that addresses numerous issues related to digital language resources, language documentation, and speech and language technologies. A particular focus of the project is on low-resourced or endangered languages. We assume that currently the majority of all languages in the world belongs to this group, since major resources in form of corpora, treebanks, transcribed and annotated speech data, or just dictionaries and formal grammars are missing.[2]

The goal of GORILLA is to bridge the gap between large amounts of available language documentation data in various archives and repositories, and the lack of speech and language data for Natural Language Processing (NLP) or Human Language Technologies (HLT) research and development needs. We expect that documentation, transcription and annotation of available language resources can be improved, if NLP and HLT communities cooperate with documentary linguists and language archives. At the same time, making language resources from low-resourced and endangered languages accessible to the NLP and HLT community will significantly improve common speech and language algorithms and technologies, and indirectly the entire language documentation and research discipline, as it already did for the most common 1% of high-resourced languages.

2. Language Documentation Resources

In their attempt to document as many languages as possible, in a race against time, with languages dying at a growing pace, endangered language documentation projects tend to create large amounts of resources that are usually archived and stored in language repositories or archives without detailed transcription or linguistic analyses that is usually provided with data and corpus annotations. The methods and approaches to language documentation differ in significant ways between researchers and schools. Some documentary linguists follow a tradition of informed and targeted language documentation that collects data that is relevant for very specific research questions. Votaries of a more "naïve" documentary fieldwork strategies collect material of endangered languages independent of concrete theoretical considerations or questions.

Apart from the differences in approach, there is a broad spectrum of formats in which materials are collected. In some archives, e.g. in the Archive of Traditional Music at Indiana University, one can find recordings on phonographic wax cylinders or magnetic steel wire. For this type of resource, the necessary first step towards general availability for the different research and interest communities is digitization. Current language documentation methods involve digital recordings using more or less adequate technology and storage formats. Given the wide accessibility of mobile recording devices and growing amounts of storage memory, the amount of language documentation data is also growing continuously.

At this time, technology alone does not solve central problems that relate to the quality of the recordings, which result from common data acquisition scenarios. Fieldworkers mostly record and acquire data in elicitation sessions, by recording stories that are told by native speakers, or by documenting ceremonies, prayers, songs, and multi-party con-

[1] See `http://gorilla.linguistlist.org`.

[2] Depending on the question and need, as Christopher Cieri (LDC) pointed out (p.c.), even high resources languages like English could be considered low-resourced with respect to some resource types that are missing currently or in the process of emerging new technologies and approaches. The classification of low-resourced is a relative and dynamic one that needs to be put in a context.

versations. Since most of these settings are located in the natural environment of the speakers – in their houses, in villages, on public squares or places, or in nature – the recordings usually contain noise or multiple speakers interacting in discourse situations. The researcher is often recorded interacting with the speakers. The common and transportable recording technology is not able to focus ideally on the targeted speakers since often device mounted or static microphones are used that are not following head or body motions. Variation of speech quality during recorded sessions is a common feature of documentary linguistic recordings. The environments most of the time do not provide ideal studio conditions for recordings of spoken language. These properties of language documentation impede further processing.

Independent of the quality and environmental conditions that impact the quality of the recorded material, the primary goal of documentary linguists is usually not to create high quality audio and video recordings of spoken language for corpus linguistic or speech technology purposes. Obtaining high-quality recordings is often of no relevance for the research questions that motivate documentary linguists to acquire the material in the first place. The recordings need to be good enough so that a trained expert who speaks the language can decipher the content. In other words, documentary linguists are not necessarily concerned with the needs of speech and natural language processing researchers or corpus linguists.

2.1. The Transcription Bottleneck

Multiple collections of recordings from low-resourced or endangered languages are archived in one of the many digital archives without transcription or further annotation. While it might be desirable to provide transcription and annotation of the resources and enable other researchers to use the material from a research perspective, more often than not the funds for this additional effort are lacking. Often the detailed annotation is not part of the stated research questions or goals and thus not funded by research funding agencies. The archives themselves do not provide any service that is related to linguistic analysis or annotation for the material.

