LREC 2018 Workshop

CCURL2018
Sustaining Knowledge Diversity in the Digital Age

PROCEEDINGS

Edited by

Claudia Soria, Laurent Besacier, Laurette Pretorius

ISBN: 979-10-95546-22-1
EAN: 9791095546221

12 May 2018

Proceedings of the LREC 2018 Workshop
"CCURL2018 – Sustaining Knowledge Diversity in the Digital Age"

12 May 2018 – Miyazaki, Japan

Edited by Claudia Soria, Laurent Besacier, Laurette Pretorius

http://www.ilc.cnr.it/ccurl2018/index.htm

Acknowledgments: This workshop is endorsed by SIGUL, the ELRA-ISCA Special Interest Group on Under-resourced Languages.

Organising Committee

- Laurent Besacier, LIG-IMAG, France*
- Khalid Choukri, ELRA/ELDA, France
- Joseph Mariani, LIMSI-CNRS, France
- Laurette Pretorius, University of South Africa, South Africa*
- Sakriani Sakti, NAIST, Japan
- Claudia Soria, CNR-ILC, Italy*

*: Main editors and chairs of the Organising Committee

Programme Committee

- Tunde Adegbola, African Languages Technology Initiative, Nigeria
- Gilles Adda, LIMSI-CNRS, France
- Shyam Agrawal, KIIT Group of Colleges, India
- Amir Aharoni, Wikimedia Foundation
- Antti Arppe, University of Alberta, Canada
- Victoria Arranz, ELRA/ELDA, France
- Martin Benjamin, the Kamusi Project, Switzerland
- Laurent Besacier, LIG-IMAG, France
- Steven Bird, Charles Darwin University, Australia
- Luong Chi-Mai, IOIT, Vietnam
- Khalid Choukri, ELRA/ELDA, France
- Chris Cieri, LDC, USA
- Thierry Declerck, DFKI, Germany
- Sebastian Drude, The Vigdís International Centre for Multilingualism and Intercultural Understanding, Iceland
- Vera Ferreira, CIDLeS - Interdisciplinary Centre for Social and Language Documentation, Portugal
- Mikel Forcada, Universitat d'Alacant, Spain
- Dafydd Gibbon, Bielefeld University, Germany
- Tatjana Gornostaja, Tilde, Latvia
- John Judge, ADAPT DCU, Ireland
- András Kornai, Hungarian Academy of Sciences, Hungary
- Joseph Mariani, LIMSI-CNRS, France
- Yohei Murakami, Kyoto University, Japan
- Satoshi Nakamura, NARA Institute of Science and Technology, Japan

- Girish Nath Jha, JNU, India
- Guy de Pauw, Textgain, Belgium
- Laurette Pretorius, University of South Africa, South Africa
- Sakriani Sakti, NAIST, Japan
- Kevin Scannell, Saint Louis University, Missouri, USA
- Claudia Soria, CNR-ILC, Italy
- Oliver Stegen, SIL International, USA
- Francis Tyers, Moscow Higher School of Economics, Russia
- Trond Trosterud, Tromsø University, Norway
- Kadri Vider, University of Tartu, Estonia
- Eveline Wandl-Vogt, Austrian Academy of Sciences, Austria

Preface

The diversity of languages and cultures is a distinctive footprint of the way humans have been interacting with the environment over time; unique visions of the world, as well as unique knowledge about the environment resides with indigenous cultures and languages. Now, preservation and sharing of the traditional knowledge encoded by languages is being increasingly recognised as an important step towards accumulating valuable knowledge that can help a sustainable and durable interaction of mankind with the environment. However, as language diversity is decreasing, the maintenance and transmission of such knowledge is at risk. Unfortunately, the vast majority of this knowledge is poorly represented in digital form. If we take the case of Wikipedia as an example, very few indigenous languages are available on the platform (according to a recent report, only four out of the 522 indigenous languages of Latin America are represented by Wikipedia projects).

Digital language resources can play a crucial role to avoid the disappearance of the diverse knowledge systems, to ensure their preservation and transmission, and to foster their cross-fertilisation with other knowledge systems. The digital representation of traditional knowledge, however, on the one hand shares many of the issues that concern under-resourcedness: as this knowledge lies with under-resourced (minority, endangered or minoritised) languages, specific methods and models of resource development are required to circumvent the problems affecting low-resourced languages, such as low investments, data sparsity, fragmentation of efforts, lack of involvement of the speaker communities, to cite just a few. On the other hand, specific problems arise: most notably, the issue of the extent to which this knowledge is shareable with community outsiders, community ownership and control over content, which brings about the need to involve community representatives from the very beginning of the resource creation process.

This workshop aims at gathering together academics, industrial researchers, knowledge experts, digital language resource and technology providers, software developers, but also language activists and community representatives in order to identify the current capacity and the difficulties in creating and sustaining the digital representation of traditional knowledge. In particular, the focus of the workshop intends to be on: current initiatives on documenting, as well as presenting experiences related to traditional knowledge representation analysing the request for such knowledge resources, their impact and potential use identifying best practices, lessons learned as well as sensitive issues related to traditional knowledge management and representation.

C. Soria, L. Besacier, L. Pretorius May 2018

Programme

Opening Session

09.15 – 09.30 Welcome and Introduction

09.30 – 10.30 Keynote Talk – Laurent Besacier
Computational language documentation: some results from the BULB project

Morning Session

11.00 – 11.30 Mat Bettinson and Steven Bird
Image-Gesture-Voice: a Web Component for Eliciting Speech

11.30 – 11.50 David Nathan
Beyond Protocol: Indigenous Knowledge Resource Circulation in the Digital Age

11.50 – 12.20 Martin Benjamin
Hard Numbers: Language Exclusion in Computational Linguistics and Natural Language Processing

12.20 – 12.40 Amelie Dorn, Eveline Wandl-Vogt, Yalemisew Abgaz, Alejandro Benito Santos and Roberto Therón
Unlocking Cultural Conceptualisation in Indigenous Language Resources: Collaborative Computing Methodologies

14.00 – 15.00 **Poster Session**

Aalok Sathe
A Rule-Based System for the Transcription of Sanskrit from the Devanagari Orthography to the International Phonetic Alphabet

Aidan Pine and Mark Turin
Seeing the Heiltsuk Orthography from Font Encoding through to Unicode: a Case Study Using Convertextract

Mathias Coeckelbergs
Classifying and Searching Resource-Poor Languages more Efficiently. Using the FastText Word Embeddings for the Aramaic Language Family

Dmitri Dmitriev
Digitizing National Cuisines: Cooking Recipes as Conceptual Graphs

Amel Fraisse, Ronald Jenn and Shelley Fisher Fishkin
Building Multilingual Parallel Corpora for Under-Resourced Languages Using Translated Fictional Texts

Shweta Sinha and Shyam S Agrawal
Sustaining Linguistic Diversity through Human Language Technology: a Case Study for Hindi

Afternoon Session 1

15.00 – 15.30 Dorothee Beermann, Lars Hellan and Tormod Haugland
Convergent Development of Digital Resources for West African Languages

15.30 – 16.00 Delyth Prys and Dewi Jones

Gathering Data for Speech Technology in the Welsh Language: a Case Study

Afternoon Session 2

16.30 – 16.50 Lászó Grad-Gyenge and Linda Andersson
The MediaBubble Dataset: a Crowdsourcing Dataset for Topic Detection Tasks for the Hungarian Language

16.50 – 17.10 Adrian Doyle, John McCrae and Clodagh Downey
Preservation of Original Orthography in the Construction of an Old Irish Corpus

17.10 – 17.30 Alice Millour and Karën Fort
Krik: First Steps into Crowdsourcing POS tags for Kréyòl Gwadloupéyen

17.30 – 18.00 Discussion and Conclusions

Table of Contents

Image-Gesture-Voice: A Web Component for Eliciting Speech

Mat Bettinson[1] **and Steven Bird**[2,3]

[1]Department of Linguistics, University of Melbourne
[2]Northern Institute, Charles Darwin University
[3]International Computer Science Institute, University of California, Berkeley

Abstract

We describe a reusable Web component for capturing talk about images. A speaker is prompted with a series of images and talks about each one while adding gestures. Others can watch the audio-visual slideshow, and navigate forwards and backwards by swiping on the images. The component supports phrase-aligned respeaking, translation, and commentary. This work extends the method of Basic Oral Language Documentation by prompting speakers with images and capturing their gestures. We show how the component is deployed in a mobile app for collecting and sharing know-how which was developed in consultation with indigenous groups in Taiwan and Australia. We focus on food preparation practices since this is an area where people are motivated to preserve and disseminate their cultural and linguistic heritage.

Keywords: language documentation, procedural discourse, mobile apps, crowdsourcing, web technologies

1. Introduction

The program of language documentation is one response to the rapid decline in linguistic diversity (Himmelmann, 1998; Woodbury, 2010). We are challenged to design scalable methods for documenting thousands of languages while there is still time. More generally, the wider problem space is the search for digital knowledge preservation at scale.

One promising avenue is in seeking collaboration with the wider audience speakers, or *crowdsourcing*. Crowdsourcing in other domains, from Google Maps to Wikipedia, is an established pattern for collective intelligence. These patterns are seen in countless user-contributed content apps. The architecture of crowsourcing apps is no different from any (Chatzimilioudis et al., 2012). However there has been little attention towards minority community use cases.

Practical crowdsourcing depends on a confluence of interests. App developers may seek collection and preservation of knowledge, while the target audience are seeking solutions to their problems. This calls for a process of exploration, negotiation and user-centered design. In 2016, a series of app design workshops explored ways to include linguists, technologists and speech community members (Bird, 2018). One of the emerging designs was an app for capturing talk about food preparation (Mettouchi et al., 2017; Bettinson and Bird, 2017; Bettinson, 2017).

This paper describes an approach to documenting procedural knowledge, or any kind of know-how. We report on the Android mobile app Zahwa, which we have tested with speakers of endangered languages in Taiwan and Australia. We also seek to bootstrap the creation of similar knowledge preservation apps in the future. To this end we discuss ongoing efforts to develop a library of reusable software components based on the emerging Web Component standard.

This paper is organised as follows. In Section 2. we discuss previous work. Then Section 3. proposes a new method for documenting procedural knowledge. The Zahwa app implementation is described in Section 6. We discuss recent work on Web Components for knowledge preservation in Section 6.

2. Previous work

There are very few apps specifically intended to let app users document and share their know-how. It's much more common to find apps that serve as a vehicle to publish content compiled by experts, such as dictionaries, or language teaching apps. Digital knowledge preservation lags the *Web 2.0* trend towards user-contributed and socially contextualised participation.

Crowdsourcing acoustic data is one established genre of mobile app. *Voice App* collects regional speech data for Swiss-German to study dialectal variation (Goldman et al., 2014). *English Dialects App* includes a dialectal prediction feature a form of 'gamification' to encourage people to use the app (Leemann et al., 2016; Leemann et al., 2018).

Dictionaries are a popular genre of mobile app. While many have been made for endangered languages, the majority don't allow user-contributions. The *Ma! Iwadja* app incorporated a "crowdsourcing lexicon development system" but the app is no longer functional. We have found it sadly quite common that language apps vanish without a trace. This serves as a reminder of the ongoing challenge of sustainable development.

The *Aikuma* mobile app is mobile app capable of crowdsourcing natural language (Bird et al., 2014). *Aikuma-LIG* is a further development aimed at collecting data speech processing (Blachon et al., 2016). While capable of crowdsourcing, these apps are intended for researcher-driven use-cases. They are not designed with general audiences in mind and don't readily support sharing with other users.

The web platform is a viable choice for building the next generation of digital knowledge preservation tools (Bettinson and Bird, 2017). The acoustic waveform display library Wavesurfer[1] is an example of a successful open source software component in the web domain. The Web Component[2] (WC) standard is now supported or 'pollyfilled' for all major web browsers. There's growing momentum behind a library of library of freely available WC components[3].

[1] https://wavesurfer-js.org/
[2] https://developer.mozilla.org/en-US/docs/Web/Web_Components
[3] https://www.webcomponents.org/

Commercial apps live or die by their ability to engage users. The ability to share content via existing social networks is ubiquitous across popular mobile apps. App users are motivated to share for a number of reasons including reciprocity, social engagement and reputation building (Oh and Syn, 2015). Aside from fulfilling user expectation, sharing also serves a helpful marketing function akin to *viral marketing* (Subramani and Rajagopalan, 2003). Finally, sharing meets the need of disseminating cultural knowledge. Digital dissemination is particularly helpful for maintaining indigenous knowledge (Chikonzo, 2006).

3. Documenting procedural knowledge

Procedural discourse is defined as "an explanation or description of a method, process, or situation having ordered steps. Examples of procedural discourses include recipes, instructions, and plans" (Johnson and Aristar Dry, 2002). In the context of endangered languages, traditional practices and rituals are falling out of use. The practices include how to prepare medicinal remedies, how to build a canoe, how to grow yams, the stages of initiation, and so forth. Of particular significance are the traditions concerning food preparation, since these connect with the local environment, the seasons, the calendar of rituals, family ties, and cultural identity. Documenting procedural knowledge is a recognised part of language documentation (Pollock, 2011).

Many speakers of endangered languages are mindful that the younger generation are not learning traditional crafts, rituals, methods for food preparation, and so on. For instance, by the time a girl grows up and has a family of her own, it may be too late to ask her grandmother how to prepare the dish that marks a particular rite of passage. Similarly, geographical separation between generations can make it difficult to transmit knowledge.

In literate societies one could email the instructions. Another approach would be to record a video of the procedure, where it may be later shared on YouTube. However, video recording demands either a single fluid performance or a video editing process in order to achieve a succinct result. It is difficult to self-record video. When multiple participants are involved, it becomes more of a performance and needs to be planned. Later, in translation, it is harder to replace the audio track and keep everything synchronised.

The mobile phone platform provides the ability to take photos to use as prompts, in a way that is similar to the use of stimulus in non-digital documentary methods (Lüpke, 2010, p.58). Further more, mobile phones offer a tactile user interface that offers the additional benefit of capturing gesture at the same time as spoken voice. In the following section we discuss a design process of an app that incorporates a new method for documenting procedural knowledge.

4. App design

Attention to the design process is important where our goal is to create apps that people are self-motivated to use. We are particularly at risk where app development and app deployments take place in different cultures. In this section we illustrate the design process from an app workshop we organised in Darwin in September 2016.

Following the basic principles of user-centered design, we describe the "personas" involved in the app (Norman and Draper, 1986; Bird, 2018). We make these as real as possible, inventing names, and discussing motivations.

Taos left the village of her childhood when she was 8. Now she is 23 and lives in Algiers with her family and studies at the university. When she is in the dorms she cooks meals with other young women. One time while preparing food they were talking about their childhoods and Taos spoke about her holidays in her village and her grandmother's delicious cooking. Because they had similar experiences, the young women decided to collect and share these traditional recipes. The next holidays, Taos is visiting her grandmother, and decides to photograph the stages of preparing each recipe.

Zahwa has always lived in the village, and has learnt to cook with her own mother and grandmother. She enjoys cooking, and is very happy when her granddaughter comes to visit her from Algiers and brings her news of the capital and her life at university. Zahwa doesnt know how read or write, so when Taos asks her about her recipes, she tells her that the best way to learn is to watch and do it with her. But then Taos says it would be better if she recorded her grandmother making the recipe, so that she would be able to do it herself, and that she would be proud to share it with her friends and maybe other people too.

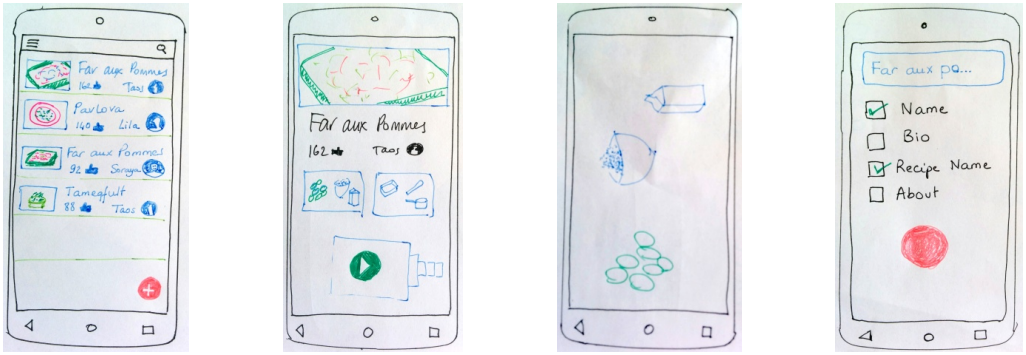
Next, we write out the value proposition, i.e. how would our proposed solution connect with the jobs, pains, and gains experienced by one of the personas (Bird, 2018). In this case, we take the perspective of Taos, as the one who drives the creation of the content (see Figure 1).

Finally, we express the design in terms of a series of app screens (see Figure 2). When the app is opened we see a list of popular recipes (Fig 2(a)) Perhaps we can follow other people, or see recent changes, etc. For each recipe we can see the number of "likes". There's a + button that lets us add a new recipe. We can open an individual recipe to see a larger image (Fig 2(b)). There are pictures which show the ingredients and utensils, so you can quickly see if you have everything you need in order to make this recipe. If we're interested we can press play to playback the recording. During playback the app goes full-screen in landscape mode, and we can touch the screen to pause or resume (Fig 2(c)). We see a series of images. We can swipe the screen to navigate to a different image and resume playback from that point. When we want to create a new recipe, we are prompted to enter a title (Fig 2(d)). Then we do a summary recording including name of the speaker, short bio, name of the recipe, something about it e.g. when it is cooked, for what ceremony, or in what season.

In field-testing early designs in rural Taiwan, we noticed that app users would touch the screen while talking. This motivated us to add a feature whereby we opportunistically capturing touch-screen gestures during recording. In the following section we describe the Image-Gesture-Voice method incorporating gesture.

Jobs: What does this persona want to do?	*Pains: What challenges does this persona face?*	*Gains: How might the app help?*
• cook for friends in traditional style • reconnect with traditional culture • connect with grandmother • document family culinary tradition so she can pass it on	• doesnt know how to make grandmothers recipe • doesnt have way to share grandmothers recipe with friends that captures whole process • cannot carry around all the recipes of her family's heritage • cannot easily compare ingredients, different methods	• provides a template for documenting grandmothers recipes including photos • allows capturing alternate versions of same recipe for comparing • provides a series of cherished recipes, told in the voice of an older relative • makes it possible to compare her traditional recipe with that of friends families

Figure 1: Value Proposition Table, enumerating jobs, pains, and gains from the standpoint of a persona (here, Taos), to better understand why someone would want to use an app



(a) Home: popular recipes.

(b) Recipe summary view.

(c) Playback: audio slideshow.

(d) Audio metadata.

Figure 2: Initial app design involving pen-and-paper drawings for four screens.

5. Image-Gesture-Voice

We propose a new Image-Gesture-Voice (IGV) documentary method, extending Basic Oral Language Documentation (BOLD) with images and touch-gestures. The method was motivated by the use case of documenting procedural discourse by recording spoken language linked to a series of still image prompts. IGV consists of separate recording and annotation activities. We first describe the recording activity via the use case of documenting a cooking recipe.

Before we begin recording, we obtain a series of still images. Photos can be taken before, during and after the food preparation activity. We might begin by photographing the ingredients and cooking utensils, then take photos of each step of preparation, and finally an image of the completed dish. A benefit of this approach is that it removes the pressure of spoken performance from the procedure.

The IGV record activity displays a slideshow of images. Recording can be initiated, paused, and resumed. When recording, only forward navigation to the next image is possible. When paused, the user can navigate forwards and backwards between images. When beginning or resuming a recording, the slideshow will seek to the first image that has not been discussed. The recording is completed when the user records to the final image.

When recording, the user's gestures on the current image are captured, and visual feedback is given. The same visual effect is used during playback. Our implementation uses a 'particle effect' which serves to reducee the precision of a gesture, and to enhance the sense of region, and of motion.[4] Capturing voice and gesture inputs simultaneously enables a dual expressive modality. Speech and gesture are semiotic resources that speakers can coordinate (Kendon, 2004; Kendon, 2008), and which have been found to boost precision of referencing in user interfaces (Bolt, 1980). For example, a user might gesture over a bowl of ingredients in a circular motion while describing the mixing action. This data may increase the precision of linking appropriate 'mixing' verbs with the audio signal. Aside from referentiality, the gesture gives additional information such as the speed and direction of mixing.

The basic layout of the IGV differs for landscape and portrait view ports. In landscape, we recommend showing three images, previous, current and next, with the previous and next images only being partially displayed. In portrait view, the current image is best displayed full width, accompanied by a three-image row of previous, current and next images beneath the main image. The previous and next images are important for navigation, including selecting the next image while recording.

After accepting a completed recording, IGV begins review playback. During playback the slides will advance automatically, and recorded gestures will be displayed on screen. If the user selects slides during playback, the audio will seek to the appropriate point. Finally, the user has the

[4]We have made available a generic web-based gesture recording and visualisation library called Gestate which was designed specifically for this purpose

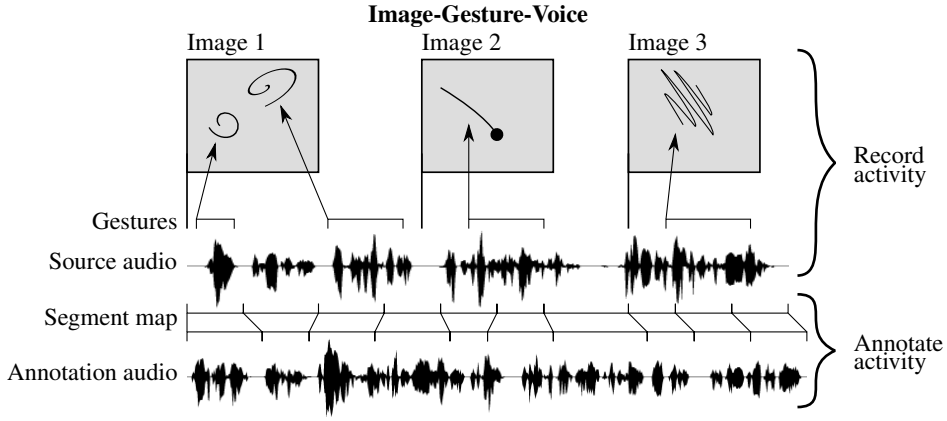


Figure 3: The IGV record activity aligns image prompts and touch gestures to the source audio signal. The annotate activity results in a segment map of source audio segments to annotation audio segments.

option to re-do (clear) or accept the recording.

The second IGV activity is for producing phrase-by-phrase oral annotations (Bird, 2010; Reiman, 2010). A common use case is performing a respeaking or a spoken translation. The activity is accomplished by a series of alternating play and record actions. During source playback, the prompt image will change automatically and gestures are displayed with the same visual effects as seen in the IGV record activity. The resulting data consists of an audio file and a segment map which associates source audio playback spans with recorded audio spans, see Figure 3.

Our mobile implementations have utilised separate playback (left) and record (right) controls. The operator plays by pressing and holding the play button, pauses by lifting off, replays by re-pressing play. Similarly, recording is a press and hold action, which also allows for momentary pausing. Play and record are alternated until all of the source audio has been played.

We may wish to play back the respeaking or translation with time-aligned images and gestures. We are able to display images at the correct time thanks to the segment map from the IGV annotation activity. However gestures occur within the image spans, and the operator may refer to things in a different order, especially when translating into a language that has a different word order. The only true way to ensure good gesture alignment for audio annotations is to recapture gestures during the annotation activity, which goes beyond our current implementation.

In the following sections we describe two implementations of IGV. Section 6. presents Zahwa, an Android app which implements both IGV activities. Section 7. discusses our recent efforts to implement IGV as open source Web Components.

6. The Zahwa App

This section presents a production quality Android app, called Zahwa, along with discussion of Web technologies and hybrid mobile app development. Zahwa has been co-designed and field tested with speakers of the endangered Austronesian language Saisiyat in rural North Western Taiwan. The app has also been field tested with indigenous Australian communities in Far North Australia. The app, and project web site, is offered in English and Traditional Chinese to support these fieldwork projects.

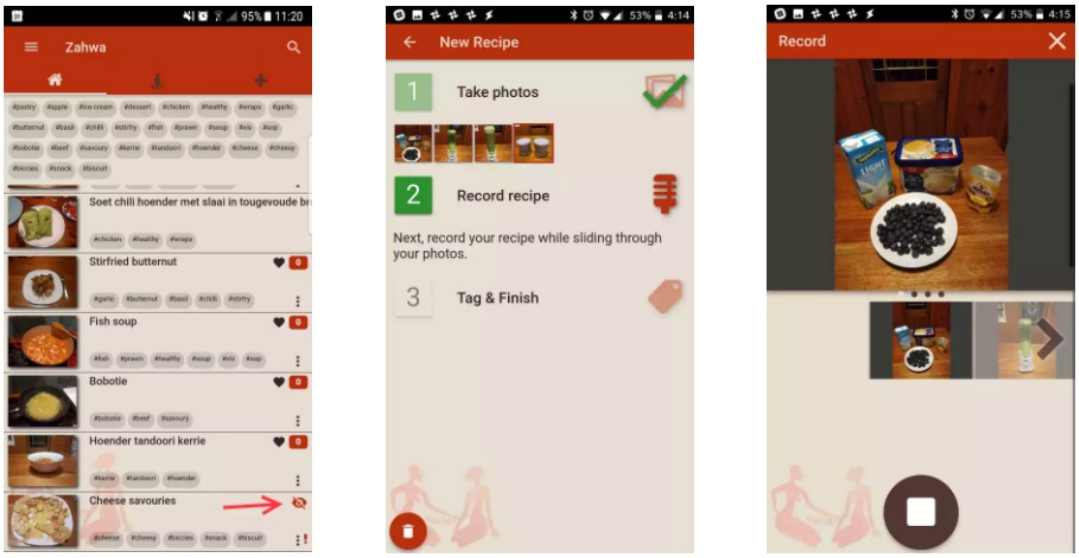
On first run, a new Zahwa user authenticates using their phone's existing Google or Facebook account. They are then shown a user agreement and given the opportunity to create a user display name. A pop-out menu provides access to features such as editing the user's profile (setting languages, taking a new photo), app settings, and data back-up/restore.

Zahwa's recipe record process is an implementation of the IGV method described in Section 5.. Creating a recipe is a three-step process; importing images, recording audio, and finalizing with descriptive metadata (see Figure 4(b)). The process of importing images requires selecting photos from the phone's gallery and putting them in sequence. The second step prompts for the language then proceeds to an IGV record activity, shown in Figure 4(c)). A finalization tasks has the user provide typed or spoken (speech-to-text) metadata including the recipe name, optional description and series of tags.

The Zahwa app is structured on three views accessed via UI 'tabs'. The rightmost *kitchen* view is a workflow management view where incomplete recipes appear. Creating a recipe is a multi-stage process and we handle these as stateful asynchronous tasks that users return to when convenient. The role of the *kitchen* is to guide the authoring of recipes through to a *minimally complete* form suitable for social interaction.

Completed recipes are moved from the kitchen view to the *social* view on the leftmost default tab (see Figure 4(a)). Here the user's recipes appear alongside downloaded recipes from other users. Recipes displayed in this view are presented with UI measures which draw attention to further actions we encourage users to perform, such as adding metadata including tags, and performing oral translations.

The *social* view also encourages the user to publish the recipe, so it can be found by others via in-app searches or



(a) Summary Screen: map the location of participants

(b) After importing photos...

(c) Recording a recipe...

Figure 4: The Zahwa app

the via the project website.[5] This view also allows users to share recipes,by launching the phone's native sharing activity, allowing the user to share via any installed app on the phone, including email, Facebook, text message and so on. The sharing scheme is based on a unique URL which also specifies the chosen translation for playback. The URL launches a mobile web app which fetches the recipe assets and plays the recipe in a desktop browser or on a mobile phone. Unpublished recipes can be shared by this unique URL, but are otherwise not discoverable (a feature inspired by YouTube).

An optional activity accessed via the social view is producing oral annotations, see Figure 5. Translations can be performed by other app users, such as bilingual speakers who want to make recipes more widely accessible.

The *map* view (middle tab) displays a geographical map with nearby recipes appearing as pins. Search is possible from any tab via the top right search icon. Zahwa adheres to a principle of being meaningfully usable when offline, including the ability to record recipes, and to search for recipes. The app caches nearby recipe metadata including thumbnail images. Users can find recipes by metadata and mark recipes for download when they have an internet connection. Similarly, the users actions such as publishing recipes are deferred until a network becomes available.

In field testing earlier prototypes we noticed that app users somewhat reluctantly re-oriented a phone to landscape if a particular view required it. This motivated us to support both portrait and landscape use. The IGV implementations adjust to the different aspect ratios by reorganising the position of buttons, and using different slide layout schemes as discussed in Section 5. (see also (Bettinson and Bird, 2017)). Considering that vision-impaired elders are particular candidates for apps of this type, running the app on an inexpensive Android tablet in landscape mode results in a suitably large image, about the size of a standard photograph.

6.1. The Web Technology Stack

The web technology stack allows us to deploy a common software component model across all platforms including Android, iPhone and the web. By web technologies, we refer to a single-page web app delivering rendered by a web browser engine. Zahwa is the latest product of a sustained program of research into web technologies that began with Aikuma-NG (Bettinson and Bird, 2017) in 2016.

Web technologies have been advancing at a dizzying pace. In the last few years there have been major advances in web browser APIs, JavaScript language specification[6], and the vibrant ecosystem of JavaScript 'frameworks'. Just two years ago, JavaScript frameworks suffered from significant performance issues and there was a lack of robust tooling in support of larger software projects. The situation today is much improved. Technology giants such as Google have done much to improve the capabilities of web browser APIs, and by extension the user experience of web apps.

The vibrancy of the JavaScript framework ecosystem is both a strength and a weakness. The ecosystem has arisen to support the industry of web app designers which tend to be concerned with short term projects. Remaining abreast of web technologies is a significant burden oft described by web developers as 'JavaScript fatigue'. There is a risk that the burden could exceeds the effort of developing two different apps via the Android and iOS native SDKs (or three, including the web).

We have minimized the web stack maintenance cost by

[5] http://zahwa.aikuma.org (includes a detailed how-to section)

[6] More accurately ECMAscript, as the standard is known

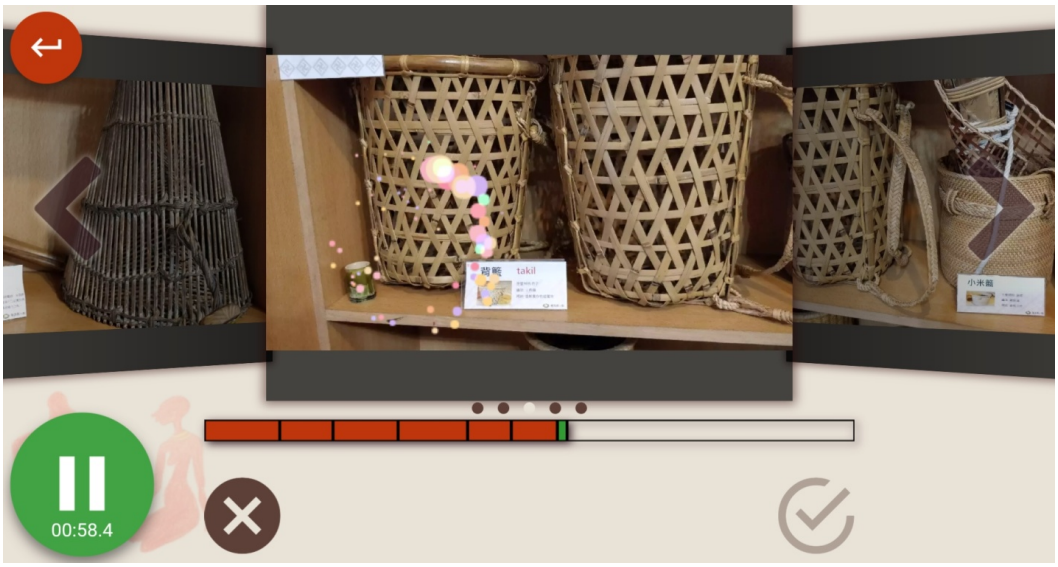


Figure 5: Zahwa's translation activity, e.g. IGV annotation. The source is a Saisiyat elder describing traditional crafts. Touch gesture particle effects are visible around the back basket (takil) currently being described. This activity was used to translate into Chinese Mandarin for the Taiwanese audience.

adopting the Ionic framework,a complete solution for deploying web and hybrid mobile applications. Ionic is based on the Angular JavaScript framework, Typescript, and a library of components that replicate common native UI elements for Android and iOS. Ionic includes 'tooling' for a complex web app, and integrates Apache Cordova to build installable hybrid mobile apps discussed in the next section. Frameworks are usually an architectural choice that applies app-wide. The risk is therefore that we might create components that require one of the several popular frameworks, such as Angular in the case of Ionic. We would hope to be able to recycle major software components to reduce the burden of developing the next app. This is hampered somewhat by frequent breaking changes in major revisions of JavaScript frameworks.

Thankfully web standards such as browser APIs are far more stable. An alternative emerging pattern is one based on the Web Component standard discussed further in Section 7.. However in the next section we discuss the more general mobile implementation pattern of hybrid mobile apps.

6.2. Hybrid Mobile

Apache Cordova is an open source tool chain that produces a native installable application which can be published and installed from app stores. However the the app views themselves are effectively web apps rendered by the mobile phone's native web view API. The resulting *Hybrid mobile* allow crafting of apps by using the same technologies as web apps, thus ensuring cross-platform compatibility across the mobile platforms and desktop web.

Cordova also offers a plug-in architecture which allows us to call native features of the mobile platform. One use case is utilising the mobile platform's helpful content sharing activity, which is automatically aware of the user's installed sharing-capable apps such as email, Facebook, Google+ or even text messages and so forth. Using native mobile features can also result in a better user experience than the web API alone. An example here is the image picker plug-in that Zahwa uses to import photos.

The hybrid development model is not without drawbacks. The Cordova tool-chain is complex and occasionally unreliable. The open source ecosystem of plug-ins is sadly mired by a large number of abandoned repositories. In one case we required a clustering feature in the Google Map plug-in used for recipe discovery in Zahwa. We required a modification so we could override the default click-to-zoom behaviour to display a list of recipes. The project maintainer was unwilling to accept a pull request, forcing us to fork the plug-in and creating a new dependency with a maintenance burden.

That said, our experience of the hybrid app software pattern is broadly positive and is a world apart in terms of development productivity. The Zahwa app demonstrates it's possible to build apps which are virtually indistinguishable from native apps. We should also point out that pure web apps, e.g. web sites rather than installed apps, are a viable proposition in many use cases. On Android, these can operate offline and full-screen just like 'real' apps, but offer the benefit of a very rapid onboard process where sharing of a URL is all one needs to recruit a participant.

7. Web Component Implementation

Web Components (WC) are a live web standard[7] that has been fully implemented in Chrome, with Firefox still in progress. All remaining browsers can be 'polyfilled' to support WC now until they finish their native implementations, which should be complete by the end of 2018. WC offers a number of advantages including performance and interoperability regardless of JavaScript frameworks.