For example, the *Documentation of Endangered Languages* (DOBES) archive,[3] the Archive of the Indigenous Languages of Latin America (AILLA) at UT Austin,[4] or The Endangered Languages Archive (ELAR) at SOAS, University of London[5] host large collections of recorded audio and video material from extinct, endangered, or extremely under-resourced languages. The archived audio and video language recordings are to a large extend not transcribed, translated, or otherwise linguistically analyzed and annotated. For some resources, even important metadata is missing.[6] The *Chatino Documentation of H. Cruz* at AILLA[7] – for example – contains almost 8 hours of audio and more than 4 hours of video recordings of Chatino, out of that only 2% of the material is transcribed. The *Balsas Valley Nahuatl Collection of J. Amith* in the same archive[8] contains more than 361 hours of recordings and no transcriptions or annotation has been made available to accompany the collection.

Documentary linguists are also not necessarily concerned with the specific formats and annotation standards that speech and language technologies might require. They are maybe also not aware or interested in corpus standards and the different tools for quantitative corpus studies. The gap between technological needs and know-how in the different sub-disciplines of linguistics and NLP/HLT is continuously growing. This might be due to little or no real interaction between these sub-disciplines. Thus, we assume that most language material from documentation projects ends up in digital language archives without transcription or any kind of linguistic analysis or annotation.

Many archives and repositories have to face the problem with this so-called *transcription bottleneck*. Transcription of audio and video resources requires expert knowledge of the particular languages and variants, an appropriate infrastructure and experts for technologies, and strategies that follow common standards and procedures for digital annotation. Transcription and annotation is extremely time consuming and resource intensive. Providing annotations as a result of linguistic analysis by experts or just an adequate translation into one of the current major languages is even more problematic and time-consuming. Unfortunately, without these transcriptions and annotations the language resources are only accessible to speakers of the languages, that is, a very narrow audience.

To summarize, we can assume that the current situation in terms of the transcription bottleneck is: the transcription of one hour audio or video recordings of speech takes 50 to 100 times real time, often even more. We did not identify any clear analysis of the time needed for these tasks for different types of languages. This can only be understood as an estimate that provides a basic transcription and maybe also part-of-speech tagging and translation. Consequently, it is prohibitively expensive to transcribe these resources and thus most of the already collected recordings are not transcribed or analyzed at all, while more and more raw audio and video material is being collected or made available via various channels.

2.2. The Language Graveyard

Another issue with the lack of transcriptions is that – given the nature of audio and video material and the current state of audio and video processing technologies – efficient and fast searching over content in the audio material is not possible. This renders any kind of research on larger amounts of audio and video based language data a very time-consuming process. Finding relevant information in the audio/video recording is not the only problem. Particular fragments cannot be understood and interpreted without a broader context, forcing researchers to listen or watch

[3] See http://dobes.mpi.nl.

[4] See http://www.ailla.utexas.org.

[5] See http://elar.soas.ac.uk.

[6] See for example the *Documentation of Effutu* collection at ELAR http://elar.soas.ac.uk/deposit/0175.

[7] See http://www.ailla.utexas.org/search/collection.html?c_id=157.

[8] See http://www.ailla.utexas.org/search/collection.html?c_id=1/.

larger sequences of the material multiple times. Again, without transcription and annotation, the material is not easily accessible to linguistic (or any other) research, not even if conducted by researchers who already have a high degree of expertise in a relevant language.

It is worth noting that experts of most endangered languages are as endangered as most of the languages themselves. Many older resources in the Archive of Traditional Music at Indiana University, for example, that are currently being digitized and archived will most likely not be accessible anymore, since the recorded languages have already been extinct for a while and there are no expert speakers available. There is most likely also not any independent motivation for researchers to gain the necessary expertise to be able to provide transcriptions and linguistic annotations.

Looking from another perspective, that of speech and language technologies, the existing resources, in spite of their potential value, do not contribute to any kind of NLP- or HLT-research, i.e. they do not constitute a corpus or language resource that can be utilized for the quantitative or qualitative study of languages or the development of speech and language technologies.

In his recent criticism of certain priorities of the endangered languages documentation movement, Paul Newman (2013) rightly refers to these archives as *language graveyards*, a concept used earlier by Christian Lehmann (Lehmann, 2001, 85) and, most surprisingly, by Nikolaus Himmelmann himself (Himmelmann, 2006, 4), Himmelmann being one of the leading exponents of the "naïve" documentary fieldwork. Newman is emphasizing the lack of contribution of the extremely valuable resources to any kind of valuable academic or research questions or goals.

The situation is even more awkward for the speaker communities. The large amounts of resources in the *language graveyards* are not available to the community either. They cannot use them for the development of dictionaries or language teaching material. Revitalization projects suffer from the lack of language resources while at the same time valuable material is in sight, but unfortunately not accessible, because it was never explicitly analyzed.

2.3. Access and Licensing

In addition to the lack of transcriptions and annotations, access restrictions imposed for specific collections are obstacles that prevent the resources from being used for derivatives or other subsequent research. They often prevent the resources from being integrated in derivatives like speech and language technologies. Interested users are often confronted with access restrictions that require personal presence in the archive or that limit the use in ways that make certain types of research or studies impossible (e.g. copies and distribution not permitted, neither extracted models).

Access restrictions or licenses vary significantly between the different data collections even within one archive. The use of language data from specific collections can be restricted such that it does not allow for the creation of language models, or raw data sets for engineering or training speech and language applications. A very common restriction is a ban of any kind of commercial use of the material or materials derived from it. This type of restriction potentially harms the speaker communities, which often would be economically challenged. Commercialization by speaker communities for the development of educational material or any kind of specific application to teach, analyze or process a particular language often cannot be developed using such restricted material. Providing the possibility to commercialize language material and products that are based on it can be a useful instrument in a sustainable language revitalization strategy. This type of restriction might effectively harm the language community and the language itself.

While some archives and repositories specify processes for the negotiation of particular access permissions and usage licenses, for most of the archived material any kind of alternation of the licensing restrictions is impossible to achieve. One of the reasons can be that restrictions were imposed by Institutional Review Boards (IRB) or research funding agencies when the data collection was approved for particular research activities. Some restrictions are a consequence of general regulations for experimental data obtained from human subjects. Other restrictions might have been introduced by decisions from speaker communities or tribal boards not to disseminate language data because of privacy concerns or religious beliefs. Often enough documentary linguists restricted access to their collections themselves. Thus, changes and negotiations of access restrictions for archived language resources are most often impossible to achieve.

Current guidelines and policies – however – proposed by research funding agencies do promote sharing of resources that were government or publicly funded. We expect the amount of accessible data to grow continuously. However, this will not necessarily impact the very common restriction to exclude any kind of commercial use of the relevant language resources.

3. The Language Resource Bottleneck

There are ever growing amounts of digital audio and video language resources not only emerging from language documentation work, but also from the normal use of modern technologies, social media, etc., by communities of language speakers. These audio and video resources are not accessible because of the lack of useful annotation. Resources like for example YouTube (`http://youtube.com`) most likely contain treasures for linguistic research from low-resourced and endangered languages that are not transparent to researchers or speech and language technology engineers.

On the other hand, speech and language technologies are facing a different kind of a bottleneck. We estimate that the majority of resources, i.e. corpora, models, speech and language processing algorithms and applications, are available for only about 1% of the world's languages. Sufficient resources in form of corpora, lexicons, time-aligned spoken language transcriptions and annotations, models of phonotactic regularities, part-of-speech tagged texts or treebanks are not freely available for the majority of approx. 7,097 living languages of the world.[9]

[9] The number of living languages is based on the list of lan-

Resources produced by documentary linguists cannot be directly used as corpora for speech and language technologies. Many languages are in fact just oral and do not have a standardized orthography or a robust collection of written texts that could serve as the base content for corpus creation, pronunciation dictionaries or language models. As a result, language technologies are currently developed using only the resources from a very limited subsection of best-documented languages. This is a situation that leads to a kind of technological mono-culture. We assume that the lack of linguistic diversity, when it comes to language resources, and the evolutionary nature of the speech and language technology solutions and engineering processes seriously limit the directions and outcomes of research in language technologies.

The evolutionary process in speech technology, for example, as laid out by the National Institute of Standards and Technologies (NIST) and other organizations, favors in annual challenges and competitions the best performing algorithms that are evaluated on a small number of corpora or language data.[10] The best performing algorithms and systems are, for example, chosen every year for subsequent research and development. Given the limitation with language resources, this selection procedure might prevent potentially valuable insights that could be triggered by experiments and evaluations based on resources from thus far understudied languages. Languages that are usually not used to evaluate current algorithms, heuristics and models frequently happen to have very unusual linguistic properties. For instance, specific rich tonal and accent systems could affect feature extraction approaches for speech recognition systems. The analysis and computational modeling of morphological and lexical properties, as for example in polysynthetic languages, might impact algorithms for lexical processing or language models in general.