[7] `https://github.com/w3c/webcomponents`

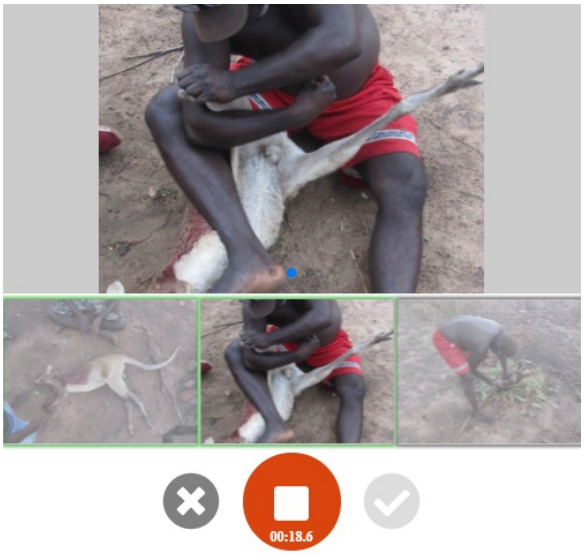


Figure 6: A Web Component implementation of the IGV record activity. Used here to document preparation of scrub wallaby in Arnhem Land, North Australia.

The typical approach for modularity in the web domain depends on directly low-level manipulation of the web browser DOM API. One example is the Dragula[8] drag 'n drop library which Zahwa uses to facilitate reordering of images. Dragula proved to be problematic given the multi-layered interface of an app, where the drag 'n drop component was merely the top most layer.

Part of the WC specification is the *Shadow DOM*, an isolated a DOM scope encapsulated within the web component. We are then free to scaffold our app with any JavaScript framework, without unwanted side-effects elsewhere in the app. Stencil[9] is a 'compiler' of Web Components, developed by the makers of the Ionic platform. Stencil allows us to continue to use current best-practices in web development, while producing standardised Web Components.

We used Stencil to implement the IGV recording and annotation activities as Web Components, see Figure 6. In the remainder of this section we demonstrate including the IGV WC on a web page, and interacting with the component via a JavaScript API.

```
1 <html>
2 <head>
3 <script src='https://unpkg.com/aikuma@0
      .0.1/dist/aikuma.js'></script>
4 </head>
5 <body>
6 <aikuma-image-gesture-voice></aikuma-image-
      gesture-voice>
7 </html>
```

The general principle is that Web Components may be selected like any other HTML element. They can also register public methods on the element. To illustrate, the following listing is an asynchronous JavaScript function:

```
1 async function doIGV() {
2   let wc = document.queryselector('aikuma-
        image-gesture-voice')
3   wc.loadFromImageURLs(['image1'...])
4   let results = await wc.waitForComplete()
5   console.log('IGV returned', results)
6 })
```

Line 2 is a standard Web DOM API call to find the first tag that matches the selector. On line 3, *wc* is now the Web Component instance and we invoke a method specific to this Web Component which imports a series of images by providing a list of URLs. Line 4 uses another method which returns a JavaScript Promise, thus we use the await keyword to yield execution back to the main thread until the Promise resolves. When the user has cancelled or completed the activity, the code block in the *then* statement is executed, printing the data structure to the web browser console.

WCs as plain JavaScript modules works particularly well with the existing JavaScript ecosystem. We have published the IGV components on Node Package Manager (NPM) repository under the @aikuma scope[10]. Please refer to the GitHub documentation for a full description of methods and IGV data structures. We also plan to release generic JavaScript libraries helpful for knowledge preservation apps.

8. Conclusion

We aim to develop mobile software that guides users through a series of steps to preserve their knowledge in the digital domain. The mobile platform offers the means for scaling up this preservation activity. We have described a general Image-Gesture-Voice documentary method and shown how it can be instantiated in a social app for preserving and sharing know-how.

Web technologies boost development productivity yet force us to engage with a complex software stack that is in a continual state of flux. This paper is a case in point: we have just built a production quality tool yet must anticipate the next implementation pattern. Thankfully, the situation is improving thanks to Web Components. As a official web standard, Web Components will have longer lifecycles than components built on this month's most popular JavaScript framework.

The time is ripe to build a common library of Web Components and other JavaScript libraries to reduce duplication and help the small community of people developing software for language documentation to focus our limited resources on well-engineered and field-refined components. We encourage others to apply and extend the work presented here, and to contribute to the improvement of software components to support the vision of scaling up digital preservation of cultural knowledge.

Acknowledgements

This research was supported by NSF grant 1464553 *Language Induction meets Language Documentation: Leveraging bilingual aligned audio for learning and preserving*

[8] https://bevacqua.github.io/dragula/

[9] https://stenciljs.com/

[10] http://www.aikuma.org/components.html

languages. We are grateful to Amina Mettouchi for supplying details of the Taos and Zahwa personas, and to Alexandra Marley for the scrub wallaby images.

9. References

Bettinson, M. and Bird, S. (2017). Developing a suite of mobile applications for collaborative language documentation. In *Second Workshop on Computational Methods for Endangered Languages*, pages 156–164.

Bettinson, M. (2017). Crafting the next generation of language documentation tools. Fifth International Conference on Language Documentation and Conservation.

Bird, S., Hanke, F. R., Adams, O., and Lee, H. (2014). Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5.

Bird, S. (2010). A scalable method for preserving oral literature from small languages. In *Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries*, pages 5–14.

Bird, S. (2018). Desiging mobile applications for documenting endangered languages. In Kenneth Rehg et al., editors, *Oxford Handbook on Endangered Languages*. Oxford University Press.

Blachon, D., Gauthiera, E., Besacier, L., Kouaratab, G.-N., Adda-Decker, M., and Rialland, A. (2016). Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app. In *Proceedings of the Fifth Workshop on Spoken Language Technologies for Under-resourced languages*, volume 81, pages 61–66.

Bolt, R. A. (1980). "put-that-there": Voice and gesture at the graphics interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, volume 14, pages 262–270. ACM.

Chatzimilioudis, G., Konstantinidis, A., Laoudias, C., and Zeinalipour-Yazti, D. (2012). Crowdsourcing with smartphones. *IEEE Internet Computing*, 16:36–44.

Chikonzo, A. (2006). The potential of information and communication technologies in collecting, preserving and disseminating indigenous knowledge in Africa. *The International Information and Library Review*, 38(3):132–138.

Goldman, J.-P., Leemann, A., Kolly, M.-J., Hove, I., Almajai, I., Dellwo, V., and Moran, S. (2014). A crowdsourcing smartphone application for Swiss German: Putting language documentation in the hands of the users. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 3444–47.

Himmelmann, N. P. (1998). Documentary and descriptive linguistics. *Linguistics*, 36:161–195.

Johnson, H. and Aristar Dry, H. (2002). OLAC discourse type vocabulary. `http://www.language-archives.org/REC/discourse.html`.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kendon, A. (2008). Some reflections on the relationship between gestureand sign. *Gesture*, 8(3):348–366.

Leemann, A., Kolly, M.-J., Purves, R., Britain, D., and Glaser, E. (2016). Crowdsourcing language change with smartphone applications. *PloS one*, 11(1):e0143060.

Leemann, A., Kolly, M.-J., and Britain, D. (2018). The English Dialects App: The creation of a crowdsourced dialect corpus. *Ampersand*, 5:1–17.

Lüpke, F. (2010). Research methods in language documentation. *Language documentation and description*, 7:55–104.

Mettouchi, A., Bettinson, M., and Bird, S. (2017). Documenting recipes. Fifth International Conference on Language Documentation and Conservation.

Norman, D. A. and Draper, S. W. (1986). User centered system design. *New Perspectives on Human-Computer Interaction*, 3.

Oh, S. and Syn, S. Y. (2015). Motivations for sharing information and social support in social media: A comparative analysis of Facebook, Twitter, Delicious, YouTube, and Flickr. *Journal of the Association for Information Science and Technology*, 66(10):2045–2060.

Pollock, N. (2011). The language of food. In Nick Thieberger, editor, *Oxford Handbook of Linguistic Fieldwork*, pages 235–49. Oxford University Press.

Reiman, W. (2010). Basic oral language documentation. *Language Documentation and Conservation*, 4:254–268.

Subramani, M. R. and Rajagopalan, B. (2003). Knowledge-sharing and influence in online social networks via viral marketing. *Communications of the ACM*, 46(12):300–307.

Woodbury, A. C. (2010). Language documentation. In Peter K. Austin et al., editors, *The Cambridge Handbook of Endangered Languages*. Cambridge University Press.

Beyond Protocol: Indigenous Knowledge Resource Circulation in the Digital Age

David Nathan
Groote Eylandt Language Centre
GEAT Building, Angurugu, NT 0822, Australia
dnathan@alcnt.com.au

Abstract

This paper considers protocol-based digital access management planned for *Ajamurnda*, a collection and access system for language and cultural items of the Indigenous Anindilyakwa people of Groote Eylandt, Australia. Ajamurnda will be a 'living' collection facilitating and regulating access and circulation of resources, based around protocol – consideration of the personal, communal, cultural, property and privacy interests of individuals, families and other culturally-relevant groupings. In the specific, highly traditional context of Groote Eylandt, standard regulation of access using accounts and passwords are ineffective. Ajamurnda will instead use a 'sanctions before barriers' strategy based on the fact that in such Aboriginal communities, acquiring and holding knowledge has consequences, and that these consequences will be best known by users themselves, and act as constraints on choice. For those of us seeking to implement a fully authentic implementation of protocol, such a 'sanctions before barriers' approach is probably the only way that access protocol can be fully informed, authentic and nuanced, and responsive to dynamics of knowledge circulation in the community.

Keywords: Anindilyakwa, access, protocol

1. Introduction

This presentation describes research and planning for *Ajamurnda*, a digital collection and access system for language and cultural resources. Ajamurnda is being designed by the Groote Eylandt Language Centre on behalf of the island's Indigenous community and will provide protocol-based access management. Anindilyakwa is the name of the language and culture of the Indigenous people of Groote Eylandt in northern Australia, and is also used to refer to the people themselves. While Ajamurnda was originally conceived as a repository for language materials, it will include a range of resources representing the community's language, culture, families, land, history and events because the boundaries between language, culture, land and history are overlapping and fluid for the Anindilyakwa community whose language, culture and lifestyle are amongst the least colonially disrupted of Australia's Indigenous peoples.

While our Centre currently hold collections of digital and waiting-to-be-digitised materials, by far the greatest amount of relevant knowledge is in the form of individuals' knowledge which is shared orally, and thus at risk of being lost as time passes. Therefore, Ajamurnda will be an ongoing participatory system including a 'crowdsourcing' function to enable community members to enrich the collection by adding resources and metadata, and correcting existing information (Christen 2011, Garrett 2014).

2. The protocol context

Our goal is to research and build a participatory platform that authentically represents Anindilyakwa methods of facilitating and regulating access. We are exploring methods which are feasible to implement in a range of real community settings and which innovatively use cultural strategies such as self-identification, cultural sanctions, language, and location, while avoiding technical barriers such as user accounts and passwords.

I use the term 'protocol' to refer to respecting materials that are sensitive, sacred, dangerous, shaming, private, or restricted in other ways so that access needs to be regulated. Ajamurnda will hide/protect materials where required, while otherwise making access as easy as possible. Protocol is dynamic over time, because sensitivities and restrictions change, just as, for example clan lands on Groote Eylandt are closed and later reopened after people pass away. Understanding and implementing protocol involves ongoing participation by a range of people to reach understandings of the cultural dynamics of knowledge holding, ownership, control, circulation and access.

The past 15 years have seen a parallel emergence of language documentation for endangered and minority languages, together with use of digital technologies to record and share the resulting documentation. A feature of the language documentation movement has been attention to the ethics of fieldwork and data collection, with increasing inclusion of native-speaker community values and participation (Czaykowska-Higgins 2009). Alongside that, several language archives were established, with varying degrees of emphasis on and implementation of access protocols meeting community expectations. The DoBeS archive[1] enables depositors (who are trusted to act on behalf of the people they have recorded) to decide whether public access to resources is allowed. The Endangered Languages Archive at SOAS University of London[2] established an innovative system of negotiated access involving exchange of information between depositors and requesters to determine whether access is appropriate (Nathan 2010). Ara Irititja[3] archives in Australia have focused tightly on providing functionality and access to Aboriginal community members. Several archives based on the Mukurtu[4] system use a nuanced system of protocols and licences to regulate addition of, access to, and usage of resources.[5] We are adapting and extending these examples of protocol implementation to suit the Anindilyakwa community.

[1] See dobes.mpi.nl
[2] See elar.soas.ac.uk
[3] See www.irititja.com and www.keepingculture.com
[4] See mukurtu.org
[5] See, for example, plateauportal.libraries.wsu.edu

3. Strategies: lowering the barriers

The potential of digital technologies to serve the needs of Indigenous communities has been long recognised (Nathan 2000). However, simply going digital is no measure or guarantee of success. Indeed, in relation to archived language resources, some have pointed out the low rate of community member access to resources (Trilsbeek & König 2014).

In planning Ajamurnda, we ask whether providing a standard system for digital cultural resource management for the Anindilyakwa community – which has very high continuity with its classical culture, values and dynamics – risks a 21st century version of ignoring, erasing, and failing to learn from Aboriginal civilisation, in the same way that Bruce Pascoe describes the many ways that colonisation of Australia has not only 'ignored ethnographic evidence of Aboriginal engineering' but erased that knowledge through blind introduction of imported practices (Pascoe 2014:65).

Two decades of work by Aboriginal lawyer Terri Janke and her colleagues have shown the inadequacy of Australian law to reflect the principles and nuances of Indigenous law, especially in regard to communal or group rights, also known as Indigenous Cultural and Intellectual Property, or ICIP (Janke 1998). The importance of recognising ICIP is heightened for groups such as the Anindilyakwa, where intangible resources (such as knowledge, stories, designs etc.) represent a much larger proportion of the stock of valued property than in 'western' communities. While there have been some (limited) accommodating changes in Australian law over these decades, current approaches of most memory institutions (archives, museums, libraries etc.) to implementing protocols for digital access remain inadequate, as they typically involve an immutable binary of 'open' vs 'closed', with elevated access only on the basis of individual accounts and logins. Although there have been efforts to develop more culturally appropriate ways of regulating access that observe community protocols, all of them are ultimately implemented using digital barriers.

3.1 The login/identification barrier

One kind of 'digital barrier' is a login process that requires presentation of a correct password or other token of permission that has been pre-arranged and verified by a digital system – an arrangement typically called an 'account'. Normally, we do not notice that such systems conflate *identification* with *authorisation*: identification (usually called 'authentication') refers to a system's confirmation that the user is who they say they are, while authorisation refers to the system's satisfaction that login has been obtained legitimately and enables access only to permitted resources. Identification and authorisation can be linked; for example, a bank teller can access information that the customer cannot. But where protocols for access reflect complex social conventions and dynamics, such as for the Anindilyakwa community, then the interplay between identification and authorisation becomes ever more complex.

Conventional barrier systems do not work well for Anindilyakwa people. Most are not living in a world of literacy (and knowledge of English is limited), so navigating web pages to set up accounts, personal profiles, and passwords can be difficult and demotivating. While smart phones are common throughout the community, passwords are frequently lost or misremembered. Devices such as phones and iPads are frequently shared and borrowed, making personal accounts an approximation at best.

We plan to use several strategies to maximise the ease of using Ajamurnda. The first is to implement identification by the presentation of screens showing photographs of individuals' faces, where an individual identifies themselves by clicking or touching on their own face image. For further discussion of this strategy, see Section 4.1. Secondly, Ajamurnda will use images wherever possible, for example icons and previews for navigation and browsing. The third strategy is called 'language first' and 'audio first'; where possible, we will provide navigation, explanation and content in the Anindilyakwa language conveyed as audio (since few people have functional literacy in the language).

The fourth strategy is pre-enrolment of users, which avoids users themselves having to set up accounts. This opportunity may be fairly unique to the Groote Eylandt situation where almost the whole community of a little over 1,000 people live on an island that is only 50 kilometres from east to west and where community visits to raise awareness about Ajamurnda. In addition, the Anindilyakwa Land Council holds lists of Indigenous residents for royalty payments, genealogical data, and other purposes, which can 'seed' the user base.

3.2 Sanctions before barriers

A second kind of digital barrier occurs when a user is denied access to a particular resource – humorously characterised as 'computer says no'. This is the most common way that access is regulated; a resource has associated metadata which indicates that it is 'closed'. The archive systems mentioned in Section 2 all recognise that access regulation needs to be more nuanced. Ajamurnda takes the opportunity to focus on serving a single primary audience – the Anindilyakwa community – to research and implement methods that best fit the community's needs.

We call a key strategy 'sanctions before barriers'. It is based on the fact that in Aboriginal communities such as the Anindilyakwa, acquiring and holding knowledge *has consequences*. Individuals, families, and groups have rights to, and consequently, potential knowledge of, particular stories, histories, ceremonies, objects, designs, places, and environmental resources. These rights are codified in terms of clan and family, ancestry, places of origin, gender, age and individual factors such as recognition of skills and seniority. The 'rules' or conventions for governing knowledge circulation are subtle, complex and dynamic, and a full description is both under investigation and beyond the scope of this paper. They go far beyond fixed attributes such as 'owner', 'gender', or 'open/closed'. For example, many cultural resources and events have 'managers' – people who bear responsibility for negotiating the transmission and integrity of resources, and who are not necessarily the owners of producers of those resources. These people are known by the Anindilyakwa (and some other nearby Aboriginal nations) as the *Jungkayi* (for a particular story, song, place, ceremony etc.). The identification of the appropriate Jungkayi to be consulted for any particular access event is a complex matter in itself.

Regulating access through 'sanctions before barriers' is a major component of meeting the complex dynamics of community-oriented access and participation. The concept was born from synthesising ethnographic observations and interviews with colleagues. It was crystallised following an account of how access to highly-sensitive men's and women's objects is implemented in the community's arts workshop. In that workshop, which is more-or-less a public space, there are two cupboards that contain, respectively, items restricted to viewing by men, and items restricted to viewing by women. Community members access these cupboards regularly, in conformance with the gender protocol. However, neither cupboard is locked, or difficult to reach or open. This shows us that observance of protocol can be driven from individuals' choices. Those choices are strongly influenced by community values and by the risk of incurring sanctions; the strong sense that events are connected means that an individual's breach of protocol is likely to result in negative consequences. Indeed, if a resource-accessing event had no consequences, the access is simply a completed transaction and the resource becomes a commodity rather than an element within the rich web of regulated knowledge distribution in a community.

Of course other forms of media and circulation involve 'consequences' ranging from zero to a level which defines the form itself. For example, a loan of a library book has few consequences for the library-using community. A radio broadcast has midrange consequences because it provides a shared daily experience to its listeners. Participation in today's social media – Facebook, Twitter etc. – not only populates and feeds their content but defines their purpose.

More work needs to be done, but for now we begin with the fact that access to knowledge has consequences, and that these consequences will be best known by users themselves, and act as constraints on choice and action. We do note that materials of recognised high sensitivity need to be pre-identified and restrictions explicitly applied. But for those of us seeking to implement a fully authentic implementation of protocol, a 'sanctions before barriers' approach is probably the *only* way that access protocol can be fully authentic, nuanced, and responsive to the dynamics of knowledge circulation in the community.

3.3 Location matters

An additional strategy is to use location-based access through simple, low-tech ways of controlling access to resources according to where a user is. Particular resources can be accessed without restriction in a supervised computer room, such as in the Language Centre (because, for example, a supervisor can ensure that only adults are using the catalogue). Many protocol-related attributes revolve around location – sacred places and stories associated with them, or ownership by clans and families who are associated with particular lands. With today's network technologies, we can make resources accessible on the basis of location, either using digital location services (where, for example, users have smart phones), or, more simply, by enabling access to specific resources through narrow-casting from physically localised wireless access points at outstations, townships or buildings.

4. Discussion

4.1 Facing identities

Using face-image selection as a way to establish users' identities solves some problems but raises interesting questions. Might people be sensitive about their faces being photographed and included? They may, but we anticipate that any such sensitivities are likely to coincide with factors that we need to take care of in any case (such as hiding images and references to people who have passed away). What if a community member selects someone else's image to indicate the 'wrong' identity? While this might be a breach of protocol, it might in some circumstances actually be culturally appropriate, since certain persons, via their kin relations, can be considered as equivalent. Even where that equivalence does not apply, the system will add access events to its 'living map' of knowledge circulation, which will be made visible to certain parties and therefore enable questions to be raised and followed up in the community. More importantly, user identification in Ajamurnda is *not* meant to be a direct proxy for individual account holding, because it is highly likely that more than one person at a time will be around a device and using it to access resources. In that case, it will be easy for those participants to identify each of themselves (by selecting their face images) and thus for their participation to be included in the system's knowledge circulation map. Finally, it should be emphasised that Ajamurnda's protocol system is necessarily a learning platform for exploring new and better ways to cater for evolving community needs and preferences, and ongoing usage will answer some of these questions.

4.2 Regulating access beyond the community

A web-based system will be potentially open to view by millions across the world. A 'sanctions before barriers' strategy can only work where the 'rules' and 'consequences' are known by a user and genuinely felt to affect his/her feelings, welfare and perhaps result in other more serious outcomes. Thus access choices by outsiders – non-Anindilyakwa people – will not reliably ensure conformance to protocols, whether or not those outsiders sympathetically respect explicit guidelines presented on the Ajamurnda website. So how can access by non-Anindilyakwa people be regulated?

There is no clear dividing line between community members and non-community members. Leopold (2013) notes, in the USA context, that 'diaspora communities and tribal members living off the reservation' are rarely considered when designing access regulation – a situation relevant to some Anindilyakwa people who have phases of residence off Groote Eylandt.

Thus for 'sanctions before barriers' to work we need to distinguish Anindilyakwa community members from 'outsiders'. To do this, we can use some of the same mechanism which identifies community members. Like a 'Captcha' which web pages use to distinguish robots from people, the system will use images as a shibboleth, by presenting selected photos of community members and asking a question about them (e.g. in Anindilyakwa, "are these people (a) cousins (b) siblings (c) partners …" or similar).

5. Conclusions

An act of accessing a cultural resource can have many consequences. Drawing attention to potentially negative consequences to guide community members' access choices is just one. Other consequences are positive: along with supplying users with the resources they seek, Ajamurnda will, by representing accessing identities and access events, also become a kind of 'living map' of the sources and circulation of Anindilyakwa knowledge.

While we expect Ajamurnda to open new horizons in protocol-managed access to resources, few of the concepts mentioned here are genuinely new. It is easy to spot other ways in which access to resources has consequences. Marshall McLuhan explained, as far back as 1959, that electronic media turns its users into participants who are creative 'co-authors' and 'co-producers' (McLuhan 1959). He thus also anticipated, 40 years earlier, the rise of social media. Today it is difficult to buy anything (most likely online) without being asked for a review of the product, which is then shared to influence others.

A recent article 'Estonia, the Digital Republic'[6] points out that it is a central ingredient of personal data protection in the upcoming 'digital societies' that all people must be able to know who has looked at their data, such as medical records.

While most existing cultural resource repositories have stuck with simple user account methods which suit academic researchers, Indigenous peoples should not be denied the potential of innovative systems that meet their values and needs.

The Ajamurnda team has had initial discussion with the Mukurtu team based at the Washington State University led by Dr Kim Christen. The Mukurtu system is an ideal springboard for Ajamurnda, since Mukurtu is based on the highly ubiquitous, open-source CMS Drupal, has had several cycles of community-influenced development, and provides a robust platform for further expansion of community-controlled protocol-based access to resources. The Ajamurnda team plans to work with the Mukurtu team to build and share new capabilities based around careful implementation that meets the Anindilyakwa community's values and dynamics, and with a focus on representing the consequences of users' interactions with digital collections. We hope that this new and ambitious implementation of community-oriented digital resource management will contribute to the Anindilyakwa community's cultural continuity and similarly inspire others.

6. Acknowledgements

The Ajamurnda Project is funded under the Australian Commonwealth Government's Indigenous Languages and the Arts grant ILAO1700002 and supported by the Anindilyakwa Land Council. Thanks to Howard Amery, Hugh Bland, Alex Bowen, Melainie Collins, Carolyn Fletcher, Hannah Harper, Salome Harris, Judy Lalara, Leslie Pyne, and Sylvia Tkac for discussion of issues raised in this paper.

[6] See www.newyorker.com/magazine/2017/12/18/estonia-the-digital-republic [Accessed 21-01-2018]

7. Bibliographical References

Christen, K. (2011). Opening Archives: Respectful Repatriation. *The American Archivist*, 74:185–210.

Czaykowska-Higgins, E. (2009). Research Models, Community Engagement, and Linguistic Fieldwork: reflections on Working within Canadian Indigenous Communities. *Language Documentation and Conservation*, 3:15–50.

Garrett, E. (2014). Participant-driven language archiving. In D. Nathan & Peter K. Austin (Eds.) *Language Documentation and Description*, 12:68–84 (Special Issue on Language Documentation and Archiving). London: SOAS.

Janke, Terri (1998). *Our Culture: our Future. Report on Australian Indigenous Cultural and Intellectual Property Rights.* Canberra: ATSIC [also online at http://www.terrijanke.com.au/our-culture-our-future, accessed 6 March 2018]

Leopold, Robert (2013). Articulating Culturally Sensitive Knowledge Online: A Cherokee Case Study. *Museum Anthropology Review*, 7(1-2):85–104.

McLuhan, M. (1959). Electronic Revolution: Revolutionary Effects of New Media. In S. McLuhan & D. Staines (Eds.) *Understanding Me: Lectures and Interviews / Marshall McLuhan*, 1–10. Toronto: McClelland & Stewart.

Nathan, D. (2010). Archives 2.0 for Endangered Languages: from Disk Space to MySpace. *International Journal of Humanities and Arts Computing*. 4(1–2):111–124. Edinburgh: Edinburgh University Press.

Nathan, D. (2000). Plugging in Indigenous knowledge – connections and innovations. *Australian Aboriginal Studies*, 2:39–47.

Pascoe, B. (2014). *Dark Emu, Black Seeds: Agriculture or Accident?* Broome: Magabala Books.

Trilsbeek, P. & König, A. (2014). Increasing the future usage of endangered language archives. In D. Nathan & P. K. Austin (Eds.), *Language Documentation and Description*, 12:151-163 (Special Issue on Language Documentation and Archiving). London: SOAS.

Hard Numbers: Language Exclusion in Computational Linguistics and Natural Language Processing

Martin Benjamin
Kamusi Project International
Place de la Gare 12 C
1020 Renens, Switzerland
martin@kamusi.org

Abstract

The intersection between computer science and human language occurs largely for English and a few dozen other languages with strong economic or political support. The supermajority of the world's languages have extremely little digital presence, and little activity that can be forecast to change that status. However, such an assertion has remained impressionistic in the absence of data comparing the attention lavished on elite languages with that given to the rest of the world. This study seeks to give some numbers to the extent to which non-lucrative languages sit at the margins of language technology and computational research. Three datasets are explored that reveal current hiring and research activity at universities and corporations concerned with computational linguistics and natural language processing. The data supports the conclusion that most research activity and career opportunities focus on a few languages, while most languages have little or no current research and little possibility for the professional pursuit of their development.

Keywords: under-resourced languages, computational linguistics, NLP, language technology

1.0. Introduction

This paper looks at technology for "under-resourced" languages by examining the amount of career opportunities and research projects in the field. Two data sets were evaluated to provide hard numbers regarding the proportion of high level research in areas related to computational linguistics. The hypothesis for the research was that high level pursuit of technological development for most of the world's languages is not a widely available career option. This hypothesis was fully supported by the data. Without a significant number of people working in the field, resources for under-resourced languages cannot be developed. The numbers in this study, which indicate where the field will be casting its gaze for years to come, give no cause for optimism that the situation will improve.

People who work on excluded languages know from experience that most of the world's languages remain outside of the technological sphere, but do not have numerical ways to demonstrate the extent of the marginalization. In principle, one could look to the amount of software that is localized in each language, but that would involve getting access to thousands of programs, installing them, and enumerating the available user interface languages – an impossible task that one already knows would reveal almost nil coverage for the supermajority of languages. One could hunt for resources per language, and find a corpus here, a spell-checker there, and a bunch of Wikipedia stub pages about asteroids somewhere else[1]. In the end, though, a spreadsheet with 7000 languages and all known technologies would show a smattering of ticks for a long tail of languages, for example where a passionate developer created an Android app[2] or where a field linguist shared a dictionary on Webonary,[3] and a huge clustering of resources for a small assortment of languages that could be guessed without looking at the data. Kornai's impressive effort (2013) to quantify existing resources for languages of the world found that 6,541 had no detectable live online presence. Following Scannell (2013) and Gibson (2014), it is possible to find instances of usage of nearly 2000 languages within communication technologies such as Twitter, but these are examples of technology as a vessel rather than an avenue for development. Because data about the topic of research activity is not obviously available, we have been left to make impressionistic assertions about the paucity of work in the field.

This study examined three sources of data that provide numerical indications of the extent to which under-resourced languages are active within the overall profession of language technology. The first

[1] The Yoruba Wikipedia, https://yo.wikipedia.org, has more than 31,000 articles listed. However, most of those contain bogus content, including thousands of pages like https://yo.wikipedia.org/wiki/23006_Pazden that are bot-generated stubs containing the names of asteroids. Clicking the link labelled "Ojúewé àrìnàkò" from any Yoruba Wikipedia page gives a random page from the project, with a high probability of landing on an asteroid.

[2] https://mothertongues.org/

[3] https://www.webonary.org/

dataset consists of all the jobs posted on Linguist List (LL) in 2017 that specify applied, computational, or text/ corpus linguistics.[4] The second dataset consists of all the papers and posters presented at COLING 2016,[5] the 26th International Conference on Computational Linguistics, organized by the Association for Natural Language Processing, in Osaka, Japan. The third consists of all the papers and posters presented at ICLDC 2017, 5th International Conference on Language Documentation & Conservation, in Honolulu, Hawaii. None of these are perfect representations of the state of the field, for reasons discussed below, but they give an overall up-to-the-moment indication of the state of attention that under-resourced languages receive among those active in the profession.

Spoiler alert: under-resourced languages receive almost no attention in work related to computational linguistics or natural language processing (NLP).

Type	Number/ Examples	Characteristics
A	4 languages (English, French, German, Spanish)	Massive investment, many existing digital resources, large monolingual and aligned corpora, somewhat functional machine translation with other languages in the group, primary focus of language technology research and development.[6]
B	About 25 languages (many official languages of the EU, Chinese, Japanese, Russian, Arabic)	Moderate to large investment and research, increasing digital resources, large monolingual corpora with bilingual alignment to A languages (especially English), rough machine translation to A languages (usually English), focus of interest for EU and national funders
C	All the rest. Almost 7000 languages, spoken by the majority of the world's 7 billion people.	Zero to mediocre investment and research. Some languages like Swahili and major languages of India, with more than 100 million speakers, have active research communities and rough machine translation, usually to English. A couple of thousand have some form of print dictionary, ranging from lists of a few hundred words to massive volumes with hundreds of pages. Most are 'embattled' – either close to extinction, or disfavored by policy or practice. Funding is usually sparse.

Table 1: Language Categories

2. How much less resourced are "less-resourced" languages?

Table 1 proposes a typology of languages, wherein languages in Category A are the ones that receive high attention in NLP research, Category B languages receive moderate attention, and Category C languages receive little or no attention. Figure 1 shows those languages at exact scale, proportional to the total number of languages in each category: a square with 4 units for Category A, 25 units for Category B, and nearly 7000 units for Category C. Figure 2 was a speculation drawn about two years prior to the present study that posited the ratio of research invested in each category as a rough inverse of the number of languages affected. This paper examines the hypothesis implicit in Figure 2. The hard numbers in this study show that the scale shown for research activity between categories A and C is about right. The representation underestimates the level of activity for Category B languages, however – mid-resourced languages, where Kilgarriff and Grefenstette's 2003 observations about trends toward increasing digital multilingualism hold true, should have a bigger box, with a gradient toward "under-resourced" that is certainly reached around 50. On the other hand, the data bears out that several languages that were cast as Category B – notably Chinese, Japanese, and Arabic – are benefiting from significant professional attention, and could now be on the borderline of Category A, which would maintain the ratio between A and B closer to the initial depiction. While a discussion of the state of Category B languages is outside of the scope of this paper, as they enjoy a host of advantages that elevate them above any notion of "under-resourced", it is worth noting that many are making strides that will redound increasingly to their benefit in the years to come.

[4] Records were laboriously reviewed by setting search criteria on the Linguist List jobs page, https://linguistlist.org/jobs/search-job1.cfm. Linguist List keeps records dating back many more years, but procuring those in a practical format would require imposing on their staff. Whether a single year's data is completely representative is therefore an open question.

[5] http://coling2016.anlp.jp/

[6] The list of "Any language" treated by DeepL (English, German, French, Spanish, Italian, Dutch, and Polish) at http://www.deepl.com/translator could be a better estimate, but discussion of levels of inclusiveness among better-resourced languages would be the subject of a different paper.

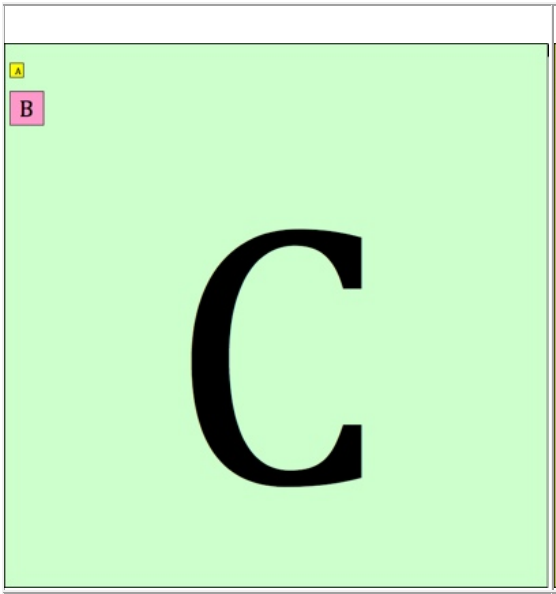


Figure 1: Languages of each type as a proportion of world total (actual numbers)

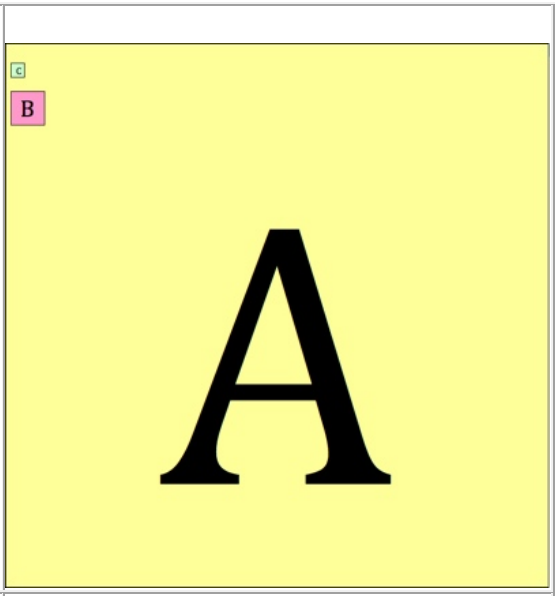


Figure 2: Languages of each type as a proportion of investment and research attention (hypothesized estimate)

2.1 Linguist List

The LL data shows all 426 jobs posted for applied, computational, or text/corpus linguistics during 2017. Of these listings, 309 mention one or more languages. A total of 42 languages are mentioned, in addition to the categories "African", "Aboriginal", "Foreign languages", "Germanic", "Indigenous languages of North America", "Multilingual", "Pacific Pidgins and Creoles", "Romance", and "Turkic". By far the most frequent language mentioned is English, with 128 listings. Second place goes to 117 unspecified listings. Random inspection shows that "unspecified" often means English, but if not English will almost certainly involve one of the other languages in Category A or B; for example, a position[7] is open to "any language, preferably the languages taught at the Center": Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Portuguese, Russian, and Spanish. The combination of the four original Category A languages and "unspecified" yields 344 total mentions (with some jobs mentioning more than one language). The next 25 languages or clusters are mentioned 160 times. Finally, 21 languages or clusters are mentioned one time each.

Many of the languages at or near the bottom appear only in jobs announced by the translation company Lionbridge or a company called Pygmalion that is working on NLP. These include several languages of the Indian sub-continent, Tagalog, Thai, and Vietnamese, each spoken by tens of millions of people. The inclusion of these languages may indicate a glimmer of recognition that some excluded languages may harbor a hidden profit potential. However, impressionistic evidence, including professional visits by the author to language technology offices in India and Vietnam, communications with linguists and technologists throughout the region, and interviews with internship candidates for NLP positions from universities around Asia, does not indicate that any of these languages other than Hindi could be considered a candidate for current or imminent inclusion in Category B.

Several languages at the bottom share the profile of most of those that do not even appear: economically and politically powerless, and not considered as candidates for technology by either producers or consumers. These include the Bajau language of Indonesia, Crow in the US, Inuktitut in Canada, Keres in Mexico, the Tok Pisin pidgin of Papua New Guinea, and the African, North American "indigenous" or "aboriginal" languages, and Turkic language clusters writ large. Closer inspection of the announcements reveals that none of these jobs involve NLP. For example, the position for African and Turkic languages[8] is for an undoubtedly fascinating project called "Discourse reporting in African storytelling" for which post docs are expected "to conduct fieldwork collecting traditional narratives, develop an annotated corpus of narrative texts, analyze selected aspects of these texts, and collaborate with other members of the team on theoretical issues related to the encoding of

[7] https://linguistlist.org/issues/28/28-5377.html

[8] https://linguistlist.org/issues/28/28-5291.html

reported discourse". Crow and Keres[9] entailed "cutting and labeling audio, data entry from handwritten notes, additional tasks relating to analysis and organization of the data, and some retyping of existing corpus." No university or corporation on the planet took advantage of the free services of LL to advertise for a linguist to work on the development of a single language of Africa or South America, nor any but the most lucrative or politically well-placed languages of Asia or Europe.

While the LL data is indicative of the global state of hiring, it should not be considered definitive for several reasons. The list only includes jobs where HR or the search committee is aware of the LL job board and considers it important. Posters include universities, translation companies, and some big technology companies such as Amazon and Google. However, many companies that seek employees for NLP elsewhere, such as Angel List,[10] are absent from LL, perhaps because they are more interested in hiring people with a computer science background than with training specifically in linguistics. Further, LL does not penetrate to many national job markets for Category B languages where conferences such as COLING demonstrate that active research is underway, such as Polish, Catalan, and Turkish. We cannot, therefore, make universal claims from the data, but we can use it as strong support for what were previously anecdotal inferences. In particular, the data is not granular enough to support conclusions about which languages belong in Category B or how extensively work is available in those languages. However, matching the LL data against the list of languages that have been identified with ISO 639-3[11] codes shows nearly 7000 zeros: it is beyond doubt that no jobs were available anywhere in 2017 for work on language technology for the supermajority of the world's languages.

2.2. COLING 2016

The COLING 2016 program listed 230 papers and posters. Of these, 16 languages were mentioned by name (English, Arabic, Chinese, German, Hebrew, Hindi, Japanese, Korean, Manipuri, Mongolian, Polish, Sanskrit, Spanish, Thai, Turkish, and Urdu), 131 did not specify a language, 53 papers were classified as "multilingual", and 10 papers were classified as "under-resourced" due to their inclusion in a special track for the topic. Only 13 specified English, but random scanning showed that as the language of analysis for many "unspecified" papers. Many of the papers in the "multilingual" category dealt with machine translation, which is inherently about more than one language, and closer inspection shows to pertain most often to Category A or Category A+B languages. For the many papers that did not specify a language, random inspection showed only Category A or B as languages of concern.

An attempt was made to estimate the language of the research based on the last name of the lead author. No bankable results were achieved, because many names could not even begin to be associated with a language, and many names that indicate the ancestry of a researcher do not indicate their current location, research interests, or available datasets. However, it is probably not a coincidence that the location of the conference in Asia, and the high participation of researchers from Chinese and Japanese institutions, coincided with 94 submissions from people with names associated with China and Japan. Although their paper titles might not have specified Chinese or Japanese as the languages of research, 15 could be identified from their descriptions as pertaining to those languages. Many papers submitted from Asian institutions that focused on deep computational issues, though, such as "Asynchronous Parallel Learning for Neural Networks and Structured Models with Dense Features", often used English as their data core, since that is where they could benefit from and measure themselves against other research; unfortunately, author's institution was only available in the processor-crushing 3500 page proceedings PDF, so the potentially fruitful inquiry of the extent to which English is central to research interests in non-English countries was not practical. In no case did inspection of an article in the proceedings that was not labeled "under-resourced" reveal research in a Category C language, and no papers were submitted to the conference by authors with a name that was discernably African or from an otherwise under-represented language area. Though more detailed research about whether NLP researchers focus on their own languages or the languages with high industry demand could reveal interesting sociological patterns, the present findings about surnames are reported in the spirit of "negative results", a hypothesis tested and found to be unsupported by the evidence at hand.

As with the LL data, COLING data was not extensive enough to make statistically valid claims about the global distribution of research on any given language. Even more than LL, many Category B languages were not represented at all at the conference, though we know that research on languages such as Danish, Dutch, and Romanian is occurring at institutions in the countries where those languages are spoken. However, as with the LL data, we can make certifiable observations about where

[9] https://linguistlist.org/issues/28/28-3570.html
[10] https://angel.co/
[11] https://en.wikipedia.org/wiki/ISO_639-3

research is generally not happening: almost all of the languages in Category C.

2.3. ICLDC 2017

ICLDC[12] is a biennial conference that attracts people working on excluded languages, especially those spoken in countries around the Pacific Rim. The conference program[13] was examined for a reverse perspective on the other two data sets. The question was, among scholars and practitioners of under-resourced languages, what proportion of research activity is given to developing technological resources?

166 papers and posters are listed in the conference program. Workshops and roundtables were not considered. The titles were judged on the single criterion of whether they pertained in a broad way to digital technology. "Creating a Digital Shell for Indigenous Language and Culture Sharing" was considered relevant, whereas "Languages, 'Languoids', and ISO-codes for Language Diversity and Variation" was judged to be outside the scope of technology development. The assumption was that all papers dealt with Category C languages. About 75 languages are indexed in the conference program, with African languages having very little representation.

Twenty-four papers, or 14.5%, met the criterion for relevance to improving technological resources. Most of these were discussions of the creation of particular data resources or learning tools. For example, "Leveraging Web Technologies to Enrich Archival Materials for Use in Language Revitalization" is a discussion of the digitalization and use of archival materials for an Alaskan language; such a corpus building activity is foundational for potential future NLP, but does not involve computational advances *per se*. Similarly, "Large-scale Language Documentation in Nepal: A strategy based on SayMore and BOLD" is about the use of software to produce data, not the development of software itself. "Re Tlli7sa ell re uqw7úq̇wis: Engaging Indigenous language learners with an epic story through a language learning app", an example of how technology can be used in the service of endangered languages, is about the use of digital tools, not their production.

The ICLDC data demonstrate that work on language technology and work on under-resourced languages are conducted by almost completely different groups of people. This is correlated to the jobs board on LL, where some (not many) positions regarding under-resourced languages can be found using "language documentation" and "lexicography" as the search criteria instead of the technology-oriented criteria stated in Section 2.1. Similarly, the fourteen chapters of Day, Rewi, and Higgins (2016) that deal with contemporary research on "besieged" languages give only glancing mention to possible inclusion on the Internet or within communication technologies. Succinctly: research activity on non-lucrative languages rarely intersects with the development of language technology resources.

2.3.1. Coda: ComputEL-2

As a counterpoint to the previous sections, mention should be made of a specific effort to assemble practitioners of technology for under-resourced languages. ComputEL 2 was the second Workshop on Computational Methods for Endangered Languages, [14] held immediately after ICLDC 2017. This workshop featured 23 presentations[15] on themes related to excluded languages and technology. Many of these had concerns similar to those of ICLDC, developing and using digital data, such as "Endangered Data for Endangered Languages: Digitizing Print dictionaries". However, a few of the papers could have been presented at COLING instead. For example, "Improving Coverage of an Inuktitut Morphological Analyzer Using a Segmental Recurrent Neural Network" and "Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection" both deal with the sorts of issues that are at the forefront of computational linguistics. Given that under-resourced languages such as East Cree and Inuktitut are, in aggregate, spoken by more than half the world's population, it should be shocking that one has to scour the planet for a small workshop devoted to their advances within language technology, not just as a question of equity but as one of research opportunity.

3.0. Conclusions

As seen at ComputEL, excluded languages are viable candidates for the computer science aspects of language research. In fact, one could argue that a topic such as East Cree verb inflection provides a challenge that is more likely to push the edges of computational linguistics than yet another foray into English data. Thousands of fascinating research questions that could push the frontiers of NLP are not being asked, because technology research is trapped in a small set of well-picked-over languages,

[12] http://icldc5.icldc-hawaii.org/

[13] http://icldc5.weebly.com/uploads/2/4/9/6/24963413/icldc_5_program.pdf

[14] http://altlab.artsrn.ualberta.ca/computel-2/

[15] http://altlab.artsrn.ualberta.ca/computel-2/computel-2-accepted-presentations/

while the limited opportunities for under-resourced languages do not involve pursuing their digital futures. From an intellectual perspective, the chasm between computation and Category C languages represents many lost opportunities for scientists to forge into fresh, uncluttered territory. For corporations, blinders about the profit potential of diverse languages could be overlooking vast markets, particularly for about 350 languages with more than a million speakers but virtually no technological presence. To give one final data set, examine the "List of LREC 2016 Shared LRs" (Language Resources);[16] or, on a blind bet, the forthcoming list for 2018; scrolling through the list, there is no need to get an exact count to see the extent to which LREC members produce resources almost exclusively for Category A and B languages. Though the datasets are too small to draw iron-clad statistical conclusions, the data evaluated in this paper gives hard numbers beneath a hard truth: not only are most languages currently neglected from the digital sphere, but today's hiring and research activity destine that exclusion to continue without end. Computational linguistics and NLP hardly intersect with the supermajority of the world's languages, job positions rarely appear to pursue such intersections, and research for most languages remains perpetually stalled.

References

Day, D., Rewi, P., and Higgins, R., eds. (2016) The Journeys of Besieged Languages. Cambridge Scholars Publishers.

Gibson, M. (2014). A framework for measuring the presence of minority languages"in cyberspace. Presentation to the 3rd International Conference on Linguistic and Cultural Diversity in Cyberspace, Yakutsk, Russia.

Kilgarriff, A., and Grefenstette, G. (2003) Introduction to the special issue on the web as corpus. Journal of Computational Linguistics – Special issue on web as corpus. Volume 29, Issue 3, September 2003, Pages 333-347.

Kornai A (2013) Digital Language Death. PLoS ONE 8(10): e77056. doi:10.1371/journal.pone.0077056.

Scannell, K. (2013). How many languages are on the web? The Crúbadán project 10+ years on, invited talk at the Workshop on Corpus-based Quantitative Typology (CoQuaT 2013), Leipzig, 14 August 2013

[16] http://lrec2016.lrec-conf.org/en/shared-lrs/

Unlocking Cultural Conceptualisation in Indigenous Language Resources: Collaborative Computing Methodologies

Amelie Dorn[1], Eveline Wandl-Vogt[1], Yalemisew Abgaz[2], Alejandro Benito Santos[3], Roberto Therón[3]

[1]Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Vienna, Austria [2]Adapt Centre, Dublin City University (DCU), Dublin, Ireland [3]VisUsal Group, Universidad de Salamanca, Salamanca, Spain

{amelie.dorn, eveline.wandl-vogt}@oeaw.ac.at

yalemisew.abgaz@adaptcentre.ie, {abenito, theron}@usal.es

Abstract

The world's indigenous languages and related cultural knowledge are under considerable threat of diminishing given the increasing expansion of the use of standard languages, particularly through the wide-ranging pervasion of digital media and machine readable editions of electronic resources. There is thus a pressing need to preserve and breathe life into traditional data resources containing both valuable linguistic and cultural knowledge. In this paper we demonstrate on the example of an Austrian non-standard language resource (DBÖ/dbo@ema), how the combined application of semantic modelling of cultural concepts and visual exploration tools are key in unlocking the indigenous knowledge system, traditional world views and valuable cultural content contained within this rich resource. The original data collection questionnaires serve as a pilot case study and initial access point to the entire collection. Set within a Digital Humanities context, the collaborative methodological approach described here acts as a demonstrator for opening up traditional/non-standard language resources for cultural content exploration through computing, ultimately giving access to, re-circulating and preserving otherwise lost immaterial cultural heritage.

Keywords: indigenous languages, cultural conceptualisation, data visualisation, semantic data modelling

1. Introduction & Background

In today's digital age, dialects, much like indigenous languages, are under considerable threat of diminishing as standard languages pervade the public domain as a means of communication, particularly in Western societies. The global decrease in indigenous languages and also in dialects or regional varieties of languages poses a considerable risk to maintaining not only linguistic knowledge diversity, but also cultural diversity and ultimately mankind's heritage. With this in mind, efforts by UNESCO (2002) have been made to sustain and foster dialogue around cultural diversity, including linguistic diversity. Similarly, educational minority- and under-resourced language initiatives have received fresh impetus and support over the past years (cf. Jones & Ogilvie, 2013). Although globalisation and the increased use of digital media as a means of communication have brought about a surge in standardisation across different fields of life, advances in computational capacities may at the same time also be exploited for maintaining knowledge diversity. Methods such as semantic modelling, the enrichment with (Linguistic) Linked Open Data (LOD)[1] or a combination of different computational processing and linking methods may enable sustainability, longevity and ultimately re-use of otherwise forgotten resources, contributing to the documentation of existing and new formations of diverse and rich knowledge networks.

In this paper, we thus showcase the potentials semantic enrichment (Section 3) paired with visualisation tools (Section 4) offer in revealing and giving access to unique traditional cultural knowledge and cultural conceptualisation contained within a non-standard language resource on the example of the Bavarian dialects in Austria, Europe, containing data from 1200 up to now, focusing on the early 1900s.

In what follows, we present work in progress and a first glimpse into the cultural conceptualisation contained in our digital resource (DBÖ/dbo@ema) (Wandl-Vogt, 2010), by looking at the original data collection questionnaires, which constituted the starting point of the collection at the time and which we therefore also take as our initial methodological case study. Our approach to tackling such resource with the methodology presented below is unique, and with our endeavour, we hope to serve as a demonstrator for other language resources of similar composition of which there are many around the world.

2. Non-standard Language Resource: the exploreAT!-case study DBÖ/dbo@ema

The questionnaire data described in this paper is part of a larger data collection, the Database of Bavarian Dialects in Austria [*Datenbank der bairischen Mundarten in Österreich*] (DBÖ) and related dbo@ema (Wandl-Vogt, 2010). The databases contain digitised data from questionnaires, related answers, as well as digitised entries from vernacular dictionaries and folklore literature. Apart from standard and non-standard German, the entries and dictionary excerpts in the database also dip into other languages such as Hungarian, Slovak, Slovene or Serbian, to name a few.

The questionnaire data we deal with here constitutes only a fraction of all data contained in the databases. It pertains to a dictionary project (WBÖ, 1970-), which aimed at capturing the German language spoken by local people from the early 20th century onwards in the area of the former Austro-Hungarian empire. Originally in analogue form (see Figure 1), the information from questionnaires and related answers (3.6 million individual digital entries) has undergone several stages of digitisation and is now available in XML/TEI formats (Schopper, Bowers & Wandl-Vogt, 2015). Apart from being a rich linguistic

[1] http://lod-cloud.net/

resource, the data also contain a wealth of historic cultural information of everyday life, e.g. customs, religious festivities, food, traditional medicine, professions, songs, etc. (cf. Wandl-Vogt, 2008). In addition, detailed information on persons (n=11,157) involved across several stages of data collection or processing is also available (cf. Piringer, Wandl-Vogt, Abgaz & Lejtovicz, 2017) as well as detailed geographical and location information (cf. Scholz, Hrastnig & Wandl-Vogt, 2018).

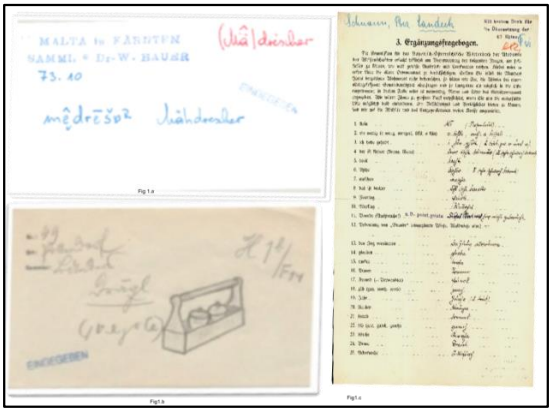


Figure 1: Left panels: examples of answer paper slips containing words, drawings, pronunciation, location and collector information. Right panel: an example of original analogue questionnaire and questions.

In the current Digital Humanities project *exploreAT! - exploring austria's culture through the language glass* (cf. Wandl-Vogt, Kieslinger, O'Connor & Therón, 2015), the cohort of these digitised data is now computationally processed from the aspects of cultural lexicography, semantic modelling, visualisation analysis and citizen science to access the traditional cultural knowledge and shed light on the knowledge system of the former society.

The questionnaires and their questions constitute the former starting point of data collection and are thus a key aspect in shaping the cultural content of the entire collection. Where projects with databases of similar content tend to entirely focus on the linguistic analysis of collected answers, we consider the questionnaires as an essential conceptual access layer to the collection. For this reason, we first aim to unlock the cultural concepts contained in the questionnaire questions, to extend the exploration to the remainder of the data in a second step.
The 120 questionnaires dealt with here thus concern three types: (1) Systematic Questionnaires [*Systematische Fragebogen*] (n=109), (2) Additional Questionnaires [*Ergänzungsfragebogen*] (n=9) and (3) Dialectographic questionnaires of the Munich and Vienna Dictionary Commissions [*Mundartgeographischer Fragebogen der Münchner und Wiener Wörterbuchkommissionen*] (n=2). Across these 120 questionnaires, we count a total of 24,382 questions asking for linguistic or cultural information or a combination of the two. The three questionnaire sets differ from one another according to form, content and purpose. In what follows we describe the application of semantic technologies (Section 3) as a first step in unlocking the cultural concepts.

3. Semantic Modelling of Cultural Knowledge Systems

For accessing the cultural content information in the questionnaires, the application of semantic modelling methods is essential. In general, various models have been designed to supplement semantics to original language resources (e.g. Chiarcos, Cimiano, Declerck & McCrae, 2013). These models don't only provide the meaning required to understand and correctly interpret these resources, but they also provide tools and techniques to effectively exploit their semantic information. To capture the semantics of the questionnaire questions, we followed a bottom-up approach (Noy & McGuinness, 2001). First, the original data collection methods, then the different types of questionnaires and questions were identified. This allowed gaining insights into the original approach taken and interpretation of the data. The questionnaires allow to aggregate and separate the resources based on the similarity of topics they address. The current semantic model (see Figure 2) captures the three types of questionnaires (systematic, additional and dialectographic questionnaires) based on the nature of the questionnaires and the type of information sought.

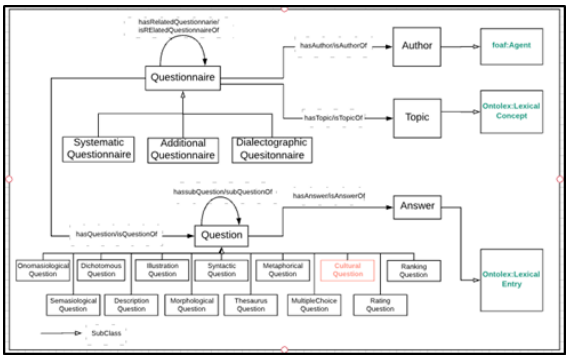


Figure 2: An initial semantic model of the questionnaires.

In addition, the model also captures authors and topics of individual questionnaires, where topics function as a means of aggregating the answers. Across the questionnaires, a total of 14 different types of linguistic questions were identified by relevant words or abbreviations contained in the questions themselves, including naming, definition, morphology, phonology, syntax, synonyms, etc. Cultural questions, on the other hand, contain significant information on representing and preserving the cultural identity of the communities and their language and were identified according to topics such as food, traditional medicine, games, songs etc., see Figure 3. In addition to the structure, we further capture patterns and examples of cultural questions which will later serve for the characterisation of questions of similar domains elsewhere.

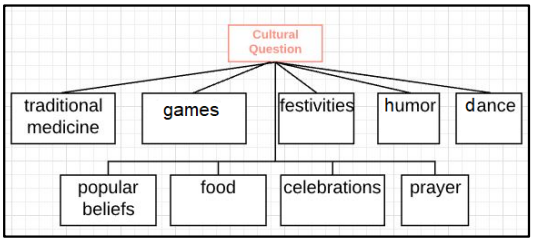


Figure 3: A data model of cultural questions.

Our model captures the original intent of the collection and provides a stable foundation for future interlinking with a variety of other resources, e.g. Europeana[2], Wikidata[3], Babelnet[4], Germanet[5] or relevant sources in the context of the co-creation approaches with user communities such as, for example, Topothek[6], Community Cooking[7] or Gastrosophie[8]. It further enables exploration of the data based on similar topics and structures and provides unique perspectives to navigate the entire collection by exploiting guided navigation to the answers of the questions. In addition, our semantic model provides the structure for the next stage of our data visualisation tools based on their semantic similarities.

4. Visual Discovery of Indigenous Cultural Knowledge Design via Concept Lights v.1.0

Data visualisation has become a key component of the Digital Humanities in recent years (cf. Benito, Losada, Therón, Dorn, Seltmann & Wandl-Vogt, 2016). By combining human-computer interaction techniques, psychology and graphic design, it can bring great insights to humanistic questions of the kind we discuss in this paper. In this line we developed the tool *Konzeptlichter/ Concept lights*[9] (Figure 4) that supports the visual exploration of the questionnaires introduced in Section 2.

The tool plays with lights and shadows to help the expert user to form a mental image of the structure of a single questionnaire by displaying common term associations and their disposition in the corpus. The foundation of the proposed linked-view system is an adjacency matrix representation, showing how many coincident terms are shared between different questions. These coincidences were detected and extracted in a previous data import stage where graph data structures were generated. Also in this step, questions were cleaned and stripped off stop words, leaving only semantically meaningful terms in each question, ready to be visualised. For example, the original question "Gesicht: Gesichtsrose, Rotlauf und andere Erkrankungen" is condensed as "Rotlauf, Gesichtsrose, Erkrankungen" in our approach. If any other question refers to the same terms, we consider these as semantically coinciding and those other concepts accompanying the matching terms are conceptually close and therefore relevant for the type of visual exploration we propose.

Two questions matching in two or more terms are considered a significant group. In turn, we use these groups to enable the exploration of the questionnaires. In Figure 4, groups are represented as coloured circles at the bottom left of the visualisation. Hovering over these groups, the terms are projected onto the matrix by illuminating the questions containing the terms.

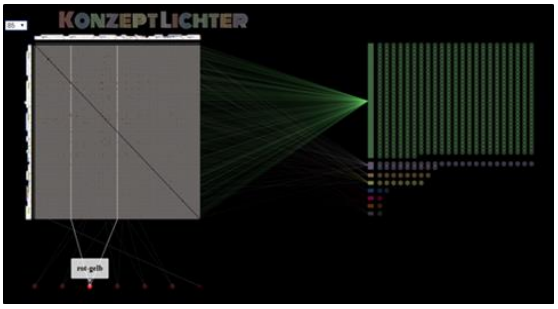


Figure 4: Konzeptlichter / Concept lights v1.0: Visual exploration prototype of words of single questionnaires. Selection of individual questionnaires (left) with the display of concept terms from questions (right).

Another way of exploring the questionnaire is enabled by the individual concepts found in the questions (see Figure 4). On the right part of the visualization, different terms found in the questionnaire are sorted by increasing order of importance. Terms found less often are placed at the top (green circles, appearing once, more abundant) whereas more common ones are moved to the bottom. (other colours, appearing two or more times, less abundant). An easy way of exploring the occurrence of terms within a questionnaire was enabled by applying the same highlighting effect described before when hovering over the individual circles representing the concepts. We also employ a magnifying effect that allows highlighting of all terms inside a group at once. These two annexe views are connected to the matrix by two-way channels, i.e. the highlighting effects also occur when selecting specific cells (questions) in the adjacency matrix. Whereas this tool is intended for the expert lexicographer with previous working experience with the questionnaires, we are designing new interaction paths tailored for the novel user that we expect to present in future research.

[2] https://www.europeana.eu/portal/ [accessed: 06.03.2018]
[3] https://www.wikidata.org/ [accessed: 06.03.2018]
[4] babelnet.org/ [accessed 06.03.2018
[5] www.sfs.uni-tuebingen.de/GermaNet/ [accessed: 06.03.2018]
[6] www.topothek.at/de [accessed: 06.03.2018]
[7] https://www.caritas-wien.at/stadtteilarbeit/ aktuelleprojekte/ community-cooking/[accessed: 06.03.2018]
[8] www.gastrosophie.at/ [accessed 06.03.2018]
[9] concept-lights.herokuapp.com [accessed: 06.03.2018]

5. Discussion & Future Work

Next steps in developing our semantic model contain digging deeper into the food domain as a case study for cultural conceptualisation: together with exploreAT!-community-groups (experts in various areas and lay-persons) a domain specific data-model for a thesaurus is co-designed, developed and evaluated.

Concept Lights v.2.0 aims to offer summarized insight into all questionnaires (not just one by one), incorporating the ontology model outlined in Section 3, which will in turn allow the proper contextualization of the displayed concepts and the retrieval of relevant content from the Semantic Web.

Concluding, exploreAT! aims to experiment with analogue and similar collection questionnaires to improve data modelling, visualisation as well as contextualisation of cultural and linguistic diversity as well as biodiversity and contribute to foster awareness about its wealth and its accessibility and documentation.

9. Acknowledgements

This research is funded by the Nationalstiftung of the Austrian Academy of Sciences under the funding scheme: Digitales kulturelles Erbe, No. DH2014/22. as part of the exploreAT! project, carried out in a collaboration with the VisUSAL Group, Universidad de Salamanca, Spain and the ADAPT Centre for Digital Content Technology at Dublin City University which is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is cofunded under the European Regional Development Fund.

6. Bibliographical References

Benito, A., Losada, A., Therón, R., Dorn, A., Seltmann, M. & Wandl-Vogt, E. (2016) "A spatio-temporal visual analysis tool for historical dictionaries." Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM). ACM. pp. 985-990. doi:10.1145/3012430.3012636

Chiarcos, C., Cimiano, P., Declerck, T. & McCrae, J. P. (2013). Linguistic Linked Open Data (LLOD). Introduction and Overview. In C. Chiarcos, P. Cimiano, T. Declerck & J. P. McCrae (Eds.), 2nd Workshop on Linked Data in Linguistics. Representing and Linking Lexicons, Terminologies and Other Language Data. Pisa, Italy, 23rd September 2013. Retrieved January, 17, 2018: http://www.aclweb.org/anthology/W13-5501.pdf

Jones, M.C. & Ogilvie, S. (Eds.) (2013) Keeping Languages Alive: Documentation, Pedagogy and Revitalization.CUP.

Noy, N. F., & McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Technical, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

Piringer, B.,Wandl-Vogt,E., Abgaz, Y., & Lejtovicz, K. (2017) Exploring and exploiting biographical and prosopographical information as common access layer for heterogeneous data facilitating inclusive, gender-symmetric research. In Wandl-Vogt, E. & Lejtovicz, K. Biographical Data in a Digital World 2017. A conference in the framework of the project APIS, 6–7 November 2017. Abstracts. [Wien]. doi:10.5281/zenodo.1041978

Scholz, J., Hrastnig, E. & Wandl-Vogt, E. (2018). A Spatio-Temporal Linked Data Representation for Modeling Spatio-Temporal Dialect Data. In P. Fogliaroni, A. Ballatore & E. Clementini (Eds.), Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017), (Lecture Notes in Geoinformation and Cartography (LNGC)). Basel: Springer International Publishing, pp. 275–282.

Schopper, D., Bowers, J. & Wandl-Vogt, E. (2015). dboe@TEI: remodelling a database of dialects into a rich LOD resource. Retrieved January 17, 2018 from Text Encoding Initiative. Conference and members' meeting 2015. October 28-31, Lyon, France. Papers: http://tei2015.huma-num.fr/en/papers/#146

UNESCO (2002) Universal Declaration on Cultural Diversity: a vision, a conceptual platform, a pool of ideas for implementation, a new paradigm. Cultural Diversity series, Vol.1 http://unesdoc.unesco.org/images /0012/001271/127162e.pdf [last access: 19.01.2018]

Wandl-Vogt, E., Kieslinger, B., O'Connor, A. & Theron, R. (2015). exploreAT! Perspektiven einer Transformation am Beispiel eines lexikographischen Jahrhundertprojekts. In DHd2015. Von Daten zu Erkenntnissen. 23. bis 27. Februar 2015, Graz. Book of Abstracts.

Wandl-Vogt, E. (2008): Wie man ein Jahrhundertprojekt zeitgemäß hält: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I Dinamlex) (mit 10 Abbildungen). In: Ernst, Peter (Eds.): Bausteine zur Wissenschaftsgeschichte von Dialektologie/ germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der internationalen Gesellschaft für Dialektologie des Deutschen, Wien, 20.-23. September 2006. Wien. pp. 93-112.

[WBÖ] Wörterbuch der bairischen Mundarten in Österreich (1970–). [Dictionary of Bavarian Dialects in Austria] Bayerisches Wörterbuch: I. Österreich. Ed. by Österreichische Akademie der Wissenschaften. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

7. Language Resource References

[DBÖ] Österreichische Akademie der Wissenschaften. (1993–). Datenbank der bairischen Mundarten in Österreich [*Database of Bavarian Dialects in Austria*] (DBÖ). Wien. [Processing status: 2018.01.]

[dbo@ema] Wandl-Vogt, E. (2010; Ed.). Datenbank der bairischen Mundarten in Österreich electronically mapped [*Database of the Bavarian Dialects in Austria electronically mapped*] (dbo@ema). Wien. [Processing status: 2018.01.] https://wboe.oeaw.ac.at/dboe/indices/

A Rule-Based System for the Transcription of Sanskrit from the Devanagari Orthography to the International Phonetic Alphabet

Aalok Sathe
University of Richmond
28 Westhampton Way, Richmond, VA 23173, USA
aalok.sathe@richmond.edu

Abstract

We propose a new system for the transcription of Sanskrit text written using the Devanagari orthography, into the International Phonetic Alphabet, and supplement it with free and open-source software. We make use of existing literature on closest known pronunciations of sounds as well as prosodic and metric rules of syllabification using the Weerasinghe-Wasala-Gamage (WWG) algorithm for Sinhala, adapted to Sanskrit. We further incorporate suprasegmental sound changes along with the assignment of syllable-weight-determined stress.

Keywords: Transcription, Sanskrit, IPA, Phonetics, WWG algorithm, Devanagari, Computational linguistics

1. Introduction

The language Sanskrit is one of the oldest classical languages, and has a large amount of literature. For this reason, Sanskrit is a topic of frequent study in literature, culture, and linguistics. One hurdle in this process of studying is often the lack of an all-encompassing system of phonetic transcription. Whereas the IAST[1] and ITRANS[2] are widely used today, and will continue to be, they are but alternate means of representation of the same text, not fully capturing the phonological and prosodic features of the language. Additionally, it is our personal experience that even though systems such as IAST exist, students of Sanskrit worldwide have varying pronunciations of the same few sounds, seemingly approximated to the inventory of their primary languages. For new learners or even existing scholars, there can be a steep learning curve in Sanskrit phonology. Hence, it would be beneficial to new learners to have a system that would guide pronunciation as accurately and consistently as possible.

Some newer tools seemingly try to address this issue. However, they either do not solve the problem at hand, or do so inaccurately. Examples include the 'ICU' system for transliteration of Indic scripts (Viswanadha, 2002) as well as the website "Ashtangayoga" (Steiner, nd). In the case of the latter, we notice a lack of any indication of stress, syllabification, as well as that of within-word and suprasegmental phonological phenomena whatsoever. One may disregard these as 'superficial' details, but they are far from being that as syllabification and stress play an important role in classical Sanskrit poetic composition. We propose an improved system, hence, which we hope will serve as a convenient tool for reference in the study of Sanskrit phonology. We describe a system for the transcription of Sanskrit text written using its Devanagari orthography into the international phonetic alphabet (IPA). We choose IPA in particular to enable near-completeness of representation of the best-known pronunciations of Sanskrit sounds, rule-based syllabification adapted for Sanskrit from the Weerasinghe-Wasala-Gamage algorithm ('WWG algorithm') originally developed for the Sinhala language, and syllable stress: a prosodic feature not captured in any modern transcription system (e.g., IAST).

In this work, we aim to develop, based on existing work, a rule-based algorithmic system, and a computer program to supplement it, which will provide a consistent transcription given well-formed[3] Sanskrit text. We develop and distribute software accompanying this system, and license it using the GNU General Public License 3, to enable anyone to access and redistribute the source code as well as develop other software with the current implementation at its base. We believe this software will in itself be a tool for preservation of traditional knowledge as well as help create newer ones.

2. Sanskrit Phonology

Sanskrit is a classical language with its origins in the Indian subcontinent, and its literature and texts being found in present-day India, Nepal, and neighboring regions. Sanskrit is one of the official languages of India and shares close common ancestry with most of the modern Indo-Aryan languages spoken in the Indian subcontinent today (Emeneau, 1956) as well as some of the older Indo-European languages. It was recorded as the mother-tongue of about 14,000 people in the 2001 census of India (Banthia, 2001). While the effective pronunciations of Sanskrit sounds differ from region to region depending on the speaker's own mother tongue and regional linguistic influence, a unified approximation of Sanskrit sounds has been proposed in several existing works based on historical as well as present-day phonetic studies. In what follows, we attempt to give a summary of Sanskrit speech sounds.

In Sanskrit, there are multiple singular vowel sounds, as well as diphthongs made by combinations of individual vowels. The simple vowels are shown in table 1, and the diphthongs in 2. All of these vowels, whether simple or compound (diphthong), may be considered as a whole unit in Sanskrit for the purpose of prosodic analysis. Table 1 also shows the vowel length, which must accordingly be considered during transcription. Diphthongs are long vowels in Sanskrit. In addition, Sanskrit uses certain approximants and semivowels and treats them in the general category of vowels. These are

[1] International Alphabet of Sanskrit Transliteration

[2] Indian languages TRANSliteration

[3] That is, one adhering to the rules of classical Sanskrit phonology and Devanagari orthography

	Front	Central	Back
High	इ ([i]), ई ([iː])		उ ([u]), ऊ ([uː])
Mid	ϕ, ए ([eː])	अ ([ə]), ϕ	ϕ, ओ ([oː])
Low		ϕ, आ ([ɑː])	

Table 1: Sanskrit speech sounds: simple vowels. Symbols on the left are short variants of the vowel, while those on the right are long. In case a variant of a vowel does not exist, 'ϕ' is shown. Blanks denote vowels not in the Sanskrit phoneme inventory.