To recapitulate, improvements of algorithms and technologies are developed and evaluated on the basis of resources from a limited set of languages. Research communities concerned with speech and language technologies lack data that would potentially help them to develop more generic and improved algorithms based on linguistic universals. The language resource bottleneck, i.e. the fact that for most languages on the planet sufficient descriptive resources are missing, as well as the lack of basic speech and NLP components for these languages, limits our abilities to research on languages from the linguistic and NLP/HLT-perspective. At the same time, language documentation and revitalization projects could benefit from speech and language technologies, if these would be available. Speech technologies could reduce the time needed for transcription of audio and video recordings. Such technologies could automatically time align available transcriptions for a given recording and reduce the time needed to produce a speech corpus, which then could improve the time and cost aspects of the transcription effort overall. Language processing tools like part-of-speech taggers, parsers or different kind of annotators could improve the production of dictionaries and education material. Machine translation technology could help making the content accessible beyond the groups of researchers and engineers, to the general public. It could also help to generate resources by translating content and texts from other languages to the under-resourced or endangered language on a larger scale, thus contributing to a revitalization effort.

The efforts to increase the availability of speech and language technologies for low-resourced languages is also hampered by the lack of their direct market value. Communities of low-resourced or endangered languages are not ideal targets for industrial investments, and research and development projects focusing on their language. While we would like to see a Chatino version of some conversational agent on a mobile phone, it is very unlikely to happen without other factors making it possible. Yet, although the private sector most likely will not engage in projects for low-resourced or endangered language resources and technologies, it might use such created resources in its technologies, if those would be accessible. Without intervention, however, without the cooperation between private sector research and development in the speech and language technology sector, the technology gap between the dominant and hyper-resourced languages and the low-resourced and endangered ones will continue growing, and the consequences could be that the process of decreasing language diversity is further reinforced.[11]

4. Goals

As discussed above, there are various serious obstacles in building bridges between documentary linguistics, speech and language technology, digital language archives, and speaker communities. While all these communities are dealing with essentially the same type of language material, the needs and standards are different and largely incompatible. At the same time, it is obvious that bridges between these communities and services could benefit everybody and improve not only our understanding of the large variety of languages and cultures, and substantially facilitate numerous language documentation and revitalization projects, but also improve algorithms, and speech and language technologies in general.

In the initial concept of GORILLA we focused on the core technological goals a.) to provide an archival service and infrastructure based on free and open standards, that facilitates transcription, linguistic analysis, annotation and corpus creation. Its goals are b.) to make high quality language data freely accessible to documentary and theoretical linguists, speech and language technology researchers and developers, and in particular the interested speaker communities. Additionally, the infrastructure needed to be c.) connected to existing networks of linked and open

guages in the 19th edition of Ethnologue (Lewis et al., 2016). See for an introduction `https://www.ethnologue.com/ethnoblog/gary-simons/welcome-19th-edition`.

[10] See for example the overview of the different NIST evaluations on the Linguistic Data Consortium (LDC) website `https://www.ldc.upenn.edu/collaborations/evaluations/nist`.

[11] This growing speech and language technology gap, as well as limitations of software and information portal localizations, potentially have broader impacts that go beyond the pure linguistic one.

linguistic data like the infrastructure provided by the Linguistic Linked Open Data movement (LLOD)[12], the Open Language Archives Community (OLAC)[13] (Simons and Bird, 2003), the Open Archives Initiative (OAI)[14], and the CLARIN infrastructure[15] with the Virtual Language Observatory.[16] This linking can potentially bring together different resource types that improve research, documentation, technology development, and also revitalization efforts.

We designed GORILLA to provide the infrastructure to archive and deposit language data in form of audio and video recordings, dictionaries or corpora. In fact, it serves as an interface to the archiving infrastructure at the Archive of Traditional Music (ATM) at Indiana University, while the GORILLA infrastructure offers an additional layer on top of the archiving services for basically any kind of digital language data. The GORILLA team assists and provides technological means for the transformation of documentary language data to speech and language corpora that can serve the documentary, theoretical, and corpus linguistics communities, as well as speech and language technology groups.