X	X+इ,ए	X+उ,ओ
अ,आ	ऐ ([ɑːi])	औ ([ɑːu])

Table 2: Sanskrit speech sounds: simplified rules of diphthong formation

ऋ ([ɻ]),	ॠ ([ɻː])
ऌ ([l̩]),	ॡ ([l̩ː])

Table 3: Sanskrit speech sounds: special vowels (sonorants). The first row shows short and long syllabic alveolar approximant sounds, respectively, while similarly, the second row shows short and long lateral approximant ones.

shown in table 3. For simplicity, we will consider all of these as vowels making up a single unit, just the way we do with "regular" vowels. Now, vowel length will be the only additional consideration other than identity, for the purposes of transcription.

We will base our transcription system upon existing literature on the phonology of Sanskrit (Jamison, 2004) as well as a system of correspondences between Devanagari text, IAST, and IPA, in the work 'The Original Pronunciation of Sanskrit' (Zieba and Stiehl, 2002). We will hence establish a mapping between Devanagari glyphs, their diacritic combinations if any, and IPA symbols (Association, 1999). Table 4 shows the correspondences used for consonants and other non-vowel sounds, while table 5 is the vowel and vowel-like sounds' counterpart. Sanskrit makes use of special symbols for several compound consonants, which we will process using their constituent components. Fortunately, Unicode character combinations for Devanagari define such compound characters in terms of their constituent components by default, making them easier to process. Although Sanskrit has many complex phonological processes where sounds interact (a popular one of which is *sandhi*), we need not encode rules of such phonetic interaction other than those implied by the orthography. This exclusion is because any phonological combination that occurs (such as from *sandhi*) results into a new phrase, which is written as-is in the orthography. It is expected of an input phrase to be well-formed, i.e., to not have any phonological inconsistencies per the rules of Sanskrit orthography. Given that this program will likely find use in transcription of existing texts, this should not be an issue in most cases.

We have taken the interpretation of sounds to be as close as possible to what is believed to have been the pronunciation in the classical Sanskrit era (Zieba and Stiehl, 2002; Jamison, 2004). One such noteworthy consideration is the differential pronunciations of a *visarga*, or the 'ः'-terminal sound. Today, the interpretation of the pronunciation of this sound is slowly shifting towards a new trend: duplicating the vowel sound of the previous syllable after ([h]). For instance, रविः would end as [-ihi] according this rule as opposed to [-h]. While this seems to be a rising trend, it did not always use to be so, and the sound was supposed to be simply a [h]-terminating one, without vowel duplication.

3. Rule-Based Transcription

In our program, we will use several one-pass processes to fully transcribe a given text in linear time. In what follows, we describe some of the orthographic intricacies that require special attention in the design of the program.

3.1. Shorthand for Nasalization

The Devanagari Sanskrit orthography has several ways to indicate the presence of a nasal sound. Presence of nasals in a word is semantic, unlike some languages where it may have a conditioned occurrence. Nasals may be one of six types: five, derived from the conventional place of articulation (velar, palatal, retroflex, dental, and labial), and the sixth, simply a nasalized articulation of any vowel. Conventionally, a nasal consonant is only explicitly written when a phrase ends, or if the upcoming character is a vowel.[4] In case the nasal sound is not explicitly shown, an *anusvara* is shown on the character preceding it, and the actual sound corresponding to it is inferred from the forthcoming sound at the time of reading. For instance, if a word ends in a nasal sound, and the word after it begins with a bilabial stop, then the nasal is inferred to be [m]. When a sound does not belong to any of the five places of articulation mentioned above (e.g., a fricative, or a vowel), it shall be called the sixth case, and in this case, the preceding vowel is nasalized, with no additional sound being added. For instance, in the word संस्कृत ([sə̃s.kɻ̩.t̪ə]), where the *anusvara*'s circumstance is not one of the five types mentioned. The specific nasal sound to be used is inferred based on the next sound, if one exists, or is taken to be [m], the bilabial nasal sound, by default.

3.2. Handling the Default Schwa

A consonant character in Devanagari Sanskrit, unless explicitly marked *halant* (i.e., a schwa-less "partial" sound marked using the diacritic ्), has an implied schwa. For instance, ग may be transcribed as [gə], while to yield [g], we would need to mark a lack of schwa as ग्. Removal of schwa is required when either explicitly marking a character *halant*,

[4] As classified in the several tables above. A vowel in the strict phonetic sense is not meant here.

	Vl. plosive	**Vl. aspirated plosive**	**Vd. plosive**	**Vd. aspirated plosive**	**Nasal**	**Approximant**	**Fricative**
Glottal							ह [ɦə]**
Velar	क [kə]	ख [k^{h}ə]	ग [gə]	घ [g^{h}ə]	ङ [ŋə]		
Palatal	च [t͡ɕə]	छ [t͡ɕ^{h}ə]	ज [d͡ʑə]	झ [d͡ʑ^{ɦ}ə]	ञ [ɲə]	य [jə]	श [ɕə]
Alveolar						र [ɹə] ल [lə]*	स [sə]
Retroflex	ट [ʈə]	ठ [ʈ^{h}ə]	ड [ɖə]	ढ [ɖ^{h}ə]	ण [ɳə]	ळ [ɭə]*	ष [ʂə]
Dental	त [t̪ə]	थ [t̪^{h}ə]	द [d̪ə]	ध [d̪^{h}ə]	न [nə]	व [ʋə]	
Labial	प [pə]	फ [p^{h}ə]	ब [bə]	भ [b^{h}ə]	म [mə]		

Table 4: Sanskrit speech sounds in Devanagari: consonants and non-vowel sounds. Merged cells indicate shared place of articulation. *Lateral approximants. **Voiced fricative.

Base	**Diacritic**	**IPA**	**Base**	**Diacritic**	**IPA**
अ		ə	आ	ा	ɑ:
इ	ि	i	ई	ी	i:
उ	ु	u	ऊ	ू	u:
ऋ	ृ	ɹ̩	ॠ	ॄ	ɹ̩:
ऌ	ॢ	l̩	ॡ	ॣ	l̩:
ए	े	e:	ऐ	ै	ɑ:i
ओ	ो	o:	औ	ौ	ɑ:u
अं	ं	əm	अः	ः	əh
ॐ		o:m			

Table 5: Sanskrit speech sounds: vowels and syllabic sounds.

or when combining it with another vowel, in which case, the vowel combination overrides the schwa. The way Devanagari diacritic combinations work in Unicode are from the point of view of typographic convenience. However, during transcription, we are required to explicitly remove the schwa, as demonstrated in the following example: गो = ग + ो is the way diacritic combination takes place in terms of Unicode characters. However, phonologically speaking, it is गो = ग + ् + ओ ([go:]), since we are removing the schwa and explicitly adding another vowel, instead of superficially dealing with diacritical marks. This needs to be taken care of during transcription, since, at the surface level, it is not explicit what underlying phonological process is taking place.

3.3. Syllabification

For syllabification, we implement the WWG algorithm (Weerasinghe et al., 2005) adapted to Sanskrit (Dasa, 2013). In the original study, the algorithm was developed to account for a majority of the Sinhalese vocabulary which has Sanskrit or Pali origins, as well as a large number of direct borrowings. In the same study, the authors note that the algorithm would be similarly applicable to Sanskrit with some modifications. As shown in algorithm 1, we use groups of vowel-consonant-vowel clusters (of the kind $V_B C_n C_{n-1} \dots C_2 C_1 V_A$, where $n \geq 1$) for syllabification. Note that a cluster is not a syllable unit, but simply a device to locate syllable boundaries. We apply rules based on the number of consonants in the middle consonant cluster, i.e., n. Based on this length, prosodic syllabification conventions, we mark the boundaries of the syllables. We reuse boundary vowels, so a vowel that was processed while considering the current cluster will be included again to spot the next cluster. We achieve this by keeping track of indices where clusters began and ended.

Algorithm 1: WWG Algorithm adapted to Sanskrit

Input: Sanskrit text to be syllabified
`initialize` scope at the beginning of text;
while *end of text not reached* **do**
  `move` to next $V_B\mathbb{C}V_A$, where $\mathbb{C}$ is a consonant cluster;
  if *length of cluster* $\mathbb{C} = 1$ **then**
    `mark` syllable break after V_B;
  else if *length of cluster* $\mathbb{C} = 2$ **then**
    `mark` syllable break after first C from left;
  else if *length of cluster* $\mathbb{C} = 3$ **then**
    if *third consonant from left* = र् or य् **or** *first and second consonants are stops* **then**
      `mark` syllable break after first C from left;
    else
      `mark` syllable break before first C from right;
  else
    if *first consonant from right* = र् or य् **then**
      `mark` syllable break before second C from right;
    else
      `mark` syllable break after least sonorous C;
end
Result: Syllabified Sanskrit text

3.3.1. Examples

In what follows, we provide some example Sanskrit words to demonstrate syllabification as carried out using algorithm 1. For ease of reading, we highlight the consonant cluster in consideration using boldface in the Devanagari text.

1. कृ**त्त**म् ([kɹ̩.t̪əm]) was split before [t̪] following the rule for a cluster of length one.

2. व**ल्क**लानि (['ʋəl.kə.lɑ:.ni]): here, the first two syllables have been demarcated from each other by splitting a consonant cluster of length two.

3. (a) मत्स्यः ([ˈmət̪.sjəh]): this cluster of length three has been split according to the rule that checks the presence of either र् or य्.
   (b) उक्त्वा ([ˈuk.t̪ʊɑː]) demonstrates the rule involving two stops. Here, क् and त्. Hence, we split it after the first stop from the left hand side.
   (c) कृत्स्नम् ([ˈkɹ̩t̪.snəm]) is useful to illustrate the 'else' condition when the conditions similar to those in 3(a) and 3(b) do not apply.
4. कार्त्स्न्यम् ([kɑːɹ̩t̪s.njəm]) contains a य्-terminal cluster of length more than three. We split it before the second consonant when scanning from the right.

3.4. Assigning Stress

Once we finish demarcating the syllables, we use traditional prosodic and metric rules to determine the syllables that should receive stress. In Sanskrit, a syllable is either 'light' (L) or 'heavy' (H) (Sridharan, 2005). A syllable may be considered to be light in the base case, which acquires the status of being heavy subject to meeting one or more of the following conditions.

1. Syllable contains a long vowel or diphthong
2. Syllable is nasal-terminated or has nasalized vowel
3. Syllable is stressed

The goal is to ensure that any syllable of the form $[C_{11}]V_1[C_{1n}...C_{13}]C_{12}[C_{21}...]V_2$ that results in a cluster of consonants because of the adjoining consonants of the next syllable (here, C_{21} and beyond), is heavy. If the syllable already satisfies at least one of the first two conditions above, it is already heavy. However, if not, we must use condition 3 and add stress to make it into a heavy one. For the sake of example, consider the syllables of the word कुरुक्षेत्र ([ku.ˈɹuk.ʂeː.t̪ɹə]). When taken independently, they have the weights L, L, H, L. However, when considered in the word, the character क्ष्, which is a compound consonant of क् + ष्, causes the previous non-heavy syllable (-[.ɹuk.]-) to end into a consonant cluster of consonants of adjoining syllables. It thus receive stress, and hence become heavy, making the weights of syllables L, H, H, L. The third syllable does not receive stress, even though the boundary of the syllable break after it, i.e., -[.t̪ɹə], contains a consonant cluster, due to having the long vowel े ([eː]), which satisfies the first condition.

4. Software

Prototype software developed as part of this work may be found at the following link: `https://github.com/aalok-sathe/sanskrit_IPA`. The program is written using Python3, primarily because of effortless inbuilt Unicode support. The program allows the user to transcribe text on-the-go using a command-prompt design. A command in the form: `transcribe` *text* may be used. The software can also read an input file externally and output it in a similarly named file. This may be especially useful for transcribing large texts. Specific implementations apart, the software has intuitively named methods and commented code that will allow anyone using it to build software on top. We observed a lack of permissively licensed software for this purpose, and would like to stress that the prototype program is free and open source software (FOSS) which may be used, modified, and redistributed by anybody in compliance with the GNU General Public License (version 3 or later). It is our hope that this licensing will encourage scrutiny, improvement, and further development in related research questions.

5. Future Work

We intend to evaluate the current work against hand-transcribed Sanskrit text. Evaluations will be hosted along with the source code. More work along similar lines will be required to create a set of tools to represent traditional knowledge in Sanskrit, as well as a large number of Indic languages. To begin with, systems need to be developed that will enable back-transcription from IPA to Devanagari, as well as all-way systems to transcribe consistently to most of the major ways of representing Sanskrit text today, such as ITRANS and IAST. Whereas developing such a system for Sanskrit is possible using rule-based decision procedures, it is not possible for most other modern Indic languages which rely largely on the speaker's cultural and experiential knowledge of the language for phonetic disambiguation. For such languages as Hindi and Marathi, statistical learning methods will need to be used in addition to rule-based systems to create transcription mechanisms that are accurate.

6. Acknowledgments

We are grateful for helpful comments by and discussion with Mukund Gokhale, Hema Kshirsagar, Dieter Gunkel, Shardul Chiplunkar, and Thomas Bonfiglio.

7. Bibliographical References

Association, I. P. (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. A Regents publication. Cambridge University Press.

Banthia, J. K. (2001). *Census of India, 2001*, volume 1. Controller of Publications.

Dasa, G. (2013). Sanskrit prosody: Syllabification with the WWG method. `sanskritstudio.wordpress.com/2013/09/13/`. Accessed: 2018-01-10.

Emeneau, M. B. (1956). India as a lingustic area. *Language*, 32(1):3–16.

Jamison, S. W. (2004). Sanskrit. *Cambridge Encyclopedia*, pages 673–699.

Sridharan, R. (2005). Sanskrit prosody, Pingala sutras and binary arithmetic. *Contributions to the History of Indian Mathematics, Hindustan Book Agency, Delhi*, pages 33–62.

Steiner, R. (n.d.). Transliteration tool. `https://www.ashtangayoga.info/sanskrit/transliteration/transliteration-tool/`. Accessed: 2018-01-10.

Viswanadha, R. (2002). Transliteration of Tamil and Other Indic Scripts. *Tamil Internet 2002*.

Weerasinghe, R., Wasala, A., and Gamage, K. (2005). A rule based syllabification algorithm for Sinhala. In *International Conference on Natural Language Processing*, pages 438–449. Springer.

Zieba, M. and Stiehl, U. (2002). The original pronunciation of Sanskrit. *Sanskritweb.net*.

Seeing the Heiltsuk Orthography from Font Encoding through to Unicode: A Case Study Using Convertextract

Aidan Pine, Mark Turin
National Research Council Canada, University of British Columbia
aidan.pine@nrc-cnrc.gc.ca, mark.turin@ubc.ca

Abstract

Across the world's languages and cultures, most writing systems predate the use of computers. In the early years of ICT, standards and protocols for encoding and rendering the majority of the world's writing systems were not in place. The opportunity to deploy less-commonly used orthographies in cross-platform digital contexts has steadily increased since Unicode became the most widely used encoding on the web in late 2007 (Davis, 2008). But what happens to resources that were developed before Unicode standards became widespread? While many tools have been created to address this problem and other issues related to transliteration and character level substitutions,[1] this paper describes the process undertaken for the Indigenous and endangered Heiltsuk (Wakashan) language, and outlines a tool (*Convertextract*) that was designed to convert not only plain text, but also Microsoft Office (pptx, xlsx, docx) documents with the goals of updating and upgrading pre-existing digital textual resources to Unicode standards, and thus preserving the knowledge they contain for both the present and the future.

Keywords: language revitalization, Unicode, font encoding

1. Introduction

With a focus on sustaining knowledge diversity in the digital age, we introduce a tool that was developed to help ensure the preservation, proper rendering and future use of the Heiltsuk writing system as new standards for compatibility emerge. The tool was developed for rapid implementation to facilitate ease of use with other font-encoded writing systems. For textual knowledge to be encoded and viewed in a digital medium, the writing system in which it is composed must be digitally legible on a range of operating systems. For years, the orthographies of many languages were not easily supported by computers, and many continue to be constrained by requiring the installation of customized fonts and other proprietary tools. Despite these challenges, members of the Heiltsuk community of Bella Bella, Western Canada, have adapted and used a variety of techniques for communicating in written form using the established orthography for their language, from font-encoded writing systems to sending images of written text and the use of ad-hoc transliterations. A new generation of field linguists are being trained to be cognizant of the challenges of dealing with legacy data and language-specific font encodings (cf. Bowern, 2015).

Now that most of the world's written languages are supported by Unicode standards, tools that convert text from earlier non-Unicode systems to Unicode standards are increasingly important for preserving and sustaining the knowledge that is recorded in earlier digital file formats. Most of the existing digital language resources in the Heiltsuk community are stored in Microsoft Office file formats using specific styling and formatting. *Convertextract* was designed to minimize the possibility for human error associated with re-typing such documents and to reduce the time burden of using a plain text transliterator and reformatting documents one at a time. With specific reference to its application in the Heiltsuk community, we demonstrate three implementations of *Convertextract* that while computationally and metaphorically simple, hold exciting potential for community impact and widespread uptake.

2. Background

2.1 Bits & Bytes

A core factor that led to the flourishing of the digital age was the development of a standardized way for the 0s and 1s (binary code) that interacted with computer hardware to be encoded into something legible by humans. In the early 1960's, the American Standard Code for Information Interchange (ASCII) was developed to achieve just that (Gorn, Bemer, & Green, 1963). ASCII prescribed that 01100001 would represent the character "a", 01100010 would represent the character "b", and so on.

An eventual if rather self-evident problem with ASCII, at least to those familiar with the level of global linguistic diversity and variation, is that since a Bit is a binary value, and because ASCII is limited to 7 Bits, only 128 (2^7) possible characters exist in an ASCII-type encoding system. Considering all of the characters that exist in the world's writing systems, the limitation of 128 characters ensured that support for non-English characters was not readily available within the ASCII system.

UTF-8, developed in the 1990's, would later become the most widely used encoding on the web. UTF-8 encodes up to 4 Bytes instead of 1, and following a restriction on the total possible combinations set in 2003, allows for 2^{21} (2,097,152) possible characters (Yergeau, 2003). The first 128 characters in UTF-8 are identical to ASCII, meaning that 01100001 still indicates "a" in UTF-8, just like it did in ASCII. But UTF-8 can also prescribe sequences such as: 11000100 10011111 which will be rendered as "ğ".

[1] Including, but not limited to : URoman https://www.isi.edu/~ulf/uroman.html, Epitran (Mortensen et. al, 2016), Chatino transliteration http://ruphus.com/chatino_transliteration/, Inuktitut Transcoder http://www.inuktitutcomputing.ca/Transcoder/, Either/Orth http://orth.nfshost.com/, Digital Linguistics Transliterator https://tools.digitallinguistics.io/transliterator/.

2.2 Hiłzaqv Language and Culture Mobilization Partnership

Following the signing of a Memorandum of Understanding in 2016 between the Heiltsuk Cultural Education Centre, the Bella Bella Community School and the First Nations and Endangered Languages Program at the University of British Columbia, the Hiłzaqv Language and Culture Mobilization Partnership <https://heiltsuk.arts.ubc.ca/> was established. The Hiłzaqv language is spoken by the Heiltsuk Nation whose traditional territory includes the administrative centre of Bella Bella in Northern British Columbia, Canada.

Despite being critically endangered with only 4.7% of the population classified as either fluent or semi-fluent speakers, the language has a deep history of community-led documentation (Carpenter et al., 2016) and a vibrant community of dedicated and accomplished learners, which comprise 11% of the total population according to the First Peoples' Heritage and Language Council of Canada (First People's Cultural Council (FPCC), 2014).

2.3 Interim Strategies

The orthography for the Heiltsuk language, designed by linguist John Rath, uses many characters that are not part of the standard ASCII character set. For example, vowels may indicate high tone (through an acute accent, as in **á**), resonants may be vocalic (marked with an underdot), carry a high tone (marked by an acute accent), or be glottalized (marked by an apostrophe or even a combination of these as in **ḿ**, **m̓̀**, **ṇ**, **ń** or **Ĺ**). An **ɬ** is used to represent a voiceless lateral fricative and **ƛ** is used to represent a lateral stop. In total, the orthography uses 44 different characters that lie outside of the standard ASCII character set. Before Unicode was in widespread use, Heiltsuk language users still needed ways to display characters like **ǧ**. Heiltsuk is not alone in this dilemma, and many language communities and "linguists have devised a variety of ingenious solutions" (Bird & Simons, 2002) to overcome similar challenges. We are aware of three strategies used by members of the Heiltsuk community to achieve this: textual images, transliteration, and font-encodings.

2.3.1 Textual Images

When the Heiltsuk language could not be easily represented digitally, some individuals resorted to taking photographs of hand-written text and sharing the image file with others. Such an approach—while creative, immediate and effective—precludes the possibility of leveraging any digital text-processing tools, and makes resulting communication somewhat cumbersome. Indeed, this approach was used and can still be seen in the welcome greeting of the Heiltsuk Cultural Education Centre's website www.hcec.ca. Using textual images is time intensive, error prone, not easily scalable, and contingent on reliable and robust internet speeds. These techniques avoid both the power and pitfalls of text-editors, choosing instead to engage with the web and social media directly by assembling and curating image files of hand-written text.

2.3.2 Transliteration

In some cases, when an orthography has only relatively few characters that are not available in ASCII, ad-hoc transliterations designed to be used exclusively in digital contexts have been created, such as the one developed by linguist John Rath for Heiltsuk. For example, if the only required character outside of ASCII is schwa (ə) in a given language, then a community linguist might opt to simply use @ in place of schwa. While this can be quite an effective interim solution provided that not too many additional characters are required, it results in a symbol like @ having more than one meaning, which can confuse speakers, digital text processing tools as well as language-independent software and search engines. Along with being visually ad-hoc, requiring additional learning, and being potentially visually jarring and confusing, such an approach also burdens speakers with having to familiarize themselves with two distinct if related writing systems.

2.3.3 Font Encoding

Another approach to representing unsupported characters before Unicode support was available involved the development of language-specific fonts, usually referred to as "font-encodings", "font-hacks" or "font-encoded orthographies." The Heiltsuk orthography has two distinct font encodings: Heiltsuk Duolos and Heiltsuk Times. These fonts were specifically designed to render the 44 non-ASCII/ISO 8859-1 characters needed by the Heiltsuk orthography. For example, the Heiltsuk Duolos font was created to deliberately disregard 8-bit encoding ISO 8859-1's stipulation that 10101001 should render as ©, and instead render this sequence as **ǧ**. In order for the characters to be viewed, the customized font must be correctly installed. Such a work-around allows almost any character to be represented regardless of the underlying encoding. While this strategy works well if both the author and reader have the font installed, one result is that the language cannot be mobilized on social media or through web applications. Without the required font installed, 10101001 will appear as © and be illegible to users. With the development of UTF-8 encoding, it is now possible to type **ǧ** and expect that it will render correctly in most mainstream fonts. With this development, there is no longer a need for font encodings, despite their having played an essential interim role for Heiltsuk and for many other languages around the world (cf. Hall, Ghimire & Newton 2009 on the Preeti font for Nepali).

3. Implementation in the Heiltsuk Community

Within the Heiltsuk community in Bella Bella and beyond, many text documents (in both txt and Microsoft Word docx formats), Excel spreadsheets, PowerPoint presentations and lessons had been composed using the Heiltsuk Duolos and Heiltsuk Times fonts. Textual images were also fairly widely used on social media platforms.

The first goal of the Hiłzaqv Language and Culture Mobilization Partnership was to develop a cross-platform Unicode input system and keyboard to replace the font-encoded fixes of Heiltsuk Duolos and Heiltsuk Times. Within a matter of minutes of the Heiltsuk Unicode

keyboard being released, community members began tweeting in the language, opening up a vibrant, online

Figure 1 Rory Housty tweeting in Heiltsuk.

digital space where the language could be shared, as seen above in Fig. 1.

All earlier Heiltsuk digital materials, however, remained in the non-Unicode Heiltsuk Duolos and Heiltsuk Times fonts, making them effectively un-readable to users without the fonts installed and much harder to share. *Convertextract* was used to convert over 70 megabytes of text files from Heiltsuk Duolos and Heiltsuk Times to Unicode, including eight PowerPoint presentations, an Excel spreadsheet dictionary containing 10,005 entries, and 103 Microsoft Word files including several books used by school teachers in the Bella Bella Community School Native Language Program. In total, these files contained 103,056 characters of text which, assuming a typing rate of 100 characters/minute would have taken over 17 hours to retype, not including the time it would take to re-format the files.

4. Convertextract: Implementation

Given the context described in Section 2.2 above, it was apparent that the Heiltsuk community could make good use of a tool to perform a series of character conversions to upgrade text to Unicode standards from either the ad-hoc transliteration strategy described in Section **Error! Reference source not found.** or from the font-encoding strategy described in Section 2.3.3. The following three sections describe implementations of *Convertextract* as of version 1.3.

4.1 Python CLI

The *Convertextract* command line tool is a MIT licensed Python library built from a fork of Dean Malmgren's Textract library.[2] Textract is a library that extracts text from a wide variety of different file formats. Leveraging this work, *Convertextract* performs a specific list of find/replace transformations on any source text, and saves a new converted file without altering the style formatting of the original document (font size, underlining, boldness etc).

As *Convertextract* expects to be converting from a 'hacked' font, it delivers the converted text in Times New Roman by default, although any other font may be specified. Out of the box, *Convertextract* currently supports three conversions: from Heiltsuk Duolos, Heiltsuk Times and Tsilhqot'in Duolos to Unicode. *Convertextract* also supports user-defined conversions which can be described in an Excel document passed as an argument to *Convertextract*. The correct ordering of each substitution is essential to producing the correct output. In order to prevent the incorrect sequencing of substitutions, they are ordered according to their length from longest to shortest.

The Python Library can be called either in a Python script or directly through the command line. Documentation on how to install and use the command line tool is available on the public repository[3]. As of version 1.3, plain text files (txt), Microsoft Word documents (docx), Excel spreadsheets (xlsx) and PowerPoint presentations (pptx) are all accepted file formats for *Convertextract*.

4.2 Web

As many potential users of *Convertextract* might not be familiar with command line interfaces, a web tool that accomplishes the same task was developed and released for Heiltsuk Duolos and Heiltsuk Times. The tool operates by allowing a user to either upload a file to be converted (and subsequently returning a converted file), or paste or type text directly into a text input box and then select a conversion. Text typed in the input box is directly and instantaneously transformed using the chosen conversion. The web UI has also been responsively designed for mobile use.

4.3 Chrome Extension

Convertextract has also been released as an extension for the Chrome web browser. This implementation allows the instantaneous conversion of published websites. While online publishing using font-encoded orthographies online is fairly uncommon (as font-encodings are not typically supported by web browsers), transliterations do sometimes appear online, as was common with the Preeti font for Nepali, and may need to be converted. The Chrome extension is likely to be most useful for converting text between multiple orthographies that are in circulation and use for the same language.[4]

4.4 Limitations

Convertextract has a number of requirements in order to work properly. First, conversions must be able to be applied independent of context. For example, consider a hypothetical orthography for a language in which the underlying vowel **ɔ** is represented as **ɔ** before rounded consonants, but as **ʌ** before unrounded consonants. As the distribution of **ʌ** is predictable, when designing a new orthography, a decision might be made to represent both as **@**. When, at a later date, both symbols become renderable in most fonts thanks to the widespread rollout of Unicode,

[2] https://github.com/deanmalmgren/textract.
[3] https://github.com/roedoejet/convertextract.
[4] As with the Inuktitut Web Page Transliterator http://www.inuktitutcomputing.ca/Transliteration/webpage/info.php?lang=en

the community wishes to use a tool like *Convertextract* to convert documents written using **@** to the correct and original orthography that uses **ʌ** and **ɔ**. Unfortunately, in the current version of *Convertextract*, no functionality exists that would support context-dependent conversions. Instead of using methods like regular expressions which can provide this functionality, substitutions are defined in spreadsheets. This was done for the ease of adding additional languages as described in Section 4.5 and because context-dependent conversions were not required by Heiltsuk.

At present, *Convertextract* is not able to convert documents that contain multiple languages or writing systems within the same document. Depending on demand, these features could be developed and included in future releases.

4.5 Adding support for other font encodings

Convertextract was designed to be easily extended to other font encodings without requiring a significant investment of time, energy or computational expertise. If, for example, it is known that a font-encoding like Heiltsuk Duolos uses © to represent **ǧ**, then that correspondence can be entered into a spreadsheet with © in one column and **ǧ** in the other. By including a relative path to that spreadsheet as an argument, *Convertextract* will order and perform the same character substitutions using that spreadsheet. In that way, adding support for more font encodings can be as simple as preparing a spreadsheet. Indeed, this is how support was implemented by the Tsilhqot'in National Government for the Tsilhqot'in Duolos font encoding. The same lookup table that is used to inform the command line tool can also be used to build the web implementation described in Section 4.2 or the Chrome extension converters described in Section 4.3. Additional Excel lookup tables for font encodings are welcome and may be submitted either as pull requests or by email.

As illustrated in Section 3 above, we hope that the combination of supporting popular Microsoft Office formats and easy integration of custom lookup tables will result in *Convertextract* being used by more communities to save time preserving and mobilizing their digital language resources in years to come.

5. Conclusion

We have introduced a tool designed to convert text, Word, Excel and PowerPoint documents composed in non-Unicode compliant, customized and now legacy fonts into Unicode without altering existing file formatting and styles. This tool, while technically uncomplicated, has the potential to greatly reduce the hours previously required of communities to manually re-type and re-style these files in order to preserve their digital resources and broaden access in the information age.

6. Acknowledgements

We thank John Nenniger for his work developing the Chrome extension and Jennifer Carpenter for her invaluable edits and suggestions for improvements to this paper. We are grateful for funding from SSHRC Partnership Grant #895-2012-1029 (PI: Marianne Ignace) and SSHRC Knowledge Synthesis Grant #421-2015-2076 (PI: Mark Turin).

7. References

Bird, S., & Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, 79(3), 557-582.

Bowern, C. (2015). *Linguistic fieldwork: A practical guide*. Springer.

Carpenter, J., Guerin, A., Kaczmarek, M., Lawson, G., Lawson, K., Nathan, L. P., & Turin, M. (2016). *Digital Access for Language and Culture in First Nations Communities*. Vancouver, BC.

Coulmas, F. (2003). Writing systems. *An introduction to their linguistic analysis*, 249-268.

Davis, M. (2008). Moving to Unicode 5.1. Retrieved January 8, 2018, from https://googleblog.blogspot.com/2008/05/moving-to-unicode-51.html

First People's Cultural Council (FPCC). (2014). *2014 Report on the Status of B.C. First Nations Languages*. Brentwood Bay, B.C.

Gorn, S., Bemer, R. W., & Green, J. (1963). *American standard code for information interchange. Communications of the ACM*, 6(8): 422–426.

Hall, P., Ghimire, G., & Newton, M. (2009). Why don't people use Nepali language software?. *Information Technologies & International Development*, 5(1): 65-79.

Jones, M. C., & Mooney, D. (2017). *Creating orthographies for endangered languages*. Cambridge University Press.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. (2016). Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3475-3484).

Pine, A, & Turin, M. (2017). Language Revitalization. *Oxford Research Encyclopedia of Linguisitics*. Oxford University Press.

Rath. J. C. (1985). *Ways of Writing*. Heiltsuk Cultural Education Centre.

Rath, J. C. (1981). *A practical Heiltsuk-English dictionary with a grammatical introduction (Vol. 2)*. National Museums of Canada.

Rath, J. C. (n.d.). *Elements of Heiltsuk Grammar*. Bella Bella: Heiltsuk Cultural Education Centre.

Yergeau, F. (2003). *UTF-8, a transformation format of ISO 10646*.

Classifying and Searching Resource-Poor Languages More Efficiently. Using the FastText Word Embeddings for the Aramaic Language Family.

Mathias Coeckelbergs
Université libre de Bruxelles
Avenue F.D. Roosevelt 50, 1000 Brussels, Belgium
mathias.coeckelbergs@gmail.com

Abstract

Within the context of the FastText initiative, pre-trained word embeddings have been made available for 294 language, based on the respective Wikipedia corpus for the particular languages. One of these languages is Aramaic, which is currently conceived of as endangered, since it is only spoken by a few minority groups in the Middle East. Nevertheless, this language has a rich history of culture and literature, being among others also the language of Jesus and the main language of the Syriac culture, which lives on today as a liturgical language in several church denominations in among others India, Syria and Iraq. This paper wants to provide first insights into the usefulness of these word embeddings to connect the separate parts of Aramaic culture, and to study them as one language with many facets and influence, a subject which hitherto has only seen separated scholarship along the lines of research questions limited to a specific time frame. Using some of the specific assets of the FastText algorithm, we show how traditional difficulties in bringing together the Aramaic literature from a computational perspective, such as limited training resources and significant lexical richness due to external influences throughout the centuries, can now accounted for.

Keywords: word embeddings, Aramaic, FastText, Out-Of-Vocabulary words

1. Introduction

In recent years, word embeddings have developed into a varied research field within natural language processing. The techniques, among which Word2Vec, developed by Mikolov et al. (2013) is the most widely used, are best known for their ability to derive word analogies from a strictly unsupervised machine learning approach. Stated in other words, this means that - as the standard example goes- when one subtracts the 'man' vector from the 'king' vector and adds the 'woman' vector, the resulting vector is found to be most similar to that of 'queen'.

Although word embeddings have proven their worth, an important problem is that a large corpus is needed to learn useful embeddings. Hence, many corpora do not suffice, certainly those containing historical corpora or other resource-poor languages. As machine learning techniques, which are important for information retrieval tasks among others, require numbers, traditionally resource-poor languages lack application value for these techniques, providing one of the main reasons why they are understudied regarding these new wave of methods prevalent within the digital humanities practice. Finding solutions to this problem has yet seen several proposals, including for example methods to (artificially) generate new sentences based on the limited amount which is readily available. For historical texts such approach is methodologically problematic, because, since most texts have a intricate redactional history, it is difficult to provide training data which is undoubtedly representative of a certain language phase.

In this article we show how the study of the Aramaic language in its broad sense -including the various contemporary dialects as well as the stages of its long history- can be leveraged by the FastText word embeddings created by the Facebook AI group. The reason for studying the Aramaic language is that it is currently, despite its rich tradition spanning about 3100 years, considered an endangered language (Naby, 2013). It is spoken throughout various communities in Iran, Iraqi Kurdistan, Syria and South Turkey, but the population who speaks it fluently is growing older, with fewer young people learning the language due to the prevalence of Arabic in the region (Sabar, 2002). Nevertheless, small communities are being formed in Israel and the Netherlands.

The main research question we treat in this article is how to be able to search and classify Aramaic documents more efficiently using the FastText word embeddings. As we will explain further, research into the Aramaic language is splintered according to the specific time period groups of researchers are interested in. Most of the documents available in the various branches of the language are mutually intelligible. FastText uses the Wikipedia articles available in Moderen Aramaic, being the sole resource of machine learning tools for the language to our knowledge. In this article, we show how relatively small amount of training data, and the severe amount of Out-Of-Vocabulary words, two main reasons why machine

learning for Aramaic is difficult, is resolved by the FastText approach.

2. The FastText project

The FastText project provides researchers with 300-dimensional word vectors for (currently) 294 languages, which includes Aramaic. The vectors are trained on the corresponding set of available Wikipedia pages for the respective languages.

The idea behind the use of pre-trained word vectors is that users no longer have to train the embeddings on the corpus which they want to model. This approach can have its drawbacks, for example for research into diachronic variation, for which it is useful to train word vectors on several language phases, after which the vectors can be used to measure semantic shifts across time. Of course, for diachronic time frames for which too few data exists, this approach is not applicable. On the other hand, the FastText approach allows users to start their exploration with salient word vectors for the language under scrutiny, so that corpora with insufficient data to use traditional machine learning methods on, can nevertheless be investigated.

The method of the FastText algorithm is -apart from one condition- exactly the same as that of the Word2Vec algorithm, developed by Mikolov et al. (2013), which brought word embeddings to the forefront of machine learning research, since it was first coined by Bengio et al. (2003). The sole aspect in which both algorithms differ is that Word2Vec takes words as basic entities for which the algorithm assigns vectorial representations, whereas the FastText algorithm does so for n-grams. The best known example for which Word2Vec has received its fame, is that it showed that the algorithm can recognise word analogies, without having any explicit semantic knowledge. The standard example goes that if one starts out from the vector king, subtracts the one for man and adds that for woman, the resulting vector is shown to be closed to that for queen.

A drawback of this approach, however, is that when a word is not encountered in the training phase, it is an Out-Of-Vocabulary (OOV) word, and will be assigned the null vector, because it cannot learn new (and reliable) word vectors in real-time (Chen et al., 2015). It still is useful within the context of the FastText algorithm to speak of an OOV word, because the n-grams are still evaluated within the context of a word (see also Wieting et al. (2016)). As the creators of FastText describe themselves, this allows the algorithm to infer similar meanings between morphemes (Bojanowski et al., 2016).

3. Brief History of Aramaic Languages

The Aramaic language has spanned a long history of contact with various other languages, leading to the development of a web of strongly related languages, which show mutual intelligibility with the exception of specific words, which represent the various external influences. The development of Aramaic languages, which leads up to today, spans about 3100 years, although discussion as to the date of the oldest fragments persists (Beyer 1986).

In summary, we can state that both the modern variety of Aramaic, as well as its many historical phases all have a large part in common, though many differences persist, most notably on the vocabulary level (Creason, 2008). In the next chapter we will show how applying the FastText word embeddings to Aramaic can provide a new point of view for discussing lexical differences and distinct influence on the Aramaic language family.

The contemporary form of Aramaic, denoted by the term Neo-Aramaic or Modern Aramaic, comprises two main forms, being Eastern and Western Aramaic, with the former being much more prominent than the latter. The western variety is today solely spoken in the vicinity of Maalouly, a Syrian city close to the border with Lebanon. Within its history more varieties of the western dialects are attested, but the eastern variety has produced more documents, since most of the literature is written in an eastern variety, with the most well-known text written in the Aramaic language commonly referred to as Syriac, written from the 4th till the 8th century. The literature in this phase of the Aramaic language is so extensive that it comprises about 90% of all Aramaic writings. This language is strongly connected to the varieties of Aramaic spoken most widely today, being Assyrian, Chaldean and Surayt/Turoyo Aramaic. Needless to say, the modern Aramaic on which the Fasttext algorithm was run, hence has strong connection to all of these languages. It has to be noted that the study of the Aramaic languages is done in a haphazard way, being that scholars tend to specialise in a specific area of Aramaic, leaving the comparison between all varieties of the language vastly understudied.

4. Applying FastText to Aramaic

4.1. Language phases characterised by OOV words

As we can conclude from the previous sections, the main reason why advanced classification methods such as word embeddings have not yet seen applications for low-

resource languages such as Aramaic, is because its different language phases rarely have enough training data to achieve salient computational models. Of course, for modern Aramaic more sources exist and can be produced because it is a modern language which is still spoken, lending the application to such resources as Wikipedia articles, on which the pre-trained model for Aramaic by FastText is based.

The algorithm has a particular way of dealing with Out-Of-Vocabulary (OOV) words, which will prove very useful for the purpose of dealing with Aramaic texts which are strongly related to the language used in Wikipedia, but which nevertheless can be considered as a separate dialect (other eastern Aramaic variants, and the historical Syriac). In regular machine learning tasks we would encounter the problem of a significant amount of OOV words, which in the case of word embeddings would either result in no vector (or null vector) corresponding to the OOV word, for which no example in context was presented during training, or in the creation of a random vector. Although this latter option makes sure that each word has a corresponding vector, properties related to the vectorial representation are no longer preserved, meaning that no semantic information, such as needed for the word analogies or words closest in meaning described above, can be derived from them (Joulin et al., 2016).

Since the FastText algorithm creates vectors for n-grams in stead of words, this solves the traditional problem of encountering words for which no vectorial representation was made during the training phase, since the algorithm does not deal with text on the word-level. Moreover, since the newly encountered words are given a vector presents a good estimation of its semantics, this provides an invaluable tool to discuss language variation among the Aramaic language family, and provides a novel viewpoint for issues regarding hapax legomena, words occurring only once in the corpus.

4.2. Assets and Drawbacks of the FastText Algorithm for Aramaic

Apart from the clear asset of dealing with OOV words, which is a general positive - and therefore applicable to all languages - difference of FastText, in comparison with other word embeddings algorithms, other points apply more specifically to the study of Aramaic using these word embeddings.

A first asset of this approach is that lexical and morphological corrections can be performed. Similar applications of word embeddings have previously been explored by Luong et al. (2013). Since Aramaic words are -like all other Semitic languages such as Arabic, Akkadian and Hebrew- based on three basic consonants, which constitute the root of the word and represents the core semantic meaning, this means that once the vowels are added to this basic root, the resulting vectorial representations for the different words derived from the same root will also lie close to each other. This solves an important problem occurring in the modeling of Semitic languages, namely to automatically infer the lexeme (or base root) of every word, including rare ones. As we have pointed out, this is a difficult task, certainly when during training phase we could not include a (clear) context. For example the root ܣܩܠ (SQL) means to take, whereas the derived noun ܫܘܩܠܐ (SWQL') means arrogance (taking to much). An assessment of the vocabulary of the corpus of a random set of Aramaic documents shows that these relationships are found. Concerning hapax legomena, words which occur only once in a given corpus, and comprise between 40-60% of the corpus according to Zipf's law, we find partly salient results. About 70% of hapax legomena in a Semitic resource contains a root which also occurs in other words, and which makes it likely to share enough n-grams with better-known words. A related drawback is that for unique roots, or for weak roots (which loose at least one characteristic consonant in most of the conjugations), will not achieve salient results. Possible ways for the future to counter this is to use a lemmatizer to discern the lexeme for weak roots.

Secondly, we find that many documents of potential historical importance have not yet been published and/or translated and edited. This is an important problem for Aramaic in particular, and historical languages in general. We currently lack the knowledge to know to which points of interest the unpublished text are relevant, because studying them manually requires too much time. However, the method of the FastText algorithm allows us to efficiently search a large amount of Aramaic texts, once they are digitised, based on the ability to query of semantic relatedness in a given collection of documents. This makes it possible to find the most relevant documents for a given query, which can then be further analysed, as previously explored by Levi et al. (2015).

The sole drawback we have discovered is that when we apply the FastText algorithm on place names, that it does not discover the fact that these substantives indicate a location. The same is true for personal names, making the identification of named entities difficult, also due to the fact that traditional methods of discerning named entitiess automatically, such as the fact that they start with a capital

letter, do not apply to Aramaic. Traditional word embeddings, which look at the context in which a word occurs, have a certain conception of the name indicating a person or a location, leaving the algorithm to discern a semantic correlation between for example Germany, Switzerland and France, on the sole basis of the words among which they tend to occur. This is one of the assets of taking words as the basic unit to assign vectorial representations, rather than n-grams. However, for modeling Aramaic we have a vast amount of named entities available through the Syriaca platform (Kiraz 2005), which through term matching allows to recognise the se named entities.

5. Conclusions and Areas for Further Research

As we have seen throughout this paper, the FastText algorithm has provided an important step towards the NLP treatment of the Aramaic language family. However, as with all unsupervised machine learning methods, it is difficult to discover salient methods of evaluating the algorithm, certainly if we want an objective standard to evaluate the quality of the in real-time created vectors for OOV words. Therefore, our main further research will involve evaluating the semantic relationships between words with work on the lexical semantics of Aramaic, which has already been done using traditional methods.

6. Bibliographical References

Bengio, Y., Ducharme R., Vincent P. and Janvin C. (2003). A neural probabilistic language model. *The Journal of Machine Learning Research* 3 (1): 1137-1155.

Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2016). Enriching word vectors with subword information. Retrieved from http://arxiv.org/abs/1607.04606.

Chen, X., Xu, L., Liu, Z., Sun, M. and Luan H. (2015). Joint learning of character and word embeddings. *Proceedings of the 24th International Joint Conference on Artificial Intelligence,* pp. 1236-1242.

Creason S. (2008). Aramaic. In R.D. Woodard (ed), *The Ancient Languages of Syria-Palestine and Arabia*. Cambridge University Press, 45-48.

Joulin, A., Grave, E., Bojanowski, P. and Mikolov, T. (2016). Bag of tricks for efficient text classification. Retrieved from http://arxiv.org/abs/1607.01759.

Kiraz, G.A. (2005). Computing the Syriac lexicon: historical notes and considerations for a future implementation. In A.D. Forbes and D.G.K. Taylor (eds.), *Perspectives in Syriac Linguistics I: Colloquia of the International Syriac Language Project*. Piscataway NJ, Gorgias Press, pp. 93-104.

Levi, O., Goldberg, Y. and Dagan, I. (2015). Improving distributional similarity with lessons learned form word embeddings. *Transactions of the Association for Computational Linguistics* 3(1), pp. 211-225.

Luong, T., Socher, R. and Manning C.D. (2013). Better word representations with recursive neural networks for morphology. *Proceedings of the 17th Conference on Computational Natural Language Learning,* pp. 104-113.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved from http://arxiv.org/abs/1301.3781.

Naby, E. (2013). From lingua franca to endangered language. The legal aspects of the preservation of Aramaic in Iraq. In J.A. Argenter and R. McKenna Brown (eds.), *Endangered Languages and Linguistics Rights on the Margins of Nations,* pp. 197-206.

Sabar, Y. (2002). A Jewish Neo-Aramaic dictionary: dialects of Amidya, Dihok, Nerwa and Zakho, Northwestern Iraq. Harrasowitz, Wiesbaden.

Wieting, J., Bansal, M., Gimpel, K. and Livescu, K. (2016). CHARAGRAM: embedding words and sentences via character n-grams. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1504-1515.*

Xu, P. and Fung P. (2013). Crosslingual language modeling for low-resource speech recognition. *IEEE Transactions on Audio, Speech and Language Processing* 21(6): 1134-1144.

7. Language Resource References

FastText (2016). Facebook AI Research Team, https://github.comfacebookresearch/fastText/

Digitizing National Cuisines: Cooking Recipes as Conceptual Graphs

Dmitri Dmitriev
Institute of Linguistic Studies, Russian Academy of Sciences
St Petersburg, Russia
dmitri@globbie.net

Abstract

The use-case of digitizing national cooking recipes through consistent use of a foundational semantic ontology is considered. GSL-based formal representations of recipes are stored in Knowdy, an open-source graph database. Cross-cultural sharing of knowledge of the underresourced languages via intermediate semantic framework with NLP adapters appears to be the most efficient strategy to preserve the unique cultural background of different geographies.

Keywords: conceptual graph, cooking recipe, Knowdy graph DB, GSL, NLP

1. National cuisine as cultural phenomenon

Preserving ethnic and cultural diversity in general becomes a great challenge nowadays similar to the task of preserving the biological diversity of our planet. The endangering factors include the ubiqitous spread of globalized digital media that are unlikely to make allowances for local national traditions. National cuisine is widely agreed to be a key cultural phenomenon that forms national identity.

The millions of combinations of tastes, aromas, techniques, as well as historical, religious, and cultural allusions incrementally form the heritage with references in literature, folklore, music - all that makes national cuisines unique and interconnected with other areas of human activities.

Fabio Parasecoli rightly applies a concept of "signifying networks" to national cuisines: "Each element in a culinary tradition is thus also part of several interconnected networks of meaning, practices, concepts and ideals; the full extent of its meaning and value cannot be grasped without analysing its interaction with other apparently unrelated domains. We can define these networks as "signifying" because they help us make sense of reality, allowing us to comprehend our cultural environment and to act within its rules and boundaries" (Parasecoli, 2005).

The notion of signifying networks allows us to model a graph of numerous components that could help us keep track of unique and shared features of national cuisines: "These signifiers therefore define local, regional or even national identities, and include ingredients, techniques, trade, location, time and media, all of which give rise to variations and, eventually, differences that are interpreted as national. Yet national cuisines remain a complicated part of the world of globalization (and, in the European context, of pan-European administration). Russia is one country where the broad array of influences on national cuisines is evident" (Smith, 2012).

As the world becomes smaller in terms of travelling and communication, we get a lot of opportunities to discover new cultural dimensions to ourselves. One can tell quite a lot about a national group just by trying its famous dishes. Foreign visitors are often quite keen to try local cuisine but might find it risky unless enough explanatory information is provided.

Sharing cooking recipes not only encompasses a list of plain ingredients and cooking directions but also the environment where food products grow. National cuisines involve a big number of factors that make the dishes special, including specific ways of whole food processing, the use of utensils, applying cooking techniques etc.

Thus the use case of digitizing cuisine is quite instructive for understanding the principles of present day cross-cultural knowledge exchange.

Today's big challenge is to encode this information into a proper digital form so that the data exchange can open doors to foreign tourists, boost economic ties, and bring cross-cultural communication to a much higher level.

2. Foundational ontologies and digital formalisms

Existing technologies of formalized knowledge representation fall into several groups of frameworks. These include Semantic Web approach in its original form of OWL and RDF. Online collections of interrelated datasets using Semantic Web instruments are known as Linked Data. Many digitizing projects are built around an idea of using some kind of foundational ontology that can be extended by knowledge engineers in a particular specific field of expertises. We shall consider the use-case of applying these tools to digitizing the national cuisines. A comprehensive overview of approaches to developing ontologies related to culinary area is given in (Ribeiro et al., 2006).

2.1 RDF

The dominant rationale of RDF is that "the Web is moving from having just human-readable information to being a world-wide network of cooperating processes. RDF provides a world-wide lingua franca for these processes"[1]. As its name suggests, RDF is a framework for expressing information about *resources* – primarily web documents and various entities. Its formalism is based around an idea of static classes and properties. The question naturally arises as to what extent it is sensible to treat a recipe as an entity rather than as a complex process with arguments, timings, nested complexity etc.

[1] https://www.w3.org/TR/rdf-concepts/

2.2 Schema.org Initiative

Schema.org is a collaborative community activity with a mission to create, maintain, and promote schemas for structured data on the Internet. This initiative aims at providing a standardized vocabulary for shared metadata of published web resources.

Web resources related to cooking can use the metadata fields of a Recipe class, maintained here: http://schema.org/Recipe. Let's consider an example of a recipe markup provided by Google.[2]

```
 <script type="application/ld+json">
    {
"@context": "http://schema.org/",
"@type": "Recipe",
"name": "Strawberry-Mango Mesclun Recipe",
"image": [
"https://example.com/photos/1x1/photo.jpg"
  ],
"author": {
"@type": "Person",
"name": "scoopnana"
},
"datePublished": "2008-03-03",
 "description": "Mango, strawberries, and
sweetened dried cranberries are a vibrant
addition to mixed greens tossed with an oil
and balsamic vinegar dressing.",
"aggregateRating": {
  "@type": "AggregateRating",
  "ratingValue": "5",
  "reviewCount": "52"
},
    "prepTime": "PT15M",
    "totalTime": "PT14M",
    "recipeYield": "12 servings",
    "nutrition": {
 "@type": "NutritionInformation",
 "servingSize": "1 bowl",
 "calories": "319 cal",
 "fatContent": "20.2 g"
    },
"recipeIngredient": [
  "1/2 cup sugar",
  "3/4 cup canola oil",
  "1 teaspoon salt",
  "1/4 cup balsamic vinegar",
  "8 cups mixed salad greens",
  "2 cups sweetened dried cranberries",
"1/2 pound fresh strawberries, quartered",
  "1 mango - peeled, seeded, and cubed",
  "1/2 cup chopped onion",
  "1 cup slivered almonds"
 ],
"recipeInstructions": "\n1. Place the
sugar, oil, salt, and vinegar in a jar with
a lid. Seal jar, and shake vigorously to
mix.\n2. In a large bowl, mix salad greens,
sweetened dried cranberries, strawberries,
mango, and onion. To serve, toss with
dressing and sprinkle with almonds."
    }
  </script>
```

Code listing 1: Linked Data Recipe representation using JSON format and Schema.org's vocabulary.

The vocabulary and format adopted by Schema.org is primarily oriented towards representing high level metadata of web documents. In our view, a properly normalized semantic graph requires much more explicit representation of concepts. Most string values in Schema.org's fields are natural texts requiring human cognitive interpretation. Such a text can not be directly read by digital systems without special NLP tools that for the most part are fairly error-prone. As a funny example, we tried to apply Google own's machine translation to a single string item from a list of ingredients of their example recipe:

'*fresh strawberries, quartered*'

English-to-Russian translation produces an utterly incorrect sense disambiguation:

'*свежая клубника, расквартированная*'

with 'quartered' used in a sense of 'lodging' as in "Our troops were quartered in Boston" instead of `cut in four.'

The lack of clearly defined roles of objects and explicit identification of methods makes the task of machine translation of a recipe a lot more difficult.

Another futuristic challenge for digital recipes is to foresee a scenario where a robotic machine could follow the instructions to cook the dish. By no means could a raw string text representation assist in this task.

Let's consider a typical description of an ingredient:

'*8 Granny Smith apples - peeled, cored and sliced*'

It is pretty obvious that this text string contains a lot of classifying information: a raw food ingredient as a class, a specific cultivar of apples, quantity in pieces, a list of methods to be applied to each piece in order to actually use the ingredient – removing the seed center, the skin, cutting an object into parts of a specific shape.

The formalism of Schema.org is not expressive enough to help us encode the cooking process with some sort of programming code. To properly digitize recipes for robotic cooking, we need to separate ingredient descriptions from logic.

2.3 Cognitive Modeling of a Recipe

The traditional way to give directions of a recipe is to start with things and operations that are needed to be executed first. This order and style of description is known as imperative or procedural. The declarative or functional

[2] https://developers.google.com/search/docs/data-types/recipe

style of describing the logic of a process usually starts from the top of the execution pyramid, the expected useful result that we wish to achieve.

Consider the following simplified representation of a classical Russian recipe of a mushroom soup. Arrows denote subprocesses (sometimes alternative) needed by the parent process.

Figure 1: Representing a cooking recipe as a hierarchy of processes.

An instructive classroom experiment described in (Sam et al., 2014) aimed at the intuitive construction of an ontology for cooking recipes. The cognitive challenge of this task looks hard straight from the outset when we even try to define what the 'recipe' is. "The term *recipe* has several contextual meanings. It can be defined in a general sense as a method to obtain a desired end. When used in the context of cooking, it is generally considered to be a set of instructions on preparing a culinary dish. As such, it could be viewed as an object with properties such as ingredients and time needed. Alternatively, it could be viewed as a *process*, which takes in some input, has a series of steps to be executed, and produces some output. The time taken to execute the steps and the utensils needed also help describe the recipe" (Sam et al. 2014).

Once faced with the task of explicit explanation of concepts that are mostly known to us from our daily life experience, we as human beings tend to come up with different semantic segmentations of the shared reality. To some people the answer to this problem lies in imposing as many global standards as possible. However, in our view it's pointless to try and enforce any homogeneity in our cognitive shaping of foundational ontologies. The possible way of cross-cultural semantic integration in our view is to promote the use of ontologies that are tightly coupled with natural language processing.

3. Knowdy Project

Since 2006, in order to develop an optimal formalism for expressing complex semantics of natural language, we've been conducting research advocating for the interlingua approach to cognitive analysis of data (Dmitriev, 2006). Our research program was densely coupled with production-ready software development. One of our latest projects is focused on graph data management.

3.1 Open source software tools

Knowdy[3] is an open source software project of our research and development group in St Petersburg aiming at developing an ultra-fast graph database that allows direct and efficient manipulation with conceptual graphs, bypassing any intermediate representations like SQL tables. The database engine is implemented in plain C and can be used both as a network service as well as a standalone library for the embedded environment.

After several years of research and development our team of data scientists came up with a custom data format for Knowdy DB called GSL (an acronym for Globbie Semantic Language). GSL is optimized for compact storage of conceptual graphs. It is used for data storage, message passing and information exchange. This format is not so excessively verbose as XML and even slightly more compact than JSON. The language takes some features of S-expressions of Lisp, but with major modification of semantics since one should keep in mind that graphs are not lists! Bracket notation in GSL has a special meaning, allowing users to express not only the multilevel grouping but also the CRUD operations within a database storage system.

3.2 Coding recipes as GSL graphs

In GSL notation a process is coded as a first class function that can be named or anonymous, supports inheritance from a base function, has arguments and subprocesses that can run in parallel. For the sake of saving space we'll limit ourselves to a couple of examples. The processes below describe some of the logic behind the above mentioned Russian mushroom soup recipe.

```
(proc prepare mushroom soup mix
     [_gloss {ru подготовка заправки
                грибного супа}]
     (base cooking by boiling)
   (arg cut-mushrooms
        {run prepare mushroom mix})
   (arg cut-potatoes
        {run prepare potato mix})
   (arg cut-onions
        {run prepare onion mix})
   {run _put
      [_gloss {ru Все ингредиенты выложить
                 в контейнер и перемешать.}]
      {obj _all}
      {into_loc container}})
```

[3] https://github.com/globbie/knowdy

```
(proc prepare mushroom mix
    [_gloss {ru подготовка
                грибной заправки}]
    (arg clean-mushrooms
         {run clean mushrooms})
    {run _cut
      [_gloss {ru Нарезать грибы
                  мелкими кубиками.}]
      {obj clean-mushrooms}
      {form slice {size 1.5 {unit cm}}}})
```

Code listing 2: GSL declarative descriptions of cooking processes.

4. Knowledge sharing via crowd-sourcing

4.1 Web technologies

What we'd like to offer to the community of linguists, anthropologists, knowledge engineers, and all other interested parties is a set of methodologies, digital formats, and software tools to help establish a collaborative platform for knowledge sharing. Surely, we are far from thinking that such complex formalisms like GSL semantic graphs should be directly used by local communities that wish to share their cultural heritage with the rest of the world. For this purpose a different class of web based authoring tools can be applied (Dmitri Dmitriev, 2014).

4.2 Natural language processing

We advocate for the wider use of natural language processing to convert the free text input provided by a user, eg. the cooking directions from the original language into a set of interrelated semantic propositions that can be generated on the fly and presented in a user-friendly graphical interface. The propositions can be restated in another natural language or the same original language but in a more generic way. If any of the automatically parsed propositions seem incorrect or ultimately wrong, one can either try to restate the original instruction or redirect this issue to a support team.

One of our current projects aims at producing a semantic transcription of the traditional cuisine of the Mari.[4] It is a semi-automated process in which the unique lexical items of Mari are mapped by the language experts to a semantic ontology, eg. an atomic lexical item **түрлаш** means 'crimping the edges of an unleavened dough to seal a pie with decorative patterns' (Yuadarov, 2009). All this information must be explicitly linked with the interlingual representation of the ontology concepts:

```
(proc sealing a pie with decorative patterns
   [_gloss {mari түрлаш}
            {ru защипывать узоры на пирогах
                из пресного теста}]
   (base sealing a pie
       (arg obj (class Unleavened rolled up
                       dough))
       (arg method (class  Decorative Pattern))))
```

Code listing 3: GSL mapping between lexical items and a foundational ontology

[4] https://en.wikipedia.org/wiki/Mari_people

5. Use cases and practical applications

Nowadays, the culinary topics are of vivid interest in social networks and mobile applications. Our GSL-based formalism was tested in various commercial and non-profit projects. In 2017, a database of more than 3,000 recipes for multicooking devices was compiled with Russian as a primary language, and 5-6 other local national languages (eg. Mari) being added currently.

6. Conclusion

The process of digitizing national cuisines remains an important and challenging task of modern civilization. Cross-cultural sharing of knowledge via intermediate semantic framework using NLP adapters appears to be the most efficient strategy to preserve the unique cultural background of different geographies.

7. Bibliographical References

Dmitriev, Dmitri. Semantic Interlingua for the Knowledge Base Creation // Electronic Government - Workshop and Poster Proceedings of the Fourth International EGOV Conference 2005, August 22-26, 2005, Copenhagen, Denmark. Schriftenreihe Informatik 13 Universitätsverlag Rudolf Trauner, Linz, Austria 2005, ISBN: 3-85487-830-3. P. 11—18.

Dmitriev, Dmitri. Web lexicography for and by non-tech-people / Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-resourced Languages SLTU-2014. St. Petersburg, Russia, 2014. P. 247-251. ISBN: 978-5-8088-0908-6.

Parasecoli, Fabio. Introduction to : Culinary cultures of Europe: identity, diversity and dialogue. Ed. Merkle, Kathrin; Goldstein, Darra; Mennell, Stephen; Council of Europe. Committee for Educational Research. Strasbourg: Council of Europe, 2005. ISBN: 9287157448.

Ribeiro R., Batista F., Pardal J.P., Mamede N.J., Pinto H.S. Cooking an Ontology. In: Euzenat J., Domingue J. (eds) Artificial Intelligence: Methodology, Systems, and Applications. AIMSA 2006. Lecture Notes in Computer Science, vol 4183. Springer, Berlin, Heidelberg. ISBN: 978-3-540-40930-4. P. 213-221.

Sam, Monica; Krisnadhi, Adila Alfa; Wang, Cong; Gallagher, John; Hitzler, Pascal. An Ontology Design Pattern for Cooking Recipes: Classroom Created // Proceedings of the 5th International Workshop on Ontology and Semantic Web Patterns (WOP 2014), CEUR Workshop Proceedings 1302. Aachen, Germany: CEUR-WS.org. P. 49-60.

Smith, Alison K. National Cuisines // *The Oxford Handbook of Food History.* Edited by Jeffrey M. Pilcher. Oxford, 2012. ISBN: 9780199729937.

Yuadarov, K. Mari Peasant Food. Yoshkar-Ola, 2009. [in Russian: Марийская крестьянская кухня]

Building Multilingual Parallel Corpora for Under-Resourced Languages Using Translated Fictional Texts

Amel Fraisse[1], Ronald Jenn[1], Shelley Fisher Fishkin[2]
[1]University of Lille (France), [2]Stanford University (USA)
{amel.fraisse, ronald.jenn}@univ-lille.fr, sfishkin@stanford.edu

Abstract

In this paper, we present an ongoing research project which consists in collecting all the translations worldwide of one fictional text in order to build multilingual parallel corpora for a large number of under-resourced languages. Building such corpora is vital to help preserve and expand language and traditional knowledge diversity. These corpora will be useful to handle under-resourced languages in a number of interconnected research fields such as computational linguistics, translation studies and corpus linguistics. Our project taps into a wealth of translated versions of a single fictional text spanning a period of over a century. It consists in collecting, digitizing, transcribing and aligning translations of this text. Our data collection process is fluid and collaborative. It is based on volunteer work from the scientific and scholarly communities, the power of the crowd and national libraries and archives. Our first experiment was conducted on the world-famous and well-traveled American novel "Adventures of Huckleberry Finn" by the American author Mark Twain. This paper reports on 10 parallel corpus that are now chapter aligned pairing English with Arabic, Basque, Bengali, Bulgarian, Dutch, Hungarian, Polish, Russian, Turkish and Ukrainian processed out of a total of 20 collected translations.

Keywords: under-resourced languages, parallel corpus, translated fictional text

1. Introduction

Out of the world's 6000+ languages only a small fraction, a dozen or so, currently enjoy the benefits of modern language technologies such as speech recognition or machine translation. A larger but still modest number, close to a hundred, have the so-called Basic Language Resource Kit (BLARK) : monolingual and bilingual corpora, machine-readable dictionaries, thesauri, part-of-speech taggers, morphological analyzers, parsers and the like (Krauwer, 2003; Arppe et al., 2016). This means that as mentioned by (Scannell, 2007) over 98% of world languages lack most, and usually all, of these language resources. Even for well-endowed languages, parallel corpora are a rare resource. And yet, there is great need for them. Parallel corpora are a valuable resource for linguistic research and natural language processing (NLP) applications. Such corpora are often used for testing new tools and methods in Statistical Machine Translation (SMT), where large amounts of aligned data are often used to learn word alignment models between two languages (Och and Ney, 2003). Building such corpora for endangered languages presupposes the existence of translated language materials in these languages, where there are mostly available in print and awaiting digitization. When translation or software localization does occur it is mostly into commercially important languages (Fraisse et al., 2009; Fraisse et al., 2012; Roukos et al., 1995; Koehn, 2005; Ziemski et al., 2016).

Multilingual online digital libraries and archival projects collect documents and make them available to a wide audience : the Wikisource project [1], an online digital library of free content textual sources, the Internet Archive project[2] building a digital library of Internet sites and other cultural artifacts in digital form such as books and audio records, or the Gutenberg project[3] offering over 56,000 free written and audio eBooks and especially older works for which copyright has expired in more than 50 under-resourced languages. Those ongoing projects have made and continue to make significant progress in the preservation of knowledge and language diversity. In this work, we present our research within the framework of the funded Global Huck project which consists in collecting all the translations worldwide of one fictional text in order to build multilingual parallel corpora for a large number of under-resourced languages. We conducted a first experiment on the novel "Adventures of Huckleberry Finn" by the American author Mark Twain (1885). This fictional text was chosen because we knew for a fact, thanks to previous scholarship, that it was translated early in many different languages worldwide and that continued interest in the novel throughout the 20th and well into the 21st centuries guaranteed that a great number of translations in sometimes unexpected under-resourced languages were available. What makes the translation of such a fictional text especially valuable for the construction of multilingual parallel corpora is that it uses everyday commonplace words and phrases to describe its actions and plot. It is therefore not confined to a specific domain although the novel does revolve around the universal topics of freedom, slavery, race relations, oppression, emancipation and violence, so many topics that account for its fame and popularity. Our main focus in this paper is the collection and construction of multilingual parallel corpora built thanks to this particular novel with a view to provide digital corpora that will eventually be turned into dictionaries, thesauri, lexicons, and other linguistic resources.

2. Related work

Over the last few years, there has been a growing interest and awareness among the scientific community and lo-

[1]https://wikisource.org
[2]https://archive.org
[3]https://www.gutenberg.org

cally among advocates of minority languages in sustaining and expanding the existing resources in endangered languages and digitizing them in order to preserve and promote knowledge and language diversity.

In particular in relation to parallel corpora for under-resourced languages, some research works focused on religious texts such as the Bible as a relevant source to compile massively parallel corpora (Resnik et al., 1999). This line of research, which entailed the compilation of many parallel corpora, has broken new ground and allowed computational linguistics to handle an important number of under-resourced languages. More recently a Bible corpus was created based on freely available resources with over 900 translations in over 830 language varieties (Mayer and Cysouw, 2014). In (Christodouloupoulos and Steedman, 2015), the authors built a massively parallel corpus based on 100 translations of the Bible, emphasizing difficulties in acquiring and processing the raw material.

Kevin Scanell (2007) focused on the creation of web-crawled corpora for many minority and under-resourced languages and the development of open NLP tools for these languages in collaboration with native speakers. In (Choudhary and Jha, 2014; Jha, 2010), the authors created a parallel aligned POS tagged corpora in 12 major Indian languages (including English) with Hindi as the source language in the domains of health and tourism.

For European languages, there is the JRC-Acquis parallel corpus (Steinberger et al., 2006), the first of the sentence-aligned and pre-processed corpora distributed by the European Commission. In its latest version, it comprised 22 languages, that is to say all of nowadays' 24 official EU languages except for Irish and Croatian.

There are also parallel corpora related to translated literary works (e.g. "Harry Potter", "Le Petit Prince", "Master i Margarita") or translations from the web, mostly available for a set of closely related languages (Cysouw and Walchli, 2007; Mayer and Cysouw, 2014). Most of these texts mainly concern well-endowed largely known languages.

3. The example of Mark Twain's text for under-resourced languages

Mark Twain's books are some of the most well-travelled texts on the planet. As the UNESCO Index Translationum[4] shows the American writer is ranked 15 in the top-50 of the most translated authors worldwide. His works have been translated into almost every language in which books are printed (Rodney, 1982) including under-resourced languages. The novel "Adventures of Huckleberry" (Twain, 1885) is one of the most commonly translated of his books. Rodney (1982) identified 375 translations in 54 different languages as of 1976. As UNESCO's Index Translationum suggests, hundreds of additional translations have been published in the four decades since Rodney completed his survey. Table 1 shows the scores of languages into which the book has been translated. The list includes Afrikaans, Albanian, Arabic, Assamese, Bengali, Bulgarian, Burmese, Catalan, Chinese, Chuvash, Czech, Danish, Dutch, Estonian, Farsi, Finnish, French, German, Georgian, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Kazakh, Korean, Kirghiz, Latvian, Lithuanian, Macedonian, Malay, Malayalam, Marathi, Norwegian, Oriya, Polish, Portuguese, Romanian, Russian, Serbo-Croatian, Sinhalese, Slovak, Slovenian, Spanish, Swedish, Tamil, Tatar, Telugu, Thai, Turkish, Ukranian, and Uzbek. In many of these languages, there have been multiple translations over time, reflecting different moments in history, and different ideological perspectives on the part of the translators or publishers, as well as different attitudes towards the US, childhood, minorities and minority dialects, race and racism, etc. Usually parallel corpora focus on very specific and specialized domains which can be efficient but also show limitations for machine translation. The advantage of using a work of fiction such as "Adventures of Huckleberry Finn", is that it uses a very broad vocabulary linked to every day life, which makes it a valuable asset for those languages that are currently lacking such computational resources.

[4] http://www.unesco.org/xtrans/

4. Copyright and Digitization

Copyright issues are one of the major challenges in digitizing works in print that are still under copyright. According to the Berne convention, the copyright duration is 50 years after the author's death, while local laws extend that duration to up to 70 years. Choosing a text such as "Adventures of Huckleberry Finn", first published in 1885 and so immediately popular that it was translated into many languages means that a range of versions is available in the public domain and therefore readily available for our research.

4.1. Collecting Mark Twain's translation in under-resourced languages

We started out by calling on the international community of Mark Twain scholars as well as Translation Studies scholars in order to identify existing translations in different languages. Those Twain scholars can be teachers of American studies and/or literature or work in another field but keep an interest in Mark Twain. A globalized and transnational approach to Mark Twain is currently trending within that community. There is a growing interest in how Mark Twain's ideas and texts were translated and interpreted in different languages and especially the rarer ones.

In addition to the bibliographical survey carried out by Rodney (1982), the Twain community provided us with a compiled list of additional references through, for example, field research at the UNESCO in Paris. The UNESCO has, for many years starting in the late 1920s early 1930s, carried out a yearly survey of translations around the world called the Index Translationum. Additional and even more recent translations of Mark Twain have been discovered within the framework of the Global Huck project. In the compiled list resulting from those different inputs, each item includes the title in the target language, the first year of publication, the name of the translator and the publisher, when available. Beside the numerous versions in well-endowed languages such as French, German, Italian and Spanish, the novel was translated into a large number of under-resourced languages (Table 1).