The project aims at the aggregation and dissemination of language resources and models for all possible languages. Its focus is to enable resource creation, i.e. corpus development, speech and language technology training or engineering for low-resourced and endangered languages. Also, by working out and providing resources for specific languages, we expect to be able to derive in a much more efficient way resources for related variants or entire language groups with similar linguistic properties.

We address the licensing issues by providing one uniform and free access license to the resources on the GORILLA website. The corpus and technology services are provided for resources that are made available using the *Creative Commons Attribution-ShareAlike* (CC BY-SA) license (Creative Commons, 2016) or a freer version of it. This is to ensure their accessibility for derivative academic research as well as private or commercial products. While this license form provides free access to the data and resources, it also requires citations of the original work. The *share*-restriction potentially improves the resource, the annotations' depth or quality, and provides access to derivative technologies or tools for particular languages. In order to improve the situation of at least some low-resourced and endangered languages, we do encourage commercial use of the resources. This could potentially help speaker communities to provide an economic base for the production and marketing of educational language material, technologies or services.

The GORILLA infrastructure aims to be an open platform that provides standardized and linked metadata management for language resources and technologies, global linking to existing infrastructure and archiving networks, and standardized and interoperable data formats that maximize the usability for corpus linguistic research, as well as speech and language technology development.

GORILLA may serve yet another role, in particular, it might be instrumental in the promoting of best-practice recommendations for creating language resources, by providing examples and ready tools.

GORILLA aims at the development of potential solutions to the problems related to a growing gap between language documentation and linguistic work on the one side, and corpus linguistics and speech and natural language processing on the other. The data collected in documentation research is often not made available in form of structured, standardized, interoperable formats and annotations, i.e. the data is not prepared as a corpus and often archived in the available binary format only. The existing resources need to be pre-processed, converted, and brought into some compatible format to be able to exploit common speech and language processing technologies for automatic transcription or annotation. GORILLA provides a platform and environment to achieve this.

A research goal of GORILLA is to create an infrastructure that maximizes the quantities of corpus development processes and the annotation quality, while minimizing the time and effort invested in their production, i.e. a kind of "Ford Assembly Line" for corpus development and speech and language resources for low-resourced languages. Having language data and models available from related languages can potentially facilitate the development of such resources for other related and under-resourced languages. Working with an infrastructure where interoperable and compatible resource templates are used for corpora and models for speech and language technologies, can also significantly reduce the effort and increase the output of the language resource development processes.

As mentioned above, low-resourced and endangered languages are often of little economic interest for industry. These languages are most often not of strategic and political interest for governments that could sponsor academic development of resources and technologies. Investment in these languages is less likely to occur, thus those languages will most likely not participate in the ongoing progress in language technologies development. Another consequence is that the language resources are not available to the speaker communities that could use them for resource development (e.g. the generation of dictionaries, textbooks, or grammars). For the majority of languages there is no material that could facilitate the development of a standard or educational resources. Common linguistic aids like spell checkers or speech recognizers do not exist for those languages, because the fundamental language resources are missing. The creation of corpora that would be useful for the development of education material, or speech and language technologies has been extremely costly and time consuming.

Providing these resources in form of corpora via GORILLA potentially can improve the situation for many of those "economically challenged" languages or speaker communities, and in fact stimulate language-related economic developments in these speaker communities.

[12] See `http://linguistic-lod.org/` for more details.

[13] See `http://www.language-archives.org` for more details.

[14] See `http://www.openarchives.org` for more details.

[15] See `http://clarin.eu` for more details.

[16] See `https://vlo.clarin.eu` for more details.

5. Partners

To address the problems mentioned above we have partnered with the LINGUIST List and institutions at Indiana University including the Department of Linguistics, the Archive of Traditional Music. [17] (ATM), and the Institute of Digital Arts and Humanities [18]

The LINGUIST List has already established a web- and cyber-infrastructure that was facilitating the development of the GORILLA platform. The LINGUIST List data is linked to the Open Language Archives Community (OLAC) infrastructure. The GORILLA resources will be linked to this meta-data exchange stream. We are working on an integration in the Open Archive Infrastructure (OAI) and the CLARIN Virtual Language Observatory. We aim at linking the content and annotations to platforms like the Linguistic Linked Open Data Cloud, returning not only HTML, but also common RDF or JSON-LD from the stored digital objects, and linking content to vocabulary standards such as General Ontology for Linguistic Description (GOLD) (Cavar, 2016) [19] and/or ISOCat.[20]

In an attempt to link the resources and meta-data catalogs to the Linked Linguistic Open Data (LLOD) resources, we have partnered with colleagues from Europe who are involved in the LLOD movement.