Languages			
1.Afrikaans	15.Farsi	29.Kirghiz	43.Sinhalese
2.Albanian	16.Finnish	30.Latvian	44.Slovak
3.Arabic	17.French	31.Lithuanian	45.Slovenian
4.Assamese	18.German	32.Macedonian	46.Spanish
5.Bengali	19.Georgian	33.Malay	47.Swedish
6.Bulgarian	20.Greek	34.Malayalam	48.Tamil
7.Burmese	21.Hebrew	35.Marathi	49.Tatar
8.Catalan	22.Hindi	36.Norwegian	50.Telugu
9.Chinese	23.Hungarian	37.Oriya	51.Thai
10.Chuvash	24.Indonesian	38.Polish	52.Turkish
11.Czech	25.Italian	39.Portuguese	53.Ukranian
12.Danish	26.Japanese	40.Romanian	54.Uzbek
13.Dutch	27.Kazakh	41.Russian	
14.Estonian	28.Korean	42.Serbo-Croatian	

Table 1: List of languages "Adventures of Huckleberry Finn" was translated into.

Using the title in the target languages, we crawled the web and mined online digital libraries and national archives in order to find the full texts. In some cases we came across the full online version that was in the public domain (provided by public institutions) in which case we downloaded them, whatever their format. When dealing with versions in pdf or epub format we converted them into text format that could later be processed. In other cases, such as Bengali for example, the digital version was in image format and could therefore not be processed as such. In this case we transcribed the text following an approach described in the next section of this paper. There were other instances when we knew of an existing version but it was not readily available online. In that case we turned to the national libraries and archives and asked them if they were willing to collaborate with us by digitizing their printed versions. Within the framework of this project, local institutions are crucial because they have the knowledge, the expertise and they help us determine the copyright status of the versions we deal with. This project therefore enhances language diversity by tapping into the local institutions of under-resourced languages.

4.2. A crowdsourcing approach for text transcription

Over the past few years, many crowdsourced transcription projects have been created in order to transcribe speech, typed or handwritten documents. A wide spectrum of languages, historical periods, and geographic areas are represented by this type of project. For example, the City Archive of Leuven[5] crowdsourced the transcription of more than 950,000 Dutch-language register pages from the Leuven court of Aldermen during the years 1362 to 1795. The Ancient Lives project (Williams et al., 2014) asks online volunteers to transcribe fragment of ancient Greek texts from a Papyri collection. The Rediscovering Indigenous Languages project[6] crowdsourced the transcription of historic word lists, records and other documents relating

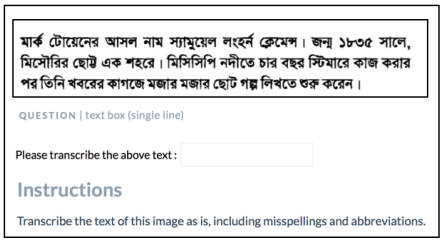


Figure 1: Example of the Bengali transcription task. The digitized image is at the top. Below it is the task instruction.

to indigenous Australian languages. As showed by several research works (Novotney and Callison-Burch, 2010; Gelas et al., 2011; Munyaradzi and Suleman, 2013), crowdsourced indigenous language transcription produces reliable transcriptions of high quality. We used CrowdFlower (Biewald, 2012) an enhanced service that feeds into Amazon's Mechanical Turk[7] and other crowdsourcing systems to transcribe digital versions that came as images, whether from local institutions or collected from the web. It provides convenient management tools that show the performance of workers for a task. The CrowdFlower User Interfaces (UIs) tend to fall into a set of tasks such us selection, categorization, text input or text transcription. The next step is the data import which may be uploaded as CSV or XLS files. Each page of each scanned translation represents a line in these data files which is associated to a task unit on CrowdFlower. The task asked workers to transcribe the text of one page as is, including misspellings and abbreviations. Figure 1 shows the example of the transcription task for the Bengali text.

5. The collected multilingual parallel corpora

In total, we collected digital translations in 20 under-resourced languages : Arabic, Basque, Bengali, Bulgar-

[5] http://itineranova.be/in/home
[6] https://transcripts.sl.nsw.gov.au
[7] http://www.mturk.com

English	Arabic	Basque
You don't know about me, without you have read a book by the name of "Adventures of Huckleberry Finn", but that ain't no matter. That book was made by Mr. Mark Twain, and he told the truth, mainly. There was things which he stretched, but mainly he told the truth. That is nothing. I never seen anybody but lied, one time or another, without it was Aunt Polly, or the widow, or maybe Mary. Aunt Polly—Tom's Aunt Polly, she is—and Mary, and the Widow Douglas, is all told about in that book—which is mostly a true book; with some stretchers, as I said before.	انك لن تعرفنى أيها القارىء الا انا كنت قد قرات كتابا بعنوان « توم سوير » ١١ ، وان كنت اعتقد الا اهمية لذلك ، فقد الف مسترر « مارك توين » هذا الكتاب وتصمنه جوهر الحقيقة ، ومعالله سمح لنفسه بان يبتدع بضع حقائق نسجها خياله ، فانه توخى الصدق بصفة عامة . وعلى اية حال ، فانى لا اذكر انى قابلت انسانا لم يكذب مرة او اخرى ، وليست استثنى من ذلك « الخالة بولى » أو « الأرملة دوجلاس » وربما « ماري » . ولقد ذكر « مارك توين » كل شيء عن « الخالة بولى » – وهى خالة توم – وماري ، والأرملة دوجلاس فى هذا الكتاب .. وهوكتاب صادق فى مجموعه مع بعض الجنوح الى الخيال كما قلت من قبل .	Ez duzu nire berri izango, baldin eta ez baduzu Tom Sawyerren abenturak izeneko liburua irakurri, baina lasai horregatik. Liburua Mark Twain jaunak idatzi zuen, eta hark egia esan zuen, gehienetan. Zenbait gauza berak asmatu zituen, baina ia beti egia esan zuen. Horrek ez du batere garrantzirik. Ez dut inor ezagutu inoiz gezurrik esan ez duenik, une batean edo bestean, izeba Polly ez baldin bazen, edo alarguna, edo beharbada Mary. Izeba Polly —Tomen izeba Polly, alegia— eta Mary, eta Douglas alargunari buruzko guztia liburu horretan esana dago eta gehien bat egia esaten digun liburua da, zenbait gehiegikeria baditu ere, lehentxeago esan bezalaxe.
Bengali	Dutch	Turkish
মার্ক টোয়েনের আসল নাম স্যামুয়েল লংহর্ন ক্লেমেন্স। জন্ম ১৮৩৫ সালে, মিসৌরির ছোট্ট এক শহরে। মিসিসিপি নদীতে চার বছর স্টিমারে কাজ করার পর তিনি খবরের কাগজে মজার মজার ছোট গল্প লিখতে শুরু করেন। মার্ক টোয়েন অনেক ছোটগল্প ও উপন্যাস লিখেছেন। 'এ কানেকটিকাট ইয়াঙ্কি ইন কিং আর্থারস কোর্ট', 'দ্য প্রিন্স এন্ড দ্য পপার', 'লাইফ অন দ্য মিসিসিপি' ইত্যাদি মার্ক টোয়েনের নামকরা উপন্যাস হলেও যে দুটো বই তাঁকে পৃথিবীখ্যাত লেখকের মর্যাদা দিয়েছে তা হচ্ছে, 'দ্যা এ্যাডভেঞ্চার অব টম সয়ার' এবং 'দ্যা এ্যাডভেঞ্চার অব হাকলবেরি ফিন'। মার্ক টোয়েনের নিজের ব্যবসা এ বই দুটিকে ফুটে উঠাতে তাঁর শৈশব জীবনের অনেক কাহিনী। মাত্র বার বছর বয়সে স্কুল ছেড়ে দেয়া মার্ক টোয়েন পরবর্তীতে আমেরিকা ও ইউরোপের বিশ্ববিদ্যালয় থেকে সম্মানসূচক ডিগ্রি পেয়েছেন। মার্ক টোয়েন আমেরিকার লেখক হলেও তাঁর পাঠক শুধু আমেরিকার ভূখন্ডে সীমিত থাকেনি। পৃথিবীর প্রতিটি দেশে রয়েছে তাঁর অসংখ্য গুণমুগ্ধ পাঠক। ১৯১০ সালে এই মহান লেখক মৃত্যুবরণ করেন।	Als gij het boek getiteld 'de lotgevallen van Tom Sawyer' niet gelezen hebt, weet gij niets van mij, maar dat doet er minder toe. Dat boek was geschreven door Mark Twain, en hij vertelde de waarheid, in hoofdzaak althans. Er waren wel dingen, die hij wat opsierde, maar in hoofdzaak vertelde hij de waarheid. Dat hindert niet. Ik heb nog nooit iemand gezien, die niet wel eens loog, of het moest tante Polly, of de weduwe, of des noods ook Mary geweest zijn. Over tante Polly - dat wil zeggen Tom's tante Polly - en over Mary en over de weduwe Douglas, over al die menschen is reeds gesproken in het boek - dat voor het grootste gedeelte een waar verhaal bevat; natuurlijk met eenige uitbreidingen, zooals ik reeds gezegd heb.	Tom Sawyer'ın Maceraları adlı kitabı okumadıysanız beni tanımazsınız, ama ziyanı yok. O kitabı Bay Mark Twain yazdı ve kendisi çoğu yerde doğruyu söylüyor. Kimi şeyleri abartmış, fakat söylediklerı doğru çoğunlukla. Ne çıkar ki zaten. Öyle ya da böyle yalan söylemeyen kimseyi tanımadım, Polly Teyze hariç, bir de dul bayan, belki bir de Mary. Polly Teyze –Tom'un Polly Teyzesi yani– ve Mary, bir de Dul Bayan Douglas'tan o kitapta bahsediliyor; demin söylediğim gibi, kitapta çoğunlukla doğru şeyler söylenmiş ama abartılmış yerler de var.

Table 2: Original text of the first paragraph of Chapter 1 of "Adventures of Huckleberry Finn" Finn and its translations in 5 under-resourced languages.

ian, Chinese (simplified), Chinese (traditional), Croatian, Czech, Dutch, Finnish, Hebrew, Hungarian, Polish, Portuguese, Romanian, Russian, Tamil, Turkish, Ukrainian and Vietnamese. The original version of the novel as well as collected translations are already structured by chapter. To perform corpus alignment, we kept the original structure by putting each chapter between an opening and closing chapter tag (<CHAPTER> </CHAPTER)). Even though we do not have a command of each and every target language, aligning the chapters was facilitated by the fact that all translations had the same number of chapters as the original English version (43 chapters) and each chapter starts on a new page and has a title of its own. In this work, we aligned 10 translations with the English source text : English-Arabic, English-Basque, English-Bengali, English-Bulgarian, English-Dutch, English-Hungarian, English-Polish, English-Russian, English-Turkish and English-Ukrainian. The remaining translations are under processing and will be included in a further version of the corpus. Table 2 shows as an example the original text of the first paragraph of chapter 1 of "Adventures of Huckleberry Finn" aligned with its translations in 5 under-resourced languages: Arabic, Basque and Bengali, Dutch and Turkish.

6. Expected use and availability

The major expected use will be to provide statistical machine translation systems with a rich parallel corpus for under-resourced languages. This current version of our corpus displays 10 under-resourced languages and others are being processed. One of our ultimate goals is to reach out to more less-resourced languages. Another goal is to engage scholars in the field of digital humanities as well as languages and Translation Studies specialists to address a number of fundamental questions. What happens in translation ? What is the impact of the linguistic and cultural transfer of the novel on its textual and iconic nature? An aligned digital corpus would allow them to evaluate the modifications and adaptations set up by translators and the translation process. A stable and reliable corpus will made available to them to conduct their own research. This research work will benefit academics both in their research and teaching activities in various areas of the humanities. More texts will be collected throughout the project's duration and thus the publicly available parallel corpus will be released. In the mean time, a first version of the corpus is available online and accessible on Github at the URL: `https://github.com/GlobalHuck/UnderResourcedLanguagesParallelCorpora/releases/tag/v1.0`

7. Conclusion and future works

In this paper we proposed and experimented a new language source to build multilingual parallel corpora for a large number of under-resourced languages. It consists in collecting all the translations worldwide of one fictional text by means of collaboration between volunteers, researchers, scholars, digital libraries and especially national archives, which are in charge of storing valuable traditional knowledge for future use. Our first experiment was conducted on the world-famous and well-traveled American novel "Adventures of Huckleberry Finn". This paper reports on 10 parallel corpus that are now chapter aligned pairing English with Arabic, Basque, Bengali, Bulgarian, Dutch, Hungarian, Polish, Russian, Turkish and Ukrainian processed out of a total of 20 collected translations that are being processed. More texts will be collected through the project duration.

Acknowledgments

This research work is conducted within the framework of the Global Huck project funded by the MESHS (Maison Européenne des Sciences de l'Homme et de la Société) in Lille, France. It is a partnership with Stanford University. We are also grateful to the Center for Mark Twain Studies in Elmira, N.Y.

8. Bibliographical references

Arppe, A., Lachler, J., Trosterud, T., Antonsen, L., and Moshagen, S. N. (2016). Basic language resource kits for endangered languages: A case study of plains cree. In *Proceedings of the the 2nd Workshop on Collaboration and Computing for Under-Resourced Languages Workshop (CCURL 2016)*, pages 1–8, Portorož, Slovenia, may 23.

Biewald, L. (2012). Massive multiplayer human computation for fun, money, and survival. *Current Trends in Web Engineering*, pages 171—-176.

Choudhary, N. and Jha, G. N. (2014). Creating multilingual parallel corpora in indian languages. In *Human Language Technology Challenges for Computer Science and Linguistics*, pages 527–537, Cham. Springer International Publishing.

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Cysouw, M. and Walchli, B. (2007). Parallel texts: Using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung STUF*, 60(2):95–99.

Fraisse, A., Boitet, C., Blanchon, H., and Bellynck, V. (2009). A solution for in context and collaborative localization of most commercial and free software. In *proceedings of the 4th Language and Technology Conference (LTC 2009)*, pages 536–540, Poznań, Poland., november 6-8.

Fraisse, A., Boitet, C., and Bellynck, V. (2012). An in context and collaborative software localisation model. In *proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, page 141–146, India, Mumbai., December 16-18.

Gelas, H., Abate, S., Besacier, L., and Pellegrino, F. (2011). Quality assessment of crowdsourcing transcriptions for african languages. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, pages 3065–3068, Florence, Italy, august 27-31.

Jha, G. N. (2010). The tdil program and the indian language corpora initiative (ilci). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 19-21.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit 2005*, september 12-16.

Krauwer, S. (2003). The basic language resource kit (blark) as the first milestone for the language resources roadmap. In *Proceedings of the International Workshop Speech and Computer*, Moscow, Russia, october.

Mayer, T. and Cysouw, M. (2014). Creating a massively parallel bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, may 26-31.

Munyaradzi, N. and Suleman, H. (2013). Quality assessment in crowdsourced indigenous language transcription. *Research and Advanced Technology for Digital Libraries.TPDL 2013. Lecture Notes in Computer Science*, 8092:13–22.

Novotney, S. and Callison-Burch, C. (2010). Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 207–215, Los Angeles, California, june 2-4.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Resnik, P., Olsen, M. B., and Mona, D. (1999). The bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1):129–153.

Rodney, R. M. (1982). *Mark Twain International: A Bibliography and Interpretation of his Wordwide Popularity*. Greenwood Press, Westport, CT.

Roukos, S., Graff, D., and Melamed, D. (1995). Hansard french/english. In *Philadelphia: Linguistic Data Consortium*.

Scannell, K. (2007). The crubadan project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, pages 5–15, Louvain-la-Neuve, Belgium.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., and Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 2142–2147, Genoa, Italy, may 24-26.

Twain, M. (1885). *Adventures of Huckleberry Finn*. Charles L. Webster and Company, Hartford, Connecticut.

Williams, A. C., Wallin, J. F., Yu, H., Perale, M., Carroll, H. D., Lamblin, A.-F., Fortson, L., Obbink, D., Lintott, C. J., and Brusuelas, J. H. (2014). A computational pipeline for crowdsourced transcriptions of ancient greek papyrus fragments. In *Proceedings of the International Conference on Big Data, IEEE Big Data 2014*, pages 100–105, Washington, United States, october 27-30.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, page 3530–3534, Portorož, Slovenia, may 24-26.

Sustaining Linguistic Diversity Through Human Language Technology : A Case Study for Hindi

Shweta Sinha, Shyam S Agrawal
KIIT College of Engineering, Gurugram, India
meshweta_sinha@rediffmail.com, ss_agrawal@gmail.com

Abstract

Language is a mean to communicate ideas, knowledge and express our cultural identity. To protect the legacy of our cultural heritage, language diversity needs to be sustained. Human language technology (HLT) can offer a lot to reduce the rate of language extinction. The focus of this paper is towards digital preservation of under-resourced languages. The discussion is apropos to the Indian languages; that almost all are under-resourced. The linguistic diversity of India is highlighted and its fate in this digital era is analyzed. This paper discusses the digital representation of the language and discusses HLT as a step towards preserving languages. Platform for online collection of speech is explained for gathering speech samples in three Indian languages; Hindi, Punjabi and Manipuri. The meta-data highlights the dialectal diversity of the speakers. These diversities have been analyzed acoustically for the Hindi speakers.

Keywords: Language Diversity, Hindi Language , Digital Preservation

1. Introduction

The United Nations Educational, Scientific and Cultural Organization defines that, cultural heritage is the legacy of physical artifacts and intangible attributes of a group or society that are inherited from the past generations, maintained in the present and vouchsafed for the benefit of future generations. The preservation of cultural heritage is not only concerned with safeguarding physical aspects of tradition but, is equally responsible for lesser physical aspect like languages, customs and beliefs. Ever since the existence of civilization the demand for communication exists. Language evolved as a sophisticated medium through which one can express thoughts that influences our society. With time, each culture evolved its own language and huge literary base for their language. In today's globalized world languages are disappearing at a very fast pace. On an average, every two weeks a language dies [1]. Language diversity is the basis of our rich cultural heritage and diversity. Of late, the loss of language diversity has grasped the attention of UNESCO also[2], as with the loss of any language, the memories and experiences of the culture are also lost. It is often observed that the positive impact of language on social, political and economical strata of the society influence the acceptance of the language in the society. Also, the colonial legacy on any country can burden the speakers and the native language with the use of exogenous language in formal and official domain. All these leads to cultural assimilation and usually results in the loss of suppressed language in years to come.

In this digital era, for the linguistic preservation and cultural redemption technology development and digital representation has become the sine qua non. Out of approximately six thousand languages of the world merely a fraction is digitally represented and efforts have to be made to reduce the exacerbation of digital divide. Hence, technology is essentially required for all and every language of the world in order to slow down its extinction. Human language technology development can offer a lot for reinvigorating and documenting any language. Till date these technological developments have been confined to the developed languages only. Technology development for any language can make life easy for its users and raise their interest for its use. Automatic speech recognition, speech to text synthesis and translation system based applications for any language can help in the growth of a language and facilitate access to textual and audio contents of the language.

Attention towards HLT is much needed for preserving under-resourced languages or other languages of the developing countries. This paper focuses on the efforts towards the preservation of some of the low resourced Indian languages, Hindi being the pivot of discussion. The paper describes unity in language diversity in context to India, section 3 presents the status of digital representation of Indian languages. Technology for sustaining the language diversity is explained in section 4. Technology development for Hindi language as a case is presented in section 5, and section 6 concludes the paper.

2. Unity in Language Diversity of India

India is a land of varied hues of culture, religion, race and languages. These variations account for the existence of different ethnic groups residing within the sanctum of one single nation. People of India speaks a large number of languages that can be divided into four families as: the Indo-European, Dravidian, Austro-Asiatic, and the Sino-Tibetan Family[10]. 73% of the Indian population speak one of the languages of Aryan group; a subgroup of Indo-European family[10]. Table 1 presents the language and speaker population of major Indian languages. Dravidian languages are spoken by 20% of the population and merely a small population speaks the languages from other two language families.

Sl. No	Name of Languages	Language Family	Speaking Population (millions)
1	Hindi	Indo-Aryan	422
2	Bengali	Indo-Aryan	83
3	Tamil	Dravidian	60.7
4	Marathi	Indo-Aryan	71.9
5	Telugu	Dravidian	74
6	Urdu	Indo-Aryan	51
7	Oriya	Indo-Aryan	33
8	Gujrati	Indo-Aryan	46
9	Punjabi	Indo-Aryan	29
10	Malyalam	Dravidian	33

Table1. A microcosm of linguistic diversity of India

In total, there are 122 major languages(spoken by more than 10K population), around 1600 distinct dialects along with 13 different scripts for writing. Out of all, Sanskrit is the most ancient language and is considered as the mother of most of the Indo-Aryan languages. It is the only language that transcended the region and boundaries of North and South India. Hindi the major language of India and has evolved from Sanskrit. Apart from this the country has developed highly sophisticated languages that mark India as a unique subcontinent that foster multiple cultures. The proverb **kos kos pər bədəlɛ pa:ni:, cha:r kos pər bədəlɛ va:Ni:**.explains that the Indian population thrive to the diversity. This can be translated as: every mile, the taste of water changes; and every four miles, the dialect changes". The unity in cultural and linguistic diversity of the country is very aptly conveyed through this proverb.

3. Digital Representation of Indian Languages

India, a land of multitudes has 30 languages that are spoken by more than a million native speakers. Many languages of the country have died in last few decades. The loss may be due to existence of fewer numbers of their native speakers, non-existence of any documentary evidence for the language, or response to new domains or media in that language. According to UNESCO's "Atlas of the world's languages in danger (2009)" [3] most of the Indian languages are vulnerable, i.e. they are mainly spoken inside the house and restricted in a particular domain. India has the largest number of endangered languages in the world[3]. Figure 1 represents the statistics of the languages of India. The categories mentioned are mainly based on the intergenerational transmission of the languages.

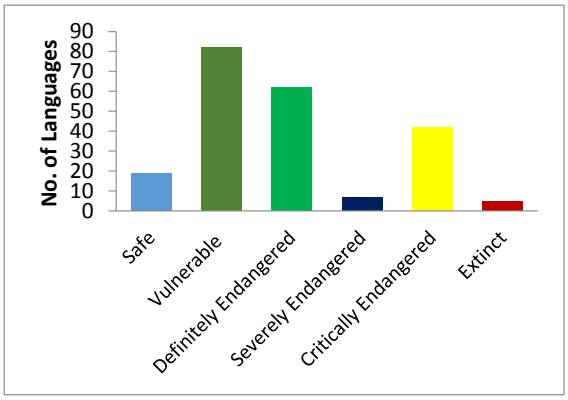


Figure 1. State of Indian Language Existence

One way to safe guard these languages is to provide more and more web resources in these languages along with technology development for human interaction. The internet user base has grown many folds in last few years. Survey by KPMG and Google show that there are 234 million Indian language internet users as compared to 175 million English internet users[4]. And the user base is expected to rise at an alarming 18% rate[4] which will generate demand for digital resources. Predicted internet user base by 2021 is represented in Figure 2 (Source: [4]). Three Indian languages, Hindi, Punjabi, and Bangla are among world's top 10 most widely spoken languages [5]. But, none of these find their place in the top ten languages on the web [6].In general the users of the web consider local language digital content to be more reliable.

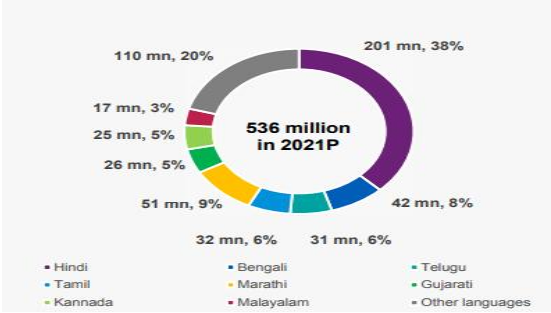


Figure 2. Internet users by native language

Limited language support either in terms of technology or content may force them to move towards English or other developed country's language. The under-representation of the Indian languages on the internet may leave them behind in the race of technological development for natural language processing techniques. Also the digital growth of the spoken languages by the means of language based applications will undoubtedly provide inclusive growth to the society and transform India into a digitally empowered country.

4. Technology Development : Efforts Towards Preservation of Language in Digital Era

Lack of digital data for the languages on the web categorizes them as under-represented language. Technology development is essentially required to revive, maintain, preserve and disseminate our traditional languages and in turn protect them from dying. Interface design for these languages will help human to communicate with computer and connect to the globalized world. Automatic speech recognition and text to speech are few of the HL techniques that have the prospects to completely alter the user's perspective for a language. Automatic language translation is the way to remove the human-to-human communication barrier. Applications based on these can make the life easier for the native language users and also document the language.

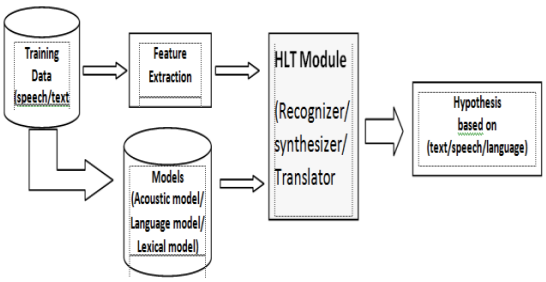


Figure 3. Architecture of HLT based system

5. Towards Technology building for Hindi : A Case Study

Till date HLTs are mainly concerned with the languages with large resources only. Indian languages with very limited digital resources lie far behind in this technological race. In a quest to protect our linguistic heritage from extinction we have initiated a small step towards development of HLT based system; to start with, ASR system for under-resourced languages. Due to lack of annotated text and speech data, nothing much has been achieved in the development of speech based applications for Indian languages. Preparing and gathering language resource is the biggest hurdle in the development of any HLT based system. High-quality ASR performance for the Indian languages is achievable only if real time data for these languages exists. For the growth of language technologies in context to Indian languages, it is essential to have corpus for speech as well as text representing pronunciation lexicon, language dictionaries etc.

Till date the efforts put forward has produced substantial data for Hindi and few other languages. The samples collected for Hindi has been acoustically analyzed for finding the dialectal diversities.

5.1 Data collection for under-resourced languages

The development of any of the HLT based application necessitates the collection of data for the concerned language to be fed as input to the system. Collection of data for the under-represented languages is a complex and tedious task. The situation is more critical when it comes to collecting recorded speech data. Data acquisition in terms of speech sample is very time and labour intensive task. Almost all the languages of India is an under-resourced language. Challenges for collecting spoken samples for these under-represented languages are many. Native speakers of under-represented languages either reside in rural areas or are distributed across a large geographical area. Portability is one of the major concerns for resource generation of these languages. Throughput, stability, latency and cost are the huge prohibitive factor for large-scale data collection. To collect speech samples from native speakers of languages residing in far-off locations we have developed a client-server based multi-lingual online speech collection system [7].

To start with, the data collection has been initiated for three Indian languages: Hindi, Punjabi and Manipuri. Hindi is the most commonly used inter-communication language. Even though this language has a large user base, due to variabilities in the use of this language its detail study with respect to articulation /pronunciation is essential.

The dialects of Hindi are categorized as the Eastern and the Western dialect. Punjabi is the 10th most widely spoken language of the world.It is spoken by people of Punjab region of Pakistan and India. Dialectal diversity exists for this language too. Manipuri, also known as Meitei is the predominant language of the Southeastern Himalayan state of Manipur. Apart from Manipur it is also spoken by people of Assam, Tripura, Bangladesh and Myanmar. This language belongs to Tibeto-Burman family. Approximately, 3 million people in the world speak this language.

Several recording specifications were set for the collection of samples. The collected database and its specifications for these languages have been summarized in Table 2.

Specifications	Indian Languages		
	Hindi	Punjabi	Manipuri
Speaker Registration	100	50	50
Male Speakers	68	27	25
Female Speakers	32	23	25
Dialects Covered	4	2	1
Sentences/Speaker	300	300	300
Isolated Utterances	200	200	200

Table 2. Corpus specification for Indian languages

5.2 Hindi Speech Sample Analysis

The major dialects of Eastern Hindi are Awadhi, Bagheli Bhojpuri and Chhattisgarhi and those of the Western Hindi dialects are Braj Bhasha, Haryanvi, Bundeli, Kannauji and Khari boli. Huge dialectal diversity exists among these varieties. Although there are about 18 classified dialects of Hindi [8]. The sample collected are predominantly from the speakers of these four dialects: Bhojpuri(BP), Bagheli (BG), Khariboli (KB) and Haryanvi(HR). Dialect influences individuals speaking style[9]. Insight into the phonological differences among the dialects can outline the factors that affect the acoustic properties.

5.2.1 Acoustic analysis of vowels in Hindi Dialects

Vowels are more often distorted than consonants in accented speech [9]. Hindi language has 10 vowels, that are categorized as 3 short vowels (/ə/, /I/, /ʊ/) and 7 long vowels (/ɑ:/, /i:/, /u:/, /e:/, /ɛ:/, /o/, /ɔ:/). To measure the dialectal influence on acoustic characteristics of these vowels duration, fundamental frequency, formants and intensity of these vowels were analyzed in reference to Hindi dialects.

The first and second formant analysis for the Hindi vowels w.r.t the four dialects outline that the second formant values for Bhojpuri dialect speakers are higher for back vowels (/ɑ:/, /ʊ/, /u:/, /o/, /ɔ:/) , for Bagheli speakers F2 is higher for all but /ɑ:/ . Haryanvi speakers and Khari Boli speakers have an approximately same value of second formant except for /ɑ:/ , where it is higher as compared to Khari boli speakers. It can be further observed that for the front vowels (/i:/, /I/, /e:/, /ɛ:/) F2 for Bhojpuri speakers are low compared to Khari boli speakers. F1 for Haryanvi speakers are high for close front vowel (/I/, /i:/). F2 value for all front vowels except

for open front vowel (/ɛ:/) is high for Bagheli speakers as compared to speakers of Khari boli dialect.

It has been further observed that prosodic features are more influenced due to dialectal influence. Duration is the highly affected prosodic feature that has been studied. Table 3 summarizes the findings of the acoustic analysis of Hindi samples. Also, to obtain the significance of dialects on the features ANOVA test was conducted on the feature parameters extracted from the samples.

Acoustic Feature	Vowels Influenced	Discrimination of Dialect
F0	Long Vowels (/ɑ:/ /i:/ /u:/ /ɛ:/ ɔ:/)	Significant for all dialect pair
F0	Short Vowels (/I/, /ʊ/, /e:/)	Not significant for any pair of dialect
F1	Back Vowels (/ɑ:/, /ʊ/, /u:/, /ɔ:/)	Significant for BG-KB dialect pair
F1	Front Vowels (/i:/, /I/, /e:/, /ɛ:/)	Significant for KB-BP dialect pair
F1	Close Front Vowels (/I/, /i:/)	Significant for KB-BP and KB-HR dialect pair
F2	Back Vowels (/ʊ/,/o/,/ɔ:/)	Significant for KB-BP dialect pair
F2	Front Vowels (/i:/, /I/, /e:/)	Significant for KB-BG pair
F3	All vowels except(/i:/,/I/,/ʊ/)	Significant for all dialect pair under study
Average Duration	All vowels at different word positions influenced due to dialect.	Significant for all dialect pair; exceptions: BP-BG not important for /ɔ:/; KB-BP not significant for /u:/
Intensity	No major distinction due to dialect	Influence selective in nature. Not significant for most of the vowels in almost all dialect pair under study.

Table 3. Summary of acoustic analysis of Hindi speech samples

5.2.2 Model Creation and automatic recognition of Speech

Based on the above analysis the features that are able to distinguish the utterances can be identified. The steps for the development of ASR requires Feature extraction, acoustic and language model and decoding techniques. These are the future work that has to be executed for the collected data.

Feature extraction techniques for the extraction of speech features have to be employed further. Using these features acoustic model need to be built. These can be Gaussian mixture models or Hidden Markov models . Recognition techniques need to be devised for the identification of utterances.

6. Conclusion

Language diversity in the world signifies the richness of our cultural heritage, To protect our culture and heritage the diversity of tongue has to be protected. Languages of developing countries are under-represented and low resourced languages of the world and are always in danger of getting lost in the coming days. The paper presents the status of Indian languages on the internet. Applications in the area of HLT has been shown as a way to protect the dying or the under resourced languages. A case study of Hindi is discussed in this paper to highlight our efforts towards sustaining the language diversity in this digital era.

7. References

[1] Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 56, 85-100.

[2] Unesco:http://www.unesco.org/new/en/indigenous-peoples/cultural-and-linguistic-diversity/[Accessed on: 24-12-17]

[3] Moseley, Christopher (ed.). 2010. Atlas of the World's Languages in Danger, 3rd edn. Paris, UNESCO Publishing. Online version: http:// www.unesco.org/ culture/en/endangeredlanguages/atlas

[4]https://assets.kpmg.com/content/dam/kpmg/in/pdf/ 2017/04/Indian-languages-Defining-Indias-Internet.pdf

[5] S. Arora, K. K. Arora, M. K. Roy, S. S. Agrawal, and B. Murthy, (2016). Collaborative speech data acquisition for under resourced languages through crowdsourcing," Procedia Computer Science, vol. 81, pp.37-44.

[6]Online : ttp://www.internetworldstats.com/stats7.htm/, [Accessed 5-March-2017].

[7] S Sinha, S Sharan, S S Agrawal,(2017). O-MARC: A multilingual online speech data acquisition for Indian languages, Oriental-COCOSDA , Nov 1-3, 2017, held at Seoul, S Korea.

[8] Arora, K. Arora, S. Agrawal, S. S. Paulsson, N. and Choukri, K. (2006). Experiences in Development of Hindi Speech Corpora based on ELDA standards, Oriental-COCOSDA 2006 held at Penang, Malaysia.

[9]. John C Wells. (1982) Accents of English, volume 1. CambridgeUniversity Press.

[10]Smriti Chand, http://www.yourarticlelibrary.com/ language/indian-languages-classification-of-indian-languages/19813

Convergent development of digital resources for West African Languages

Dorothee Beermann, Lars Hellan, Tormod Haugland
NTNU
N-7431 Trondheim, Norway
{dorothee.beermann@ntnu.no, lars.hellan@ntnu.no, tormod.haugland@gmail.com}

Abstract

We describe existing resources of the Kwa languages Akan and Ga, with a view to transfer of resources well developed for one to the other. While we can build on an Interlinear Glossed Text (IGT) corpus for Akan we have a modern digital lexicon for Ga, something we still lack for Akan, while we have only very limited IGT data for Ga. While it is normally the case that annotations from a resource rich language are transferred to a resource poor language, we are here preparing our resources to allow for a transfer approach between two resource-low but closely related languages. We envisage this to be a viable strategy also for other pairs of closely related under resourced languages.

Keywords: Akan, Ga, Interlinear Glossed Text, valence lexicon, morphological tagging, transfer learning between two resource-low but closely related languages.

1. Introduction

Akan and Ga are Kwa languages spoken in the southern and south-western parts of Ghana, and two of its official languages. Akan (ISO-639-3 "aka") is spoken by about 8 million native speakers according to the LDC[1]. The language has been studied extensively over many years (publications dating at least back to Christaller 1875, 1881), yet it still lacks most of the basic digital language resources, such as a lexicon, corpora, morphological analysers, and taggers. Ga (ISO-693-3 "gaa") is spoken mainly in the Accra area by about 745 000 speakers, according to Ethnologue[2]. It also has a literature dating back many years, starting with Rask (1828), and like Akan it lacks the basic digital resources, with one noteworthy exception, viz. a modern dictionary, compiled by Mary Esther Kropp Dakubu (Dakubu 2009), an authority in the study of West African languages and an expert of the language.

Having access to linguistic resources from two closely related Kwa languages, the line of research that we are interested in is driven by the question whether convergent development of closely related under-resourced languages, such as Akan and Ga, can create an opportunity to develop the basic digital resources for both languages more efficiently. In NLP, transfer learning is used as a methodology whereby resources from a resource rich language are transferred to a resource poor language. Can a similar approach be used whereby a digital resource from a poor resource language is transferred to a closely related resource poor language? In this paper we present our digital resources for Akan and Ga, which consist of an Interlinear Glossed Text (IGT) repository and a morphological tagger for Akan, and a digital valence lexicon for Ga, in the light of this question.

In section 2 we describe the curation of an Akan corpus and the development of a morphological tagger for the language. In both cases we combine community driven manual annotation with the automatic parsing of our IGT resources. In section 3 we describe, for Ga, the digitalization of a Toolbox lexicon and its conversion to a valence lexicon. We consider the learning from lexical data in the context of the semi-automatic valence annotation of Ga and eventually also of Akan. One of our long-term goals is to advance parsing for Akan using Ga resources, and the use of automatic annotation procedures for a more efficient enlargement of our West African IGT corpora.

2. Akan

Our Akan corpus consists of 102 IGT-style annotated texts, mostly linguistic sentence collections and small transcribed oral narratives. The corpus was created using a collaborative approach. G*raduate students* were asked to work on class projects which involved the morpho-syntactic *annotation* of their native language.

TypeCraft Akan resources	Words	Phrases
TypeCraft owned Akan resources	28 429	2689
TypeCraft hosted Akan resources	96 697	7535

Table 1: Snapshot of the TypeCraft Akan corpora

1 Linguistic Data Consortium, https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/west-african-languages.pdf (accessed 21.01.2018)

2 Ethnologue, https://www.ethnologue.com/language/gaa (accessed 21.01.2018)

For our work we used the *TypeCraft* research tool,[3] which contains two different sub-corpora for Akan (see Table 1), one sub-corpus consisting of 7535 phrases annotated by native speaker students of linguistics, and one TypeCraft-owned corpus consisting of 2689 phrases, which builds on the Akan data that was hosted at TypeCraft. In the case of the former, we were granted the necessary permissions by the owners (thus, for instance, for graduate work at our University, the students' consent was sought for use of their work in further research). In order to systematize our work with the TypeCraft owned corpus, we in 2016 started an Akan corpus curation project which in its first phase undertook the manual re-annotation of the TypeCraft-owned Akan data. At the same time, we started to enlarge that corpus more systematically. The Akan data hosted on TypeCraft was not affected by the effort, as that data is owned by individual TypeCraft users. The curation effort was accompanied by phonetic studies of Akan Tone.[4] In the project's first phase we re-annotated 2689 phrases manually. We followed a community approach receiving the help of

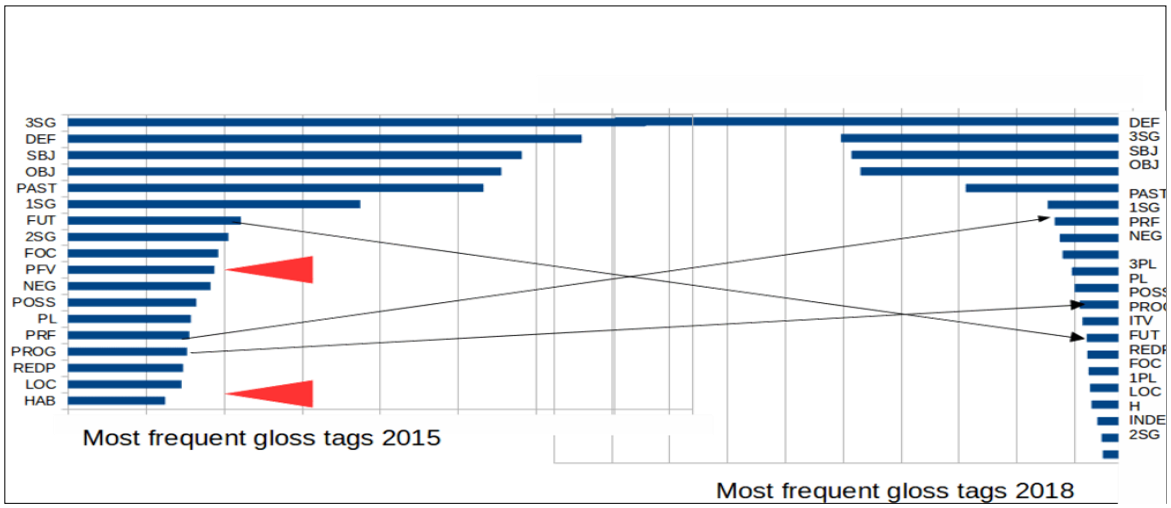


Figure 1: Comparison of the Gloss annotations for the 2015 and 2018 version of the TypeCraft Akan corpus

the Ghanaian Student Association at our university as well as the support of Akan speaking senior linguists. In addition, we hired Akan students as part-time data analysts.
To monitor conciseness and consistency of our annotations, we continuously used Annotation Profiling, a methodology based on the analysis of words and morphemes bound annotations. We will describe this effort in the next section.

2.1 Annotation Profiling

Figure 1 shows a comparison for the most frequently assigned gloss tags for Akan. To the left we show their ranking for 2015 and one right for 2018. The 2015 snapshot was taken for all Akan data then hosted by TypeCraft. The snapshot from 2018 was performed on our own Akan data. The gloss profile from 2015 still reflects the work of student annotators who were native speakers of the language. The students did not receive special annotation training as part of their linguistic studies, so that they were informed but not supervised annotators. The 2018 annotation profile reflects the work of expert linguists working together with trained student annotators. Figure 1 shows several things:

(I) The 6 most frequent tags remain the same for the 2015 and the 2018 corpus, although for all of the labels their absolute numbers and their ranking relative to each other may have changed, as noted under (II).
(II) The categorization of features for the verbal inflection has been reconsidered; an exception is the assignment of past tense which is in both 2015 and 2018 the most frequently assigned tense label. As indicated by the small black arrows in Figure 1 the ranking of the tense and aspect features future, perfect and progressive has changed. While the morphological marking is unambiguous, that is, the Akan prefix bɛ- stands for future, *a-* for perfect (unless the verb is negative), and *re-* for the progressive, prior annotations tend to reflect the tense expressed in the English translation of the sentence rather than the actual value of the Akan morpheme.
(III) Concerning again verbal features, the perfective and the habitual which figured prominently in the 2015 annotations (see red arrows), are no longer between the most frequently tagged grammatical features, as we rectified errors which for the most frequently assigned tags listed in Figure 1 concerned the difference between the perfective aspect and the perfect tense. The difference is not easily pinned down, especially when sentences appear in isolation, as perfect verb forms can have a perfective meaning. Example (1) illustrates what is meant. The sentence describes a scene

3 TypeCraft (https://typecraft.org) is a service. It can be used online by individual users and projects. As a service TypeCraft *hosts* data. The TypeCraft project is a research group which as one of its activities curates data using the TypeCraft application. The data provided by the TypeCraft project is Typecraft *owned* data.

4 This has developed into a sub project in its own right, cf. Van Dommelen and Beermann (forthcoming).

in a video clip where a cat is looking at a man for a while without him waking up. And in fact the *a*-prefix on the verb *hwɛ* meaning 'look' expresses the perfect tense (PRF), not the aspectual perfective (i.e., completed aspect, marked PFV, as wrongly marked in (1)). (Whether Akan is predominately an aspect or a tense marking language is a long standing discussion in Akan studies (Dolphyne (1988, 1996), Boadi (2008), Osam (1994)).)

(1) Wahwɛ ara nso still papa no nsɔre.

w	a	hwɛ	ara	nso	still	papa	no	n	sɔre
3SG	PFV	*look*	FOC	FOC		*man*.SBJ	DEF	NEG	*get_up*
V			PRT	PRT		N	DET	V	

"It looked (for some time) but still the man is not getting up"

Generated in TypeCraft.

One can say that the expert annotation led to an increased depth of annotation for all parts of the grammar, especially however for the verbal inflection:
(A) Preverbs such as spatial verbs serving as inchoative markers, which were mostly not annotated in 2015, now received an annotation, in Figure 1 reflected by the tag: ITV 'itive'.
(B) multi-functional formatives where now annotated in context, which in Figure 1 is reflected in a decrease in the formatives classified as focus markers, which to us seems to appear as a label when one was not so very sure what the grammatical function really was.
(C) In 2015 mainly, definite nominal modifiers were identified, now also indefinite modification is tagged.
(D) The coverage for negation and relational nouns was increased. In Figure 1, the tag LOC mainly points to relational nouns which are tagged as: POS: Nrel, Gloss: LOC.

In summary our curation effort resulted in the improved conciseness of our annotations especially for the coverage of the verbal inflection; much more work needs to be done for the nominal system. We also improved the consistency of annotations and achieved more depth in annotation.
Finally, for the evaluation of our results we also used trailing annotation profiles as heuristics (see Figure 2). To start with, trailing annotations pointed to random tags which where only assigned once or twice, such as ACC (accusative), or ADD (additive aspect) for cases of reduplication. In our present corpus, we still find trailing annotation contours with over 50 tags, however, these reflect that some

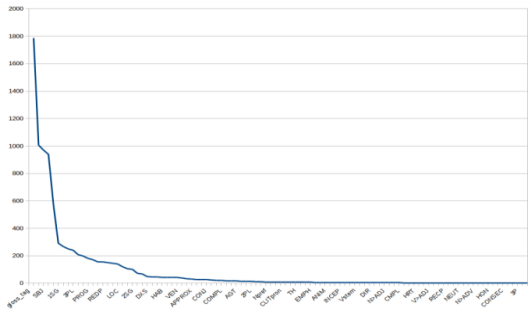


Figure 2: Trailing annotation contour

annotators chose to annotate aspects of the grammar which were not yet targeted for annotation by the project, such as reference, derivational morphemes and thematic roles.

2.2 Extending the corpus and automatic tagging

During the second phase of the curation project, we added with the radio corpus a new resource to our Akan corpus. It consists of 10 texts of between 60 and 100 sentences of transcribed and translated radio conversations between a Ghanaian radio host and his guests.[5] The material reflects contemporary spoken Akan, and prominently features code switching between Akan and English by all speakers.
In parallel, we had worked for a while on the development of an Akan tagger. In the first cycle of our curation project we had trained and tested the tagger on the material we had curated, and we proceeded to test it on one the radio texts, which was then unseen data. The challenges for the parsing of this newly acquired corpus resided in two factors. Although the radio corpus was large compared to our other IGT resources, under testing we still had to deal with the scarceness of data. Secondly, while our training data had no codeswitching, the radio corpus was a codeswitching corpus, that is native Akan speakers were alternating between Akan and English.
The tagger uses a hybrid approach to tagging for both Part-of-speech and gloss tags. It is primarily a universal context-processor, which translates a parsed and annotated sentence on the word and morpheme level into a set of context features. The feature set used by the tagger is configurable. While most taggers use a rather sparse feature set (cf. Schmid 1995, Toutanova et al. 2000), we use a rich feature set of up to several hundred distinct types of features. When supplied with training data, the tagger extracts, according to configuration, all context features observed, and stores them in a database to create a language model. When tagging on untagged data, context features are in the same manner extracted in a left-to-right fashion. That is, at word *n* we have information about the surrounding words (and possibly morphemes) and the *n-1* preceding inferred tags. These contexts are then matched with a set of tags (per context) which are assigned a probability based on their likelihood. The probability generation is performed using Bayesian inference based on occurrence count of the context. Some adjustments are made however. We adjust the feature probabilities upwards for features that are complex, and downwards for features that are rarely seen. The most probable Bayesian estimator is then selected.

On completion, the tagger iteratively reruns the tagging procedure a configurable number of times. On these subsequent runs the features available should now be richer, the idea being that the tagger should iteratively correct its own mistakes when supplying itself with more context.

5 The radio shows were recorded and transcribed by S.Brobbey 2015.

Secondarily the tagger can be configured with specific definite rules which map (a combination of) context features directly to tags. This allows the tagger to deal with noisy training data by correcting the generated language model with overriding rules. For the Akan language model, several such rules were incorporated.

We parsed the corpus using an English and an Akan language model, a process that we will not describe further here. The results were poor for our first run of one of the radio corpus texts, as shown in Table 2.

	Precision	Recall
POS tags	0.72	0.72
Gloss tags	0.70	0.80

Table 2: Classification measures for unseen Akan data.

To improve the performance, we manually re-annotated that text and re-trained the tagger again using these 60 sentences long text.
We further noticed that our effort put in the manual re-annotation did give us some improvement in precision and recall, most likely due to the reduction of inconsistency, but still left us with a flawed linguistic representation of Akan. So taking everything into consideration, in spite of a further round of re-annotation we still had noisier training data than normally is used for the creation of annotated corpora. The use of noisy trainings data is also described by Garrette and Baldridge (2013), who focused on POS tagging using 14 different tags. We dealt in our project with a considerably larger number of word and morpheme level tags, which then also meant a higher and several sources for the inconsistency of the annotation in our training data. In order to arrive also at a grammatically adequate corpus of Akan, we needed to implement on top of Bayesian inference a set of conditions reflecting the basic rules of the Akan grammar. For the present tagger development, we focused on the verbal inflection, and some very basic syntactic rules concerning the position of nouns and their modifiers. With all this in place we re-ran the parser. The considerably improved results are shown in Table 3 and 4, once with direct mapping as our rules enforced, and once without, again for POS and Gloss annotations.The results are calculated by weighted averages over total positives for each tag.

	Tag	**Precision**	**Recall**
Without rules	ADJ	0.93	1
	ADV	1	1
	CONJ	0.95	0.94
	DET	0.77	0.96
	N	0.95	0.99
	PN	0.91	0.96
	PREP	0.95	1
	PUN	1	1
	V	0.56	0.9
	TOTAL	0.83	0.93
With rules	ADJ	0.93	0.93
	ADV	0.78	0.91
	CONJ	0.9	0.94
	DET	0.55	0.97
	N	0.92	0.93
	PN	0.59	0.59
	PREP	0.91	0.95
	PUN	1	1
	V	0.87	0.9
	TOTAL	0.78	0.84

Table 3: Classification results for a selection of POS-tags for seen Akan data.

The "without rules" results are the tagger tagging with no assistance by overriding rules. The total score is calculated by weighted averages over total positives.
In both cases the total precision/recall/f1 ratings are calculated by weighted averages over total positives. Note that directly comparing the result-sets in Table 2 with the results shown in Table 3 and 4 may be misleading, as the improved result is on seen data, while the results shown in Table 2 are on unseen data.

	Tag	**Precision**	**Recall**
Without rules	<empty gloss>	0.98	0.89
	1PL	0.75	0.98
	1SG	0.62	0.94
	2PL	0.67	1
	2SG	0.51	0.96
	3PL	0.78	0.99
	3SG	0.75	0.88
	FUT	0.95	1
	NEG	0.89	0.95
	PROG	0.94	0.97
	TOTAL	0.87	0.86
With rules	<empty gloss>	0.98	0.87
	1PL	0.56	1
	1SG	0.6	1
	2PL	0.67	1
	2SG	0.51	0.92
	3PL	0.78	0.99
	3SG	0.75	0.9
	FUT	0.95	1
	NEG	0.91	0.96
	PROG	0.94	0.97
	TOTAL	0.86	0.85

Table 4: Classification results for a selection of Gloss-tags for seen Akan data.

The tagger in its present stage does not have built-in strategies reflecting syntactic structure of the strings processed, and no strategies reflecting valency information about the lexical items occurring, strategies which of course could add to parsing adequacy. To our knowledge there exist no IGT parsers of Akan, and no digital lexical resources which could be built into the current tagger.[6] In order to make such strategies in principle available to the development of the present tagger, we therefore will explore strategies of transferring information from our Ga resources.

6 Dictionaries like Christaller (1881) and Anyidoho (2006) are not amenable to digital employment.

2.3 Tagger configuration and evaluation

The most important configuration entries of the tagger can be found in Table 5.

Number of iterations per tagging	3
Max *n*-gram length	4
Max length of context feature combinations	3
Ignore empty POS	True

Table 5: Important configurations for the tagger. When combining context features into more complex context features,

When training for English, the configuration was slightly changed by letting the *n*-gram length and combination length be 2 and 1, respectively. The tagger was also configured with specific context feature type weighting. The base context features used (which are combined to more complex features) can be found in Table 6.

Word
Morpheme
Surrounding ngram (of words, POS, etc.)
Prefix ngram (of words, POS, etc.)
Suffix ngram (of words, POS, etc.)
Gloss
Citation form

Table 6: The context features used in training and evaluation. These feature fom the base, or atomic, feature types used, and are combined to more complex context features.

The tagger was first trained on and evaluated with Akan data. The training data was split up into 80%/20% training and test data (in total about 5000 word tokens), for which the tagger had an F1 score of 57%. It was then trained on English, primarily on direct word to tag features. It was not evaluated on English alone.

3. Ga

The starting point for our work with Ga is a Toolbox project holding data of the general-purpose published dictionary (Dakubu 2009). The lexicon file consists of 80,000 lines of code, with 7080 entries, of which 5014 for nouns, and 935 for verbs, of which 722 were annotated for valence. From this Toolbox repository we created a valence lexicon.

3.1 Toolbox lexicon augmented by valence information

In the Toolbox edition used, verb entries are systematically annotated for *valency* such that each entry reflects a unique valence frame. The code used in this annotation is the system *Construction Labelling (CL)* (Hellan and Dakubu 2009, 2010, Dakubu and Hellan 2017). Following the overall left-to-right order indicated in the schema in (2), the CL valency annotation 'templates' are written as illustrated in (3), with the information between each pair of slashes or underscores counting as a 'minimal construction unit' (MCU):

(2) head – valenceFrame – special properties of syntactic constituents – semantic roles of constituents – aspect, Aktionsart – situation type

(3) v-tr-suAg_obTh-CREATION

A paraphrase of (3) is: 'a verb-headed transitive syntactic frame where the subject carries an agent role and the object a patient role, and the situation type expressed is CREATION'.)
This template is applicable to a sentence like (4).

(4) E-fee flɔɔ
3S.AOR-make stew

'she made stew'

The design of a lexical entry in the amended Toolbox version is exemplified in Figure 2, for the verb *fee* as used in (4); the valence codes are written into the lexical entry following the general 'field' style of Toolbox, here as the fields \sl1, \sl2, \sl4, \sl6:

\lx **fee**
\hm 2
\ph fèê, fèé, !fé
\ps verb
\sn 1
\ge make
\de make, do, perform
\sl1 v
\sl2 tr
\sl4 suAg_obTh
\sl6 CREATION
\xv E-fee flɔɔ, samala
\xg 3S.AOR-make stew
\xe she made stew, soap

Figure 2: Example of Ga Toolbox entry enriched with CL valence annotation

A verb with more than one valence frame has one entry specified per frame, hence the verb *ba*, for instance, is represented by 15 different entries in this edition of the Toolbox file. 547 verb lexemes here received altogether 2006 entries annotated in this fashion. In Figure 2, the specification '\hm 2' indicates that this is the second lexeme entered with the form *fee*.[7]
The above resource is also available as a lexical data structure of the type used in Head-Driven Phrase Structure Grammar (HPSG) [8] implemented grammar. The present version consists of 1980 sequentially numbered entries,

7 An overview of full CL templates established for Ga can be seen at: https://typecraft.org/tc2wiki/Ga_Valence_Profile.
8 Cf. Pollard and Sag 1994, Sag et al. 2003. HPSG uses the formalism of Typed Feature Structures (Copestake 2002), whereby every object in the grammar and lexicon belongs to a type; types are organized in multi-inheritance hierarchies.

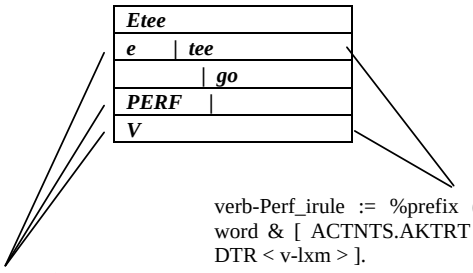
now using the style of notation in (3) in the top line of the entry to indicate the *lexical type* to which the entry belongs. The example in Figure 3 shows a direct counterpart to the Toolbox entry in Figure 2, with *fee_244* as the entry identifier (the formula part ':= v-tr-suAg_obTh-CREATION' means 'belongs to the type v-tr-suAg_obTh-CREATION'):

```
fee 244 := v-tr-suAg_obTh-CREATION &
 [STEM <"fee">,
 PHON <"fee">,
 ENGL-GLOSS <"make">,
 EXAMPLE "E-fee flɔɔ, samala",
 GLOSS "3S.AOR-make stew",
 FREE-TRANSL "she made stew, soap."].
```

Figure 3: HPSG style counterpart to the entry in Figure 2

3.2 Inferring IGT from HPSG type lexical data

It is possible to exchange information between IGT and HPSG grammars. A way of inferring information for an HPSG grammar from IGT is illustrated in Figure 4, this approach is described in Hellan and Beermann (2014) with exemplification for Ga; the implementation framework itself is called *TypeGram*.[9] Here, from a snippet of a Ga IGT like the one indicated, one can infer the grammar specification indicated underneath the snippet, being fragments of a lexical specification and an inflectional rule formulation (attributes such as 'ORTH', 'AKTRT' etc., and value categories such as *v-lxm*, *perf* and *word*, are defined in the general grammar system):



verb-Perf_irule := %prefix (* e) word & [ACTNTS.AKTRT perf, DTR < v-lxm >].

tee-v := v-lxm & [ORTH <"tee">, …].

Figure 4 Illustration of correspondences between IGT and HPSG grammar encoding

Inference of IGT from an HPSG grammar, including valence information, is in turn described in Hellan et al. 2017, the IGT being generated as part of the parse result. While this involves a full grammar, partial inference can also be done from parts of a grammar, such as valence information into an IGT from an HPSG type of lexicon, given recognition of lemma forms in the strings to be annotated. This will be feasible if the procedure can be combined with a morphological parser like the Context parser for Akan described in section 2. By extending such a parser to Ga, and supplementing its assignments with lexical information from the lexicon file, we hope in a future step to make the valence information from Ga operational for Akan.

3.3 Ga valence features

The lexicon file is by itself a large text file,[10] where lexical specifications and valence information are laid out as illustrated above. Of particular interest in a Kwa perspective are construction types quite common in the language but hardly found in European languages. Some types are mentioned below, with indication of the number of verb entries in which they appear as valence information, exemplified for class a and b in (5) :

a. *Bodypart relations* (158 entries)
b. *Identity relations* (110 entries)
c. Subject headed by relational noun (99 entries)
d. Object headed by relational noun (690 entries)
e. Object's specifier headed by relational noun (29 entries)

(5)
a.
v-tr-suIDobSpec_obBPobSpec-suAg_obLoc-COMMUNICATION
Ee-la e-daa-ŋ
3S.PROG-sing 3S.POSS-mouth-LOC"
V N
"He's murmuring incoherently to himself."
('suIDobSpec' = subject (expressed by a clitic) is coreferential with the specifier (expressed by a clitic) of the object; 'obBPobSpec' = object is bodypart of the specifier of the object)

b.
v-tr-obPossp_obBPobSpec-suAg_obLoc-CONTACTFORCEFUL
E-ŋmra e-toi-ŋ
3S.AOR-scrape 3S.POSS-ear-LOC
V N
"She slapped him."

The MCU spelled with capital letters is in each case the situation type to which the content of the sentence belongs; for language comparison of valence frames, such information is of course essential. The eight largest classes in the lexicon file are listed in Table 7:

COGNITION	(83 entries)
COMMUNICATION	(178 entries)
CONTACT	(56 entries)
EXPERIENCING	(45 entries)
MOTION	(180 entries)
MOTIONDIRECTED	(55 entries)
PLACEMENT	(53 entries)
PROPERTY	(164 entries)

Table 7: The most frequently used situation type labels in the Ga lexicon

At present we only have a very small annotated IGT corpus of Ga in TypeCraft, 90 phrases, however with

9 See http://typecraft.org/tc2wiki/TypeGram. For a related approach to grammar induction from IGT using LFG, see Beermann 2014.

10 Much of its information is also exposed at the online 4-languages valency lexicon *MultiVal*, cf. Hellan et al. 2014.

inclusion also of valency information along the lines here described.

3.4 Evaluation of the valence resource

The investigation of valence types in Ga can be related to the research into valency classes started with Levin (1993), followed up, i.a., in VerbNet and in the Leipzig Valency Classes (LVC) Project,[11] being attempts to associate commonalities in morpho-syntactic patterns with semantic factors, both language internally (like Levin op. cit. and VerbNet) and cross-linguistically (LVC). Establishing valency classes for Ga has a tie to VerbNet in aiming at a fairly large coverage of the language's verbs, and to LVC in establishing one more coordinate point in the attempt to attain a typologically broad basis for generalizations within this domain.

Preliminary comparisons of valency frame types for Ga and English suggest that they have less than 20% of their valency frames in common (see, e.g., Dakubu and Hellan (2017)). Even if situation types are common across languages, it is thus by no means a given that there is much commonality between languages as concerns valency classes.

Given the large discrepancies in valency frames between Ga and English, a good strategy may be to first explore commonalities between Ga and other West African languages. Some perspectives are here offered in Schaefer and Egbokhare. 2015, Creissels 2015, conducted in the frame of LVC. However, in the present setting, the natural step will be to build a mapping between Ga and Akan lexical information, assuming that the valency labels used for Ga are adequate also for Akan.

4. Conclusion[12]

With the Akan Context Tagger, we present the first IGT tagger for Akan. It has been used with homogeneous as well as with code-switching data. Our results are encouraging but further training with both types of data are necessary. We plan to use lexical information including valency information developed for Ga to increase its efficiency, which would allow us to tag larger amounts of text than what we have so far. Since the grammatical systems of the languages are not very different, and they are also not too distant lexically, integrating such information will be in principle a feasible task.

From the perspective of Ga, the extension of the parser technology for Akan to Ga should likewise be possible. An interesting issue is here whether an already small HPSG parser for Ga can be utilized in this process. This then would also allow us the syntactic parsing of both languages. From the viewpoint of research into valency classes per se, an alignment of the Ga resources with resources of Akan is desirable, but this is probably more a long-term research project than a matter of transfer of available resources, since this requires analysis at a level of detail far beyond what is required for establishing a large but basic vocabulary for efficient basic morpho-syntactic parsing.

11 Cf. for LVC, Malchukov and Comrie (eds) 2015 and http://valpal.info/; for VerbNet http://verbs.colorado.edu/~mpalmer/projects/verbnet.html.

12 We are grateful for the comments from the three reviewers of this paper.

5. Bibliographical References

Anyidoho, A. et al. (2006) Akan Dictionary. Pilot project. University of Ghana.

Beermann, D. (2014). Data management and analysis for less documented languages. In Jones, M., and Connolly, C. (eds) *Language Documentation and New Technology*. Cambridge University Press.

Beermann, D. and Mihaylov, P. (2014). Collaborative databasing and Resource sharing for Linguists. In: *Languages Resources and Evaluation*. Springer.

Boadi, L. (2008). Tense, Aspect and Mood in Akan. In Ameka, Felix, ed. *Aspect and Modality in Kwa Languages.* Amsterdam: John Benjamins Pub. Co., 9 – 68.

Brobbey, S. (2015) *Codeswitching on Ghanaian Radio Talk-show: "Bilingualism as an Asset".* Master's thesis, University of Bergen, Norway.

Christaller, J.G. (1875). *A Grammar of the Asante and Fante Language Called Tshi.* Gregg Press.

Christaller, J.G. (1881). *Dictionary of the Asante and Fante Language* Basel: Basel Evangelical Missionary Society

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI Publications.

Creissels, D. (2015). Valency properties of Mandinka verbs. In: Makchukov, A., and B. Comrie (eds) Pages 221-260

Dakubu, M. E. Kropp. (2009). *Ga-English Dictionary with English-Ga Index.* Accra: Black Mask Publishers.

Dakubu, M. E. Kropp. (2013). Ga Verbs and their constructions. Monograph ms, Univ. of Ghana.

Dakubu, M.E.Kropp, and L. Hellan (2017) A labeling system for valency: linguistic coverage and applications. In Hellan, L., A. Malchukov and M. Cennamo (eds) (2017).

Dolphyne, F. A. (1988). The Akan (Twi-Fante) Language: Its Sound Systems and Tonal Structure. Accra: Ghana Universities Press.

Dolphyne, F.A. (1996). A Comprehensive course in Twi (Asante) for Non – Twi Learners. Ghana Universities Press, Ghana

Garrette, D. and J. Baldridge (2013). Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages. *Proceedings of NAACL-HLT* 2013, pp. 138–147, Atlanta, Georgia,

Hellan, L. and M.E.Kropp Dakubu. (2009): A methodology for enhancing argument structure specification. In *Proceedings from the 4th Language Technology Conference (LTC 2009)*, Poznan.

Hellan, L. and M. E. Kropp Dakubu. (2010): *Identifying Verb Constructions Cross-Linguistically. Studies in the Languages of the Volta Basin* 6.3. Legon: Linguistics Dept., University of Ghana. http://www.typecraft.org/w/images/d/db/1_Introlabels_SLAVOB-final.pdf.

Hellan, L. and D. Beermann (2014) Inducing grammars from IGT. In Z. Vetulani and J. Mariani (eds.) *Human*

Language Technologies as a Challenge for Computer Science and Linguistics. Springer.
Hellan, L., D. Beermann, T. Bruland, M. E. K. Dakubu, M. Marimon. (2014). *MultiVal* – towards a multilingual valence lexicon. LREC 2014.
Hellan, L., A. Malchukov and M. Cennamo (eds) (2017) *Contrastive studies in verbal valency*. Amsterdam: J. Benjamins.
Hellan, L., D. Beermann, T. Bruland, T. Haugland, E. Aamot. (2017). Creating a Norwegian valence corpus from a deep grammar. In Vetulani (ed) *Proceedings from LTC 2017*. Poznan.
Levin, B. (1993). *English Verb Classes and Alternations.* Chicago IL: University of Chicago Press.
Malchukov, A. L. & Comrie, B. (eds.) (2015). *Valency classes in the world's languages.* Berlin: De Gruyter Mouton.
Osam, E. K. (1994). Aspects of Akan Grammar. A Functional Perspective. Ph.D. thesis, University of Oregon.
Pollard, C. and Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. Chicago University Press.
Rask, R. (1828). *Vejledning til Akra-Sproget på Kysten Ginea* (Introduction to the Accra language on the Guinea Coast).
Sag, I., Wasow, T. and Bender, E. (2003). *Syntactic Theory*. CSLI Publications, Stanford.
Schaefer, R.B, and F. O. Egbokhare. (201)5. Emai valency classes and their alternations. In Malchukov, A. and B. Comrie (eds) 2015. Pp. 261-298.
Schmid, H. (1995): Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.
Tesnière, L. (1959). *Éleménts de syntaxe structurale*. Paris: Klincksieck.
Toutanova, K. and C. D. Manning. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70.
Van Dommelen, W. and D. Beermann (forthcoming) A study of Akan Tone – the case of NA.

6. Language Resource References

The valence resources for Ga:
https://typecraft.org/tc2wiki/Ga_Valence_Profile

TypeCraft Akan corpus:
https://typecraft.org/tc2wiki/Special:TypeCraft/PortalOfLanguages

The TypeCraft Context Tagger :
https://github.com/Typecraft/casetagger

Gathering Data for Speech Technology in the Welsh Language: A Case Study

Delyth Prys, Dewi Bryn Jones
Language Technologies Unit, Bangor University
Bangor, Wales, UK
{d.prys, d.b.jones}@bangor.ac.uk

Abstract

Less-resourced languages face additional challenges in the creation of tools and resources for speech recognition applications. These include lack of funding, sparsity of data and shortage of experts with relevant skills. On the other hand there are also opportunities to be had from tapping into committed communities of language activists, and potentially developing innovative solutions to common problems that may be applied elsewhere. This paper describes a recent series of short-term projects for the Welsh language that have used crowdsourcing methodologies, together with data from Wicipedia (the Welsh Wikipedia) and existing Welsh corpora, to further advance the field. They have also borrowed and adapted open source tools, such as MaryTTS and Mozilla CommonVoice that were already freely available. In addition this paper provides some pointers towards further needs and solutions for speech technology in less-resourced languages, aiming at a coherent, long-term approach that may be applicable in many environments.

Keywords: speech recognition, less-resourced languages, Welsh

1 Introduction

Automatic speech recognition (ASR) technology is the most critical component in intelligent speech interfaces that are becoming increasingly popular amongst consumers who wish to access web-based information and services. Such platforms have recently become viable due to advances in the use of neural network methods to reduce speech recognition word error rates. However, ASR research has been primarily carried out for English and other major languages, where the vast amounts of text and speech data required for training exist, or where it is commercially viable to collect and deploy it.

It remains challenging however to develop ASR for less-resourced languages with limited or no training data, and where there is no immediate or direct commercial benefit for private sector actors. In such cases, stimulation and investment is required from public organisations or foundations that desire, or are under legal obligation, to support speakers of less-resourced languages.

In 2013, the Welsh Government published its *Welsh Language Technology and Digital Media Action Plan* that provided grants for financing projects aiding any aspect of pairing Welsh language and technology. Due to budgetary and other constraints, these grants could only fund short term projects, with no certainty of continued funding. They needed to produce outputs useful to the public, or to engage the public by other means, and they needed also to show continued innovation, rather than 'more of the same'. In addition, they had to demonstrate good fit to the Government's Action Plan and the priorities named in it.

ASR development is a long-term investment, requiring research and development by many individuals over a number of years, rather than months. In order to advance the development of Welsh ASR therefore, and in the absence of any other sources of funding, a strategy was devised to incrementally develop, through a series of possible projects, a Welsh language digital assistant that could be exploited to guide the research and development of Welsh ASR from the initial position of no data and no support, to incrementally increasing the amounts of data available, and specifically helping towards a digital personal assistant that would be able to understand and respond to oral commands and questions in Welsh.

1.1 Project 0 (2013 – 2014)

Project or stage 0 of our longer-term objectives began with the study of Welsh letter-to-sound rules and production of pronunciation dictionaries as well as the novel development of iOS and Android apps for crowdsourcing a speech corpus. Up until 2014, the Voxforge.org website had been the only possible platform for collecting transcribed speech for a new language. However, our apps brought our crowdsourcing activity up to date in order to facilitate contributions from more recently and increasingly popular mobile devices. The app, named 'Paldaruo' (Welsh for 'to chatter'), provided a small initial collection of 43, specially crafted for phonetic coverage, prompts that users could record at their leisure. Recording all prompts would take approximately 30 minutes and over 400 users duly complied with providing recordings of their voices. The speech data was successfully used to build acoustic models that successfully recognised commands to move a robot arm connected to a Raspberry Pi using spoken commands in Welsh (Cooper et al, 2014).

1.2 Project 1 (2015-2016)

After the end of Project 0, the Paldaruo app was kept in the app stores, and continued to collect speech data from the Welsh language community.

The increasing amount of speech data was put to use in the next project we conducted, namely to expand the capability of Welsh speech recognition to recognize closed and simple, but still useful, questions regarding time, weather and news. These questions in turn would be included as 'skills' into the first prototype version of a Welsh-language digital assistant named 'Macsen'. Macsen would run on a Raspberry Pi, with the Welsh speech recognition engine implemented using HTK acoustic models within Julius.