6. GORILLA resources

GORILLA does not specialize in a particular language or language family. Instead, it serves as a repository for any free and open language resource. As mentioned above language models and resources for low-resourced languages are highly desirable for various reasons. In addition to speech corpora, the resources that are collected include various resources, such as:

- Video recordings of different spoken languages
- Parallel corpora
- Morpho-syntactically annotated corpora
- Treebanks
- Semantically annotated corpora
- Lexicons and dictionaries
- Formal and computational grammars
- Language models
- Resources for language technologies.

6.1. Data Standards

The resources in GORILLA are properly annotated in standardized frameworks (e.g. XML-based annotation standards or common open formats like Praat's TextGrid, RDF, TEI) and freely available online. As a result, the resources can be made readily available to be integrated in the development of computational tools and software. Developers are able to create various computational tools such as Forced Aligners, speech recognizers, or spell checkers for low-resourced languages without having to process documentary linguistic data.

Our infrastructure at LINGUIST List is already linked to the Open Language Archives Community (OLAC) infrastructure, and we are working on an integration in the Open Archive Infrastructure (OAI) and the CLARIN Virtual Language Observatory. We aim at linking the content and annotations to platforms like the Linguistic Linked Open Data Cloud (LLOD), returning not only HTML, but also common RDF or JSON-LD fro the stored digital objects, and linking content to vocabulary standards such as General Ontology for Linguistic Description (GOLD) [21] and/or ISOCat[22].

The common formats for transcription and annotation of audio and video recordings using the SIL FieldWorks Language Explorer (FLEx), ELAN, or Praat are supported.

6.2. Available Corpora

During summer 2015 we created initial speech corpora as pilot projects to estimate the average effort for transcription and annotation, and to experiment with different speech technologies for alignment and transcription. For all languages that we tested, our goal was to create initial corpora and language models that can be potentially used in the transcription and annotation process of larger resource collections. Thus, most corpora were not created using existing data from fieldwork or documentation projects.

Among the corpora that we created is for example the Chatino speech corpus (Cavar et al., 2016a), this volume. The Chatino speech corpus is the first available speech corpus for Chatino with full time alignment of the transcription, part-of-speech tagging, and translation. It was recorded and annotated with Hilaria Cruz, a native speaker and linguistic expert of the language. Since the language does not have any standardized orthography, only a phonetic transcription schema was used.

In addition to that we created a Croatian speech corpus, as well as a Yiddish one (Cavar et al., 2016b), this volume. We also recorded and transcribed partially or fully Spanish, Russian, and Burmese resources. Currently we are working on the development of further annotations using these corpora to generate corpora that are useful for NLP technologies, as well as linguistic research or language education.

For some of the languages we have created the first – to the best of our knowledge – existing speech corpora with Part-of-speech tags and translation (e.g. Chatino, Burmese, Croatian, Yiddish).

The initial transcriptions, time alignment, annotation and translation of the recordings have been done using ELAN (Sloetjes and Wittenburg, 2008). A subsequent detailed time alignment at the word level has been done using Praat (Boersma, 2001). We have created tools to process ELAN and Praat annotations, and, for example, generate a training corpus for Forced Aligners from ELAN annotation

[17] See http://www.indiana.edu/~libarchm/.

[18] See http://idah.indiana.edu/.

[19] See http://linguistics-ontology.org/.

[20] See http://www.isocat.org for a detailed overview.

[21] http://linguistics-ontology.org/

[22] http://www.isocat.org/

files (see for example ELAN2Split[23]). These resources are freely available on the website of GORILLA or by contacting us.

The speech corpora on GORILLA have been generated using the following strategies. In order to avoid licensing problems related to the content, we have used freely available and CC BY-SA licensed text where available. For completely under-resourced languages like Chatino we have used text created by the native speaker who was involved in recording spoken language samples. The native speakers read the text to create initial audio recordings. We transcribed and time aligned the recordings using a copy-and-paste method given the pre-existing text. With the time-aligned data of at least 2 hours per language we were able to train for example the Prosody Lab Aligner[24] or any other HTK-based Forced Aligner.[25]

An initial Yiddish speech corpus has been developed in order to facilitate the transcription of Yiddish testimonials in the AHEYM collection at Indiana University. The original recordings in the AHEYM collection (Kerler, 2014), (Cavar et al., 2016b) are only partially useful to serve as training corpora for speech technologies due to limitations of the recordings done in the field, as described above. Thus we used a small sub-set of this material together with various freely available audio-books read in Yiddish. Efforts to apply Optical Character Recognition (OCR) tools to the scanned books failed due to specific properties of the Hebrew script as it is used for Yiddish. Although the texts of the audio books were only available in form of scanned images of original books, we were able to use the image representation of the text to transcribe at least 5 hours of spoken Yiddish from these sources manually.

For Burmese we used freely available text-sources to create freely licensed content. A Burmese native speaker recorded several Wikipedia articles and time-aligned the recordings into sentences using ELAN. The ELAN transcription was reduced to a copy-and-paste process and manual time alignment on the utterance level. The detailed time-alignment on the word level is done in Praat, which allows for a fine-grained alignment using the spectrogram visualization.

For some of the languages we used the Aligner that is implemented in Praat. The Praat-based aligner makes use of a so called analysis by synthesis method. It expects an audio recording and the corresponding transcription text as input. It uses a text to speech synthesis engine, i.e. Espeak[26], to generate an audio representation of the text using a specific language model (i.e. a phone inventory and specific sound-mappings). This generated audio is mapped on the recorded speech and hypotheses about time alignment are generated. This method has two advantages. On the one hand, it makes sense to use the alignment features in the annotation tool directly to simplify the corpus creation process. On the other hand, the development of a language model for a particular language using Espeak includes the specification of the phone inventory of a language, specific pronunciation regularities and lexical exceptions. We consider this a valuable language documentation contribution as such. In fact, the same is true for HTK-based Forced Aligners. They also require acoustic and language models, i.e. a model of the languages phone inventory and phonotactic regularities, and a basic grammar or model of lexical distribution patterns.

Currently we have basic Espeak models for Burmese or Yiddish for example. The corpora of all the recorded languages are large enough, i.e. more than 2 hours of time aligned speech data, to train a Forced Aligner, and for some languages, e.g. Croatian we have more than 10 hours of transcribed speech to be able to train first Automatic Speech Recognizers (ASR).

In the process of exploring common technologies for tokenization and morphological analysis we also evaluated the processes to generate models for under-resourced languages that would enable us to increase the quantity of corpora. Thus, in addition to the audio material and speech corpora, we developed Finite State Tools using the Foma toolkit and libraries (Hulden, 2009). Foma provides a platform for engineering of morphologies and tokenizers for natural languages, using a two-level morphology based approach. We created for example a tokenizer for Burmese, a Tibeto-Burman language, written in *abugida* script with no spaces between words, and a basic morphological analyzer for Croatian.

We created initial resources and models for Forced Alignment, basic acoustic and language models for speech recognition systems, and other NLP-components. These resources and the corresponding technologies will enable us to facilitate the annotation process of resources from these languages and catalyze their expansion. These resources should show how language resources can be transformed to benefit language documentation, linguistic research, and speech and language technologies.

All data, language models and technologies are being assigned meta-data using current standards and resource location technologies, e.g. CMDI and OLAC for meta-data and DOI for digital resource identifiers.

7. Conclusion

GORILLA will motivate researchers from various disciplines to contribute and share language resources by lowering the barrier for resource collection and development. It hopefully will create a best practice platform with a collection of basic material and technologies that can be used in documentation and technology related research. We hope to be able to attract speaker communities to collaborate with us, and contribute or actively make use of the resources that GORILLA provides.

Therefore, the potential impacts of GORILLA might be felt mostly on low-resourced languages by providing linguistic models, corpora, computer software and tools to support these languages. Further, the resources are distributed using free and open standards; they are not in any proprietary format and thus will still be available as snapshots in time of some languages to the future generations of researchers.

[23] See `https://bitbucket.org/dcavar/elan2split` for more details.

[24] see for example `http://prosodylab.org/tools/aligner`.