Example questions (translated) included: "What is the time?", "What is the weather for today?", "What's the news?", "Play me some music", "Play me some Welsh music" (Jones and Cooper, 2016).

1.3 Project 2 (2016 – 2017)

Project 1 gave us a valuable insight of where challenges remained and to strategize for longer term progress whilst still constrained by short term project goals.

Speech technologies are dependent on speech data consisting of audio aligned with textual annotations at segment, word and/or phonetic levels. The Paldaruo speech data thus collected was used to produce the first Welsh language forced aligner, based on Prosodylab Aligner. This would aid future research in increasing and improving the size and quality of Welsh language speech data.

Project 2 also provided an opportunity to re-base Macsen's speech recognition components on the more widely used and popular Kaldi-ASR. A prototype recipe for Macsen's questions was developed and the resulting engine was hosted online and accessible via an API, not only to Macsen but to other software such as the Kõnele app for Android.

Finally, speech interfaces such as Macsen also require text-to-speech capability. Up until project 2, the Welsh TTS options for 'Macsen' were limited to either an open source, but robotic sounding, Festival diphone voice, or a commercially available naturally sounding voice from Amazon Ivona's SpeechCloud. Project 2 provided an opportunity for Macsen to gain open source and naturally sounding voices by developing tools to simplify building unit selection Welsh voices with MaryTTS.

1.4 Project 3 (2017 – 2018)

The Welsh Government in its most recent strategy for the Welsh Language (Cymraeg 2050, 2017) emphasised again the importance of digital technologies, including investing in Welsh language speech technology, to ensure the vitality of Welsh, and the fulfilment of its ambitious plan to double the number of Welsh speakers by 2050.

As part of this commitment, it funded further work on Macsen, and the publication and open dissemination of resources used in its improvement. This led to the current project, which has as its aim extending Macsen's speech recognition capability to being able to recognise more open-ended questions that a user would typically ask a digital personal assistant in order to gain new knowledge.

Wikipedia is an invaluable source for knowledge, especially for less-resourced languages where other available digital data may be scarce. The Welsh language Wikipedia, called Wicipedia, emphasizes creating original content in Welsh, as well as localizing international content where appropriate. The Welsh Government's Cymraeg 2050: Work programme 2017-21 (2017) specifically names supporting efforts to increase the number of Wicipedia pages as one of its aims for the period to 2021. Wikipedia's regular structure also facilitates finding information in its pages, and extracting relevant material to answer questions.

A module or skill in Macsen, enabled by a more developed Welsh language speech recognition, would thus respond to oral questions such as "Pwy oedd Hywel Dda?" (*"Who was Hywel Dda?")* by reading the first paragraph of the article on Hywel Dda in Wicipedia. This has entailed crowdsourcing greater amounts of richer sound data, as well as the analysis of Wicipedia content and usage.

2. Formulating Recording Scripts for Questions

In our first project to develop speech recognition for Welsh from scratch, sound recordings of a small set of prompts had been sufficient for simple questions and commands. Our experience with the Paldaruo app provided an insight into the nature of participation and engagement levels when crowdsourcing in a less-resourced language community. It was observed that:

- Welsh language speakers were very willing to contribute recordings of their voices
- Contributors would typically record in one sitting/session
- A majority of contributors did not record all of the prompts
- The number of contributors, so far, (without any dedicated significant funding for marketing) is in the 500-600 range.

These observations guided our decisions as to how a larger and richer set of speech data, required for the larger domain of recognising questions to Wikipedia, should be realised. In contrast to assumptions in larger languages, crowdsourcing from a less-resourced language community requires extracting as much speech information as possible from the limited numbers and length of contributions.

Questions to Wikipedia however represent a large domain and thus careful planning would be required as to how the Paldaruo crowdsourcing capability could be expanded to facilitate successful recognition of questions likely to be asked by Welsh language users.

2.1 Subjects covered in Wicipedia

A question and answering module inside Macsen would be perceived as useful if it would answer questions on the subjects typically asked of it. Thus we would have to ensure that at the very minimum, models and thus speech data for these subjects would be sufficiently captured in any crowdsourcing activities.

The Welsh Wikipedia currently contains over 90,000 articles. These are not usually straight translations from the English, and cover specifically Welsh topics or have a unique Welsh slant on international or more general subjects. Judging what would be the most typically viewed subjects in Wicipedia entailed gaining the assistance of the Wikipedia UK Manager in Wales. The project was kindly furnished with a list of analytics websites, in particular Catscan (https://petscan.wmflabs.org) and Topview (https://tools.wmflabs.org/topviews/?project=cy.wikipedia.org). The summation of top views for each month from July 2016 to July 2017 gave surprising results. For example, the most viewed article in Wicipedia by far was about an international pornographic actress. Closer inspection however showed that she did not have an English language article, therefore English queries in popular search engines were pointing towards the Welsh language article.

Having manually weeded out anomalies such as this, and filtered for any articles deemed to contain inappropriate material for family audiences, an initial list of 1000 articles was filtered down to 646 that had received at least 250 visits. The top twenty articles in the filtered list are given in Table 1.

Article Title	***(Translation)***	**Visits**	**Type (object)**	**Tense (default)**
Cymraeg	*Welsh*	12105	Lang	
Saesneg	*English*	10840	Lang	
Cymru	*Wales*	8909	Place	
Unol Daleithiau America	*United States of America*	8688	Place	
Y Deyrnas Unedig	*United Kingdom*	5842	Place	
Ffrainc	*France*	4902	Place	
Yr Ail Ryfel Byd	*Second World War*	4735	Event	Past
T. Llew Jones	*T. Llew Jones*	4687	Person	Past
Lloegr	*England*	4669	Place	
Yr Almaen	*Germany*	4479	Place	
Wikipedia	*Wikipedia*	4339		
Hedd Wyn	*Hedd Wyn*	4159	Person	Past
Sioned James	*Sioned James*	4141	Person	Past
Rhyngrwyd	*Internet*	4131		
Ewrop	*Europe*	4089	Place	
Lladin	*Latin*	4030	Lang	
Caerdydd	*Cardiff*	4012	Place	
Awstralia	*Australia*	4004	Place	
Wicipedia Cymraeg	*Welsh Wikipedia*	3951		
Cynnwys rhydd	*Free content*	3896		

***Table 1** - Edited List of Most Viewed articles in Wicipedia (07/16 - 07/17)*

2.2 Enlarging the prompts set with questions for improved acoustic modelling

Analysis of the most viewed subjects led to a categorisation on the type of question one would naturally ask, such as "Pwy ydy …?" ("*Who is ….?")* / "Pwy oedd ….?" (*"Who was….?")* / "Beth ydy…?" *("What is ….? ")* in order to gain knowledge on that subject.

With additional meta-data tagged in the *Type* and *Tense* columns as demonstrated in Table 1, an initial list of questions was generated for all 646 subjects, examples of which as given in Table 2.

Example Question	**(Translation)**
Beth ydy Cymraeg?	*What is Welsh?*
Beth ydy Saesneg?	*What is English?*
Beth oedd Yr Ail Ryfel Byd?	*What was the Second World War?*
Pwy oedd T. Llew Jones?	*Who was T. Llew Jones?*
Beth ydy Wikipedia?	*What is Wikipedia?*
Pwy oedd Hedd Wyn?	*Who was Hedd Wyn?*
Pwy oedd Sioned James?	*Who was Sioned James?*
Beth ydy Rhyngrwyd?	*What is the internet?*
Beth ydy Lladin?	*What is Latin?*
Beth ydy Wicipedia Cymraeg?	*What is Welsh Wikipedia?*
Beth ydy cynnwys rhydd?	*What is free content?*

Table 2 – Generated Example Questions

This produced a rather limited repertoire of different questions and ways of asking for information. In order to cross-reference with a wider set of possible question, it was decided not to rely solely on Wicipedia, and to seek other sources for examples of questions in Welsh.

The research team undertaking this work was fortunate to have at its disposal the 100 million-word Cysill Ar-lein Welsh language corpus, collected via the free on-line

provision of its popular Welsh language spelling and grammar checker product Cysill (Prys and Jones 2016a). Over 17,000 examples of questions were extracted by simple regular expression matching for the '?' symbol, and omitting any questions that contained capital letters or numbers, in case such questions included private information such as names and/or phone numbers.

Experience with our Paldaruo crowdsourcing app however showed it to be unrealistic to expect a set of more than 17,000 questions to be recorded by thousands of contributors. If contributors numbered only in their hundreds, although a significant achievement for a small language community, there was a danger that a random selection of prompts from too large a prompt set would not provide sufficient phonetic coverage in the speech data. The size of the prompt set would have to be sufficient for a random selection mechanism to allow contributors, between them, to be able to contribute as much phonetic data as possible.

For this task, we were able to approximate a suitably sized and phonetically balanced prompt set by re-using our Welsh MaryTTS resources. Usually these resources are used in crafting recording scripts; however in this utilization we reduced the number of questions from more than 17000 to 300.

A further reduction was achieved by reducing the number of questions with the common beginning "Beth yw..?" (W*hat is*) and relocating subjects into prompts that listed individual words. Some of the individual words are mutated forms that do not usually occur in standalone words, as they are triggered by other words in a sentence. However, some of them include phonemes that were otherwise rare, and so this was deemed the easiest way of including them in a small prompt set. In all a new collection would consist of 270 prompts examples of which can be seen in Table 3.

A fedrwch chi fy helpu I os gwelwch yn dda?	*Can you help me please?*
Faint o'r gloch mae amser cinio yn gorffen?	*What time does lunch finish?*
Faint o'r gloch fydd y bws nesaf yn mynd heibio?	*What time does the next bus go past?*
Faint mae llaeth yn costio?	*How much does milk cost?*
Wyt ti mewn wythnos nesa o gwbl i arwyddo nhw?	*Are you in next week at all to sign them?*
Beth sydd yn digwydd yn y stori yn eich geiriau eich hun?	*What happens in the story in your own words?*
Beth oedd y ffeithiau mwyaf diddorol i ti a pham?	*What were the most interesting facts for you, and why?*
Ble mae cartref gofal agosaf yr Awurdod Lleol ?	*Where is the nearest Local Authority care home?*
Rhyfel Cartref America, Y	*American Civil War, The*
Brythoniaid, y Chwyldro Diwydiannol	*Ancient Britons. The Industrial Revolution*
Y Dirwasgiad Mawr, Yr Oesoedd Canol, Y Rhyfel Oer	*The Great Depression, The Middle Ages, The Cold War*
Bob Dylan, Bryn Fôn, Caryl Parry Jones	*Bob Dylan, Bryn Fôn, Caryl Parry Jones*
Ceri Wyn Jones, Cymraeg, Dafydd Dafis	*Ceri Wyn Jones, Welsh, Dafydd Dafis*
Theuluoedd, Toes, Porthaethwy, Cibwts, Rhaeadr, Lliw, Minoaidd, Nymff	*Families, Dough, Menai Bridge, Kibboutz, Rhayader, Colour, Minnoan, Nymph*
Magdalen, Cewri, Ffeuen, Clwyfau, Puw, Sipswn, Llai, Fronhaul	*Magdalen, Giants, Bean, Wounds, Pugh, Gipsies, Fronhaul*
Soia, Deuawd, Prawf, Rois, Teulu, Byw, Ddaw, Amheus	*Soya, Duet, Test, I gave, Family, Live, Will Come, Doubtful*
Bwdhaeth, Botswana, Gewyn, Heddiw, Ebwy, Lleisiau	*Buddhism, Botswana, Ligament, Today, Ebbw, Storyteller, Llangeitho, Voices*
Caerhun, Llew, Arwyllsiad, Ieithoedd, Ehangdir, Ceulan, Bontddu, Nhrwyn	*Caerhun, Lion, Discharge, Languages, Expanse, Hollow Bank, Bontddu, Nose.*

Table 3 - Example prompts with English translations

3. Crowd sourcing voice recordings

When no speech data is available, either due to none actually existing or not suitably licensed, for developing speech recognition for any language, crowdsourcing can be an effective strategy for bootstrapping initiatives. Up until November 2017, the fourth release of the Paldaruo speech corpus had collected 38 hours of speech from 536 contributors which were available according to a CC-BY license. Challenges remain in expanding the corpus, attracting more contributors and in ascertaining its quality for speech recognition development.

A recent welcomed international development regarding crowdsourcing speech data has been Mozilla's CommonVoice project (Mozilla n.d). CommonVoice aims to crowdsource large and publicly available voice datasets in order to foster innovation and healthy commercial competition in machine learning based speech technology. Launched in summer 2017, by January 2018 CommonVoice had crowdsourced, and verified, 254 hours of English language speech, provided by nearly 20,000 volunteers worldwide (The Mozilla Blog, 2017). Its ambition, with the help of open source communities, is to increase the number of hours to the thousands and to build large public-domain datasets for as many languages as possible in the world.

As our case study demonstrates, this ambition will be challenging for languages that have less human and digital

resources, where crowdsourcing dynamics do not exist or cannot scale down effectively.

CommonVoice has allowed us to fork the website code to provide a web-based version of the Paldaruo app. It also has additional features such as the ability to crowdsource verifying the correctness of other recordings. The lack of an easy mechanism for verification and quality control had been a weakness in our previous Paldaruo app. By now the entire corpus is being verified by volunteers and future release of the Paldaruo corpus will similarly claim to have verified hours of speech.

We also embraced the CommonVoice software in order to expand the Paldaruo corpus via a website and to understand how Mozilla's approaches could be scale-down and made impactful for lesser resourced languages. Figure 1 shows a screenshot of our offering at http://paldaruo.techiaith.cymru, with Welsh language text. This translates as "What is speech recognition and why Paldaruo?" with a short explanation on its increasing importance and its ubiquity e.g. in personal devices.

Figure 1 - Screenshot of CommonVoice website at http://paldaruo.techiaith.cymru to crowdsource Paldaruo

Required alterations to the CommonVoice website code included not only its localization into Welsh but also making the website functionality consistent with smaller prompt sets (i.e. to avoid providing already recorded prompts), additional profile meta-data fields, and mandatory provision of such profile meta-data before allowing recording.

To date there has not been a large publicity campaign to recruit further volunteers to record their voices through any Paldaruo interface as these are costly and time-consuming to organise. Just keeping the Paldaruo app active from the end of the original 2014 to late 2017 gathered around 136 additional recordings. However, a new appeal disseminated through social media sites and e-mails has drawn a good initial reaction, with further dissemination undertaken by volunteers retweeting, and incorporating the appeal in their own newsletters and e-mail lists. Participants who are reluctant to record their own voices now also feel able to contribute by verifying the accuracy of recordings made by others. We believe that an active publicity campaign would enable us to significantly increase both the number and length of contributions so far made.

4. Making results visible and accessible

Good visibility of the outputs made possible by crowdsourcing generates an enthusiasm amongst the public interested in supporting the development of Welsh language speech technology. All projects contain outreach activities which have been instrumental in generating a following, with past contributors keen to view and listen to the latest developments.

On a technical level, and as stipulated in the grants awarded, all project outputs to
date have been published and shared with permissive open source licenses on the Welsh National Language Technologies Portal (Prys and Jones, 2016b), and/or on widely used repositories such as GitHub. This has led the team to engage with several developers, companies and enthusiasts interested in utilising the speech technology outputs in their own Welsh language provisions. This includes electronics students attempting to create a 'body' for Macsen as a personal assistant, and teacher trainers developing primary school lessons on coding in the context of Welsh language (Prys and Jones, 2017).

Further attempts at reaching out to developers and users have led to a website dedicated to supporting anyone wanting to obtain, create and develop their own Macsen. Based on a translation and fork of the Jasper project, the website can be found at https://projectmacsen.github.io as seen in Figure 2.

The widest audiences have been reached however via television and radio interviews in the Welsh language media when questions arose with the appearance and increasing popularity of Amazon Alexa and Google Home as to why there were no Welsh language versions. In the meantime, team members continue to present the work at societies and cultural events such as Hacio'r Iaith (*'Hacking the Language'*) the National Eisteddfod. Each outreach occasion presents opportunities to appeal for contributions to the Paldaruo speech corpus.

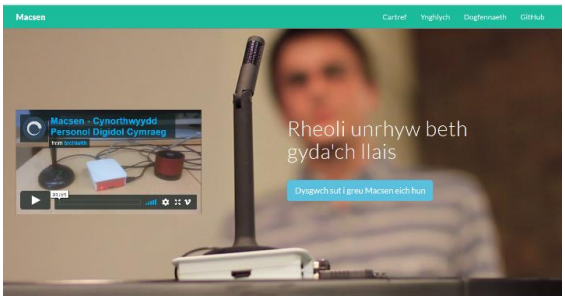


Figure 2 - Screenshot of website for Macsen users and developers (https://projectmacsen.github.io)

5. Conclusions and Future Work

No research project is undertaken in isolation. At their best, projects are undertaken to increase the sum of human knowledge and support human life and culture. Crowdsourcing brings researchers and their public closer together, and the greater intimacy offered by small language communities can bring positive benefits, not least in the field of speech technology.

On the other hand, generic, global projects, such as Wikipedia, MaryTTS and Mozilla Common Voice, offered on generous open licences, are of vital importance to less-resourced languages, enabling knowledge and resources to be shared and built up to the benefit of all. Combining use of local, language-specific knowledge and resources with global tools and initiatives can provide the help needed to level the playing field for digitally excluded language communities, and such combinations are to be welcomed.

As with speech technology for many other less-resourced languages, much remains to be done for Welsh. Within the current project, as well as gathering and processing additional data, improved acoustic and language models need to be built. Assessing quality, and answering the question 'how much data is enough, or at least sufficient' is becoming increasingly urgent as we seek to improve on the first on the first generation of outputs.

The increasing pace of technological developments, especially neural networks for speech recognition, is creating new challenges for less-resources languages, especially as truly enormous datasets are needed to gain the best results. However, smarter ways of working, use of both generic global and local language-specific knowledge can provide a way ahead for many less-resourced languages. We feel privileged to be part of such global and local communities.

6. Acknowledgements

The projects reported on in this paper were made possible with the financial support of the Welsh Government, through its Technology and Digital Media in the Welsh Language Fund and S4C. The authors would also like to thank the contributors from various hackers and communities of users that assisted us on the projects, as well as Robin Owain, Wikimedia UK Manager in Wales and Michael Henretty from the Mozilla Common Voice project for their aid.

7. References

Cooper, S. Jones, D. B. and Prys, D. 2014. *Developing further speech recognition resources for Welsh*. In: Judge, J., Lynn, T., Ward, M. and Ó Raghallaigh, B. eds. Proceeedings of the First Celtic Language Technology Workshop at the 25th International Conference on Computational Linguistics (COLING 2014), 23 August 2014, Dublin, Ireland. pp. 55-59.

DFKI. 2016. http://mary.dfki.de/ [accessed 12/01/2018]

Gorman, Kyle, Jonathan Howell and Michael Wagner. 2011. Prosodylab-Aligner: A Tool for Forced Alignment of Laboratory Speech. Canadian Acoustics. 39.3. 192–193.

Jones, D.B. and Cooper, S. 2016 *Building Intelligent Digitial Assistants for Speakers of a Lesser-Resourced Language*. p74-79 Proceedings of the LREC 2016 Workshop "CCURL 2016 – Towards an Alliance for Digital Language Diversity", Claudia Soria, Laurette Pretorius, Thierry Declerck, Joseph Mariani, Kevin Scannell, Eveline Wandl-Vogt.

Mozilla (n.d.) Common Voice https://voice.mozilla.org/. [accessed 12/01/2018]

Prys, D., Prys G., and Jones, D.B. 2016a. *Cysill Ar-lein: A Corpus of Written Contemporary Welsh Compiled fro m an On-line Spelling and Grammar Checker*. Proceedings of the Tenth International Conference on Language Resour ces and Evaluation (LREC 2016) Portoroz, Slovenia.

Prys, D., and Jones D. B. 2016b. *National Language Technology Portals for LRLs: A Case Study.* Language Technologies in Support of Less-Resourced Languages, (LRL 2015).

Prys, D., Jones, D.B & S. Ghazzali. 2017. *Using LT tools in classroom and coding club activities to help LRLs* Language Technologies in Support of Less-Resourced Languages, (LRL 2017).

The Mozilla Blog (2017) https://blog.mozilla.org/blog/2017/11/29/announcing-the-initial-release-of-mozillas-open-source-speech-recognition-model-and-voice-dataset/ [accessed 12/01/2018]

Voxforge http://www.voxforge.org/ [accessed 06/03/2018].

Welsh Government. 2013. Welsh language Technology and Digital Media Action Plan. http://gov.wales/docs/dcells/publications/230513-action-plan-en.pdf [accessed 12/01/2018]

Welsh Government. 2017. Cymraeg 2050: A million Welsh speakers. Work programme 2017-2021. http://gov.wales/docs/dcells/publications/170711-cymraeg-2050-work-programme-eng-v2.pdf [accessed 12/01/2018]

The MediaBubble Dataset

A Crowdsourcing Dataset for Topic Detection Tasks for the Hungarian Language

László Grad-Gyenge, Linda Andersson
Creo Group, TU Vienna
Budapest,Vienna
laszlo.grad-gyenge@creo.hu, linda.andersson@tuwien.ac.at

Abstract

The paper presents the MediaBubble Dataset. Developing the dataset, our primary aim is to fill the gap of political topic detection dataset for Hungarian, a low density language. The dataset contains 1 000 political articles appeared in the Hungarian on-line media in political topics on the major news portals between 26.04.2017 and 29.04.2017. The dataset contains the topics and topic assignments created by 3 annotators. In addition, the dataset is initiated as a crowdsourcing dataset. It means that although the dataset is publicly available, in order to download it, a dedicated amount of annotations has to be conducted as a contribution to research.

Keywords: topic detection, crowdsourcing, dataset, Hungarian

1. Introduction

The importance of understanding political discourse on on-line platforms is becoming increasingly clear. There is several work in this direction done on high density languages such as German and English, but very few cover low density languages such as Hungarian. The purpose with the project MediaBubble [1] is to develop an adequate dataset for topics reflecting different political opinions of on-line news articles for the Hungarian language. Political preference on the same topic is referred to as *framing* (Card et al., 2015). Framing is related to the bias in a political discussions which emphasize or favor the speaker/writers opinion on a specific topic. Framing has been a central concept in political science and journalism for many decade (Goffman, 1974).

The primary goal of the project is to aid on-line news readers to eliminate / extend their filter bubble by recommending news articles on the same topic, but with a different political preference, i.e. frames, on the article the user is interested in. The frame for a news will have a different source frame aiming for a change in the perception of the issue among the reader (Scheufele, 1999). In (Fulgoni et al., 2016), they discover within the frame of *police violence* that the liberal press would use term as *uprising*, meanwhile the conservatives press would refer to the same event as *riot*.

In end-application of the MediaBubble project, we will utilize various semantic representation techniques on the articles appearing on-line. In order to train and evaluate different methods we first need to establish a dataset.

The Hungarian language belongs to the group of Uralic languages. Hungarian is an agglutinative language and unlike Germanic languages, does not follow a strict word order. In addition, to mention some of its properties, the language is rich on grammatical cases, lacks on grammatical gender, uses postpositions, involves specific plural markers, uses possessive suffixes and numeral expressions are singular. In order to develop semantic representation methods for such a language, specific techniques are to be involved. Our interest in creating this data set is two-folded. At first, we think that an adequate dataset for political mining of low density language is interesting by itself. At second, to be able to compare state-of-the-art mining algorithms involving distributional semantic methods on Uralic in comparison with Germanic language would contribute to the computational linguistic research community. Our contribution to the MediaBubble Dataset can be summarized as:

- collecting the news text from on-line portals with different political views,
- the initial annotations in the dataset,
- the maintenance of the underlying infrastructure,
- the user interface and work-flow logic to serve further annotation tasks.

The rest of the paper is organized as follows. Section 2. discusses related research conducted. Section 3. presents how the data was collected, pre-processed and its format. Section 4. introduces the user interface to prepare and to contribute to the dataset. Section 4.1. presents the format and the statistical properties of the dataset. Section 5. concludes the paper and gives insight into our future plans.

2. Related Work

In recent years, there has been a significant interest of mining social media for political preference. Tweets are by far the most popular, there have been several studies regarding tweets associated with elections in different countries e.g. Germany (Tumasjan et al., 2010), Ireland (Bakliwal et al., 2013) and U.S. presidential election (Williams and Gulati, 2008).

In (Tumasjan et al., 2010), the focus was on the federal election in order to investigate if the twitter flow could predict the outcome of the election. In (Bakliwal et al., 2013),

[1] The MediaBubble project has been carried out by Creo Group (creo.hu) in cooperation with Mérték Médiaelemző Műhely (mertek.eu). The project has been supported by the Google Digital News Initiative (digitalnewsinitiative.com)

they addressed sentiment analysis of the Irish General Election 2011, the goal was to classify tweets on a specific topic as positive/negative/neutral and also to see if it was possible to detect tweets as sarcastic i.e. if the literal sentiment was different to its actual sentiment. In (Williams and Gulati, 2008) they studied the effect of using social media platforms such as Facebook for vote sharing in the presidential primaries 2007.

Political tweets have also been used to discover party loyalty, in (Calzolari et al., 2016), they investigated if it was possible to discover if social and behavioural information available on Twitter would give sufficient data to train a classifier in order to identify *aisle-crossing politicians* i.e. those politicians who vote against their party. They collected 184,914 tweets from members of the U.S. Congress (both the House of Representatives and Senates) utilizing frames. Each tweet could be classified with one or more frames. They had a predefined list of 17 possible frames (e.g. Economic, Capacity & resources, Quality of Life, Culture identity, etc.

Frames have also been explored in public statements, congressional speeches, and news articles (Tsur et al., 2015; Baumer et al., 2015; Fulgoni et al., 2016). Tsur et al (Tsur et al., 2015), observed the language of framing in agenda setting campaigns. In (Baumer et al., 2015), they developed computational techniques in order to detect different framing on various political issues. Their work is based upon the fact that framing can have significant impact on the readers perception and therefore it is important to draw the reader's attention to the language of framing. They collected data from 15 political news feeds, and lay annotators have been used in order to reflect the frames among the general public. Fulgoni et al (Fulgoni et al., 2016), studied 17 different topics ranging from climate change to common core and from abortion to police violence. They observed divergence of themes between each partisans sides (conservative versus liberal), each sides would use different frames in order to appeal to their readers. For instance, in the abortion debate the conservative press use pro-life and the liberal press use anti-life.

We would like to contribute to research regarding language of framing for other languages than English such as Hungarian. It is of interest to study and to draw the readers' attention to the language of framing due to the pervasive influence of framing have on the reader perception on an issue.

3. Collecting the Dataset

The initial MediaBubble Dataset has been conducted on 1 000 news articles. The set of articles to be annotated is defined as the articles appeared on the major Hungarian news portals in the interval 26.04.2017 - 29.04.2017.

Table 1 summarizes the concrete portals involved in the initial annotation process of the dataset. The column denoted "Name (Url)" contains the name and the URL of the portal. The column denoted "Att" contains the political attitude of the particular portal according to the European political scale. The attitude is represented on a 1-5 scale. Value 1 represents the left attitude. Value 2 represents the moderate left attitude. Value 3 represents the centered political attitude. Value 4 represents the moderate right attitude. Value 5 represents the right attitude. The political attitude values are the subjective opinion of the author of this article and should not be considered as the ground truth. The column denoted "Cnt" contains the number of articles involved in the annotation process from the specific news portal.

Name (Url)	Att	Cnt
24.hu (`24.hu`)	2	99
PestiSrácok (`pestisracok.hu`)	5	28
B1 BLOGCSALÁD (`b1.blog.hu`)	1	4
mandiner (`mandiner.hu`)	5	52
hvg.hu (`hvg.hu`)	2	99
Index (`index.hu`)	3	105
Kettős Mérce (`kettosmerce.blog.hu`)	2	2
Magyar Idők (`magyaridok.hu`)	5	102
Magyar Narancs (`magyarnarancs.hu`)	2	34
Magyar Nemzet (`mno.hu`)	4	102
Népszava (`nepszava.hu`)	2	134
ORIGO (`www.origo.hu`)	5	129
444 (`444.hu`)	1	62
888.hu (`888.hu`)	5	46
atlatszo.hu (`atlatszo.hu`)	1	1
Ténytár (`tenytar.hu`)	2	1

Table 1: List of news portals involved in the annotation process.

3.1. Format

The dataset can be downloaded as a compressed archive in tar.gz format. The files contained in the archive are in a tabular format and are the following.

- annotators.csv – The annotators involved in the project. The file contains an `id` and a `nick` column. The columns stand for the unique identifier of the annotator and for the nick name of the annotator, respectively.
- topics.csv – The topics defined by the annotators. The file contains an `id` and a `title` column. The columns stand for the unique identifier of the topic and for the title of the topic, respectively.
- articles.csv – The articles to be annotated. The file contains an `id` and a `title` column. The columns stand for the unqiue identifier of the article and for the title of the article, respectively.
- assignments.csv – The topic assignments of the articles. The file contains an `article`, a `topic` and an `annotator` column. The columns stand for the unique identifier of the article, the unique identifier of the topic and for the unique identifier of the topic assignment, respectively.

The concrete tabular file format is csv. The delimiter character is comma, the titles are quoted with double quotes. Double quotes in strings are escaped with repeated double quotes.

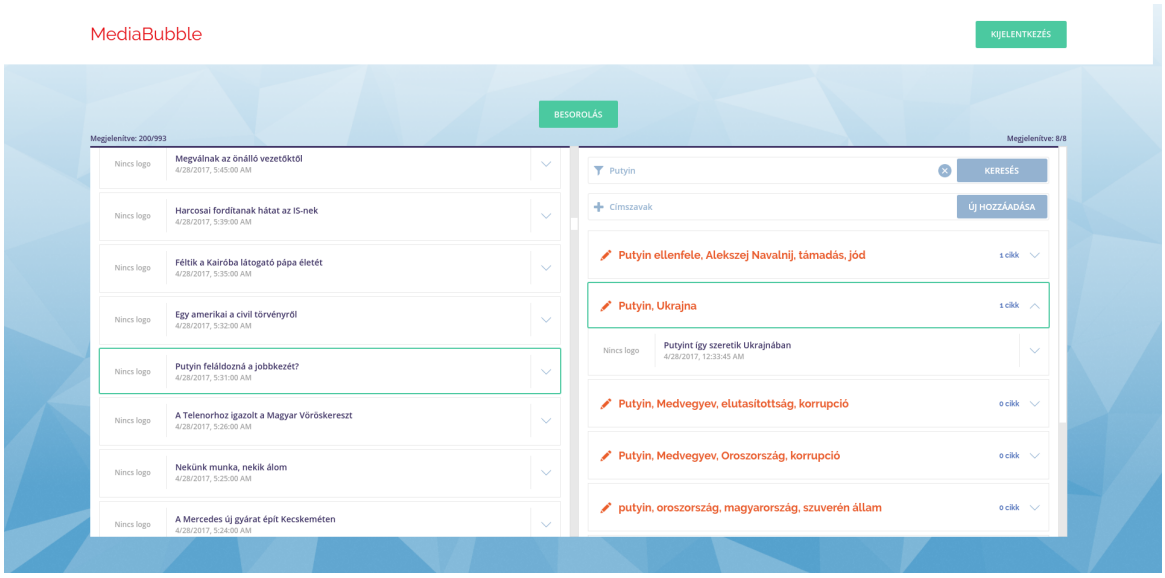


Figure 1: The user interface of the annotation tool.

4. The Annotation Process

The annotation process has been initiated involving 3 annotators. All of the annotators are highly educated and are outside of the computer science field. Before conducting the annotations, the goal and scope of the MediaBubble project has been described to them. The concrete meaning of the term ”topic” has not been described explicitly and typically has been referred to in an intuitive manner as our intention was not to influence the annotators into any specific direction.

Figure 1 presents the user interface developed specifically for the MediaBubble Dataset project. The annotation interface can be reached by conducting a login process. This is how the current annotator is identified. The user interface is clean and provides essential tools to support the annotation process. The main components of the user interface are the article list on the left and the topic list on the right. A counter is presented on the top of both lists to inform the annotators about the status of their annotation roadmap.

The article list contains the articles to be annotated. The list of the articles is fixed. Each article is presented with its title and the date of its appearance. In the case the title does not contain sufficient information, in order to get detailed information, the down arrow on the right of the title is to be clicked. The detailed view shows the abstract of the article and contains a link to display the article in its original location.

The topic list contains the topics the articles can be assigned to. The detailed view of a topic shows the assigned articles in a list below the title of the topic. The detailed view can be displayed by clicking the down arrow on the right of the title of the topic. Unlike the list of articles, the list of topics can be altered by the annotator. In a typical annotation work-flow, the annotator processes the articles sequentially. In the case no corresponding topic is available, the annotator can create a new topic by specifying its title. The title is used to help the annotators to identify the topic. The topics are shared among the annotators. In order to help the actual annotator to find a particular topic, a keyword-based search tool is created. By entering a search term, the annotator can filter the list of topics to the topics containing the search term.

An article can be selected by clicking on its title. The selection is indicated by an emphasizing border. The corresponding topic can be selected similarly, by clicking on its title. Having both items selected, the annotator has to click the assign button in order to finalize the topic assignment. In the case, the user clicks on an already assigned article, the title of the assign button changes to remove. The purpose of this button is to let the annotator undo mistaken assignments.

The novel contributions are to be conducted via the user interface described above.

4.1. The Statistical Properties of the established Dataset

In order to give a thorough overview of the dataset, the statistical properties of the dataset have been calculated from three different aspects as the empirical distribution of the cluster sizes, the empirical distribution of the agreement levels and the pairwise inter-annotator agreements of the annotators. The annotators are presented anonymously and are denoted with letter A, B and C.

Table 2 presents the count of topic sizes per annotator. The column ”Topic Size” denotes the size of the topic. Columns denoted as ”Annotator X” present the amount of topics of the corresponding size in the case of the specific annotator. The primary property of the dataset is that there is a specific amount of single articles assigned to a separate topic. On the other side, topics over size 10 are represented sparsely in this dataset. Topics containing only one article

are validated. The purpose of these single articles is to provide control on false topic assignments of machine learning based methods.

Topic size	Annotator A	Annotator B	Annotator C
1	348	417	328
2	69	92	72
3	39	35	37
4	24	23	16
5	11	14	12
6	14	6	13
7	11	5	9
8	4	0	7
9	2	3	2
10	1	0	0
11	0	2	1
12	1	1	2
13	1	0	1
14	0	0	1
15	0	0	1

Table 2: Empirical distribution of the topic size per annotator.

In order to have an overview on the topic assignments of the articles, the annotator agreement level (AAL) is calculated for each article. The articles assigned to the same topic by all the annotators are denoted as having AAL 3. The articles having a majority vote meaning that two of the three annotators vote for the same topic. These articles are denoted as having AAL 2. Those articles assigned to three different topics by the annotators are marked as having AAL 1.
Having the measures calculated, the articles are separated into three different sets based on their AAL assignment. Table 3 presents the histogram of the articles regarding the AAL value. The column "Annotator Agreement Level" contains the AAL value. The column "Count of Articles" presents the amount of articles having the particular AAL value. Considering that a typical numerical experiment in the topic detection domain involves majority voting to determine the final / aggregated topic assignment of an article, the amount of articles having a final assignment is 736 which is 74% of the sample. This value could be looked upon as a kind of confidence level of the dataset, thus the dataset shows potential of be involved into further research projects