[25] See for the Hidden Markov Model Toolkit (HTK) `http://htk.eng.cam.ac.uk` and Young et al. (2006).

[26] See `http://espeak.sourceforge.net` for details.

Among the goals of the GORILLA project is to foster a paradigm shift with respect to the way linguistic data are shared within the interdisciplinary community of linguists, anthropologists, computer scientists, and others. For most of the existing resources it might not be possible to establish a CC BY-SA based license and distribute them freely. However, smaller samples of data collections can be prepared and made usable under such a license. Where this is not possible, we try to acquire enough new resources under conditions that allow for publication under the CC BY-SA license and work on speech and language technologies that can be used for non-free resources in digital language archives.

This project is still in an early phase. We encourage the community to provide us with feedback, suggestions and comments related to the ways that we can improve the platform.

8. Acknowledgments

This project emerged as a spin-off from the NSF-funded AARDVARC project (NSF number 1244713). We are grateful to a large group of colleagues who participated in workshops and conferences organized by the AARDVARC team. We are most grateful for ideas, suggestions, help, and comments from Helen Aristar-Dry, Anthony Aristar, and Douglas H. Whalen, the main team members and organizers of AARDVARC.

The director of the Archive of Traditional Music (ATM) at Indiana University, Alan Burdette, agreed to co-operate with the GORILLA project and provide the necessary archiving infrastructure. He established the connections with different other divisions and services at Indiana University that allow GORILLA to provide a professional service to the community.

The GORILLA project is also endorsed and supported by the Department of Linguistics at Indiana University.

We are grateful to the interns and visiting students at LINGUIST List in the summer of 2015 for their help with corpora and the initial definition of the GORILLA infrastructure. Many graduate students from the Indiana University Department of Linguistics have participated and contributed to the resources. We are in particular grateful to Aaron Albin, Seyed Asghari, Edvard Bigbaev, Zac Branson, Petar Garžina, Clara García Gómez, Levi King, and Alec Wolyniec for their help and support. Michael Abramov helped us transcribe and time align Chatino in ELAN. Anya Quilitsch and Avi Lang invested a lot of time in transcription and annotation of the Yiddish speech corpus, and Dov-Ber Kerler provided us with the material and valuable suggestions. Raphael Finkel from the University of Kentucky provided us with a Yiddish dictionary and phonetic transcriptions, as well as tools for the transfer of Hebrew script Yiddish to YIVO romanization. All editors and graduate assistants of LINGUIST List were maximally helpful with suggestions, comments, contributions: Sara Couture, Andrew Lamont, and Anna White.

9. Bibliographical References

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Cavar, D., Cavar, M., and Cruz, H. (2016a). Endangered language documentation: Bootstrapping a Chatino Speech Corpus, Forced Aligner, ASR. In *Proceedings of LREC 2016*. ELRA.

Cavar, D., Cavar, M., Kerler, D.-B., and Quilitsch, A. (2016b). Generating a Yiddish speech corpus, forced aligner and basic ASR system for the AHEYM project. In *Proceedings of LREC 2016*. ELRA.

Cavar, D. (2016). General Ontology for Linguistic Description (GOLD). http://linguistics-ontology.org/.

Creative Commons. (2016). Creative commons attribution-sharealike 4.0 international. https://creativecommons.org/licenses/by-sa/4.0/, Mar.

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for? In J. Gippert, et al., editors, *Essentials of Language Documentation*, pages 1–30. Mouton de Gruyter, Berlin.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics.

Kerler, D.-B. (2014). Surviving remnants of yiddish folksinging and creativity in contemporary ukraine [in yiddish]. *Afn Shvel*.

Lehmann, C. (2001). Language documentation: A program. In Walter Bisang, editor, *Aspects of Typology and Universals*, pages 83–97. Academie Verlag, Berlin.

Lewis, M. P., Simons, G. F., and Fennig, C. D. (2016). Ethnologue: Languages of the world. Online version.

Newman, P. (2013). The law of unintended consequences: How the endangered languages movement undermines field linguistics as a scientific enterprise. Paper presented at the Linguistics Departmental Seminar Series, SOAS, University of London.

Simons, G. and Bird, S. (2003). The open language archives community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18(2):117–128.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category: ELAN and ISO DCR. In Nicoletta Calzolari, et al., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (2006). *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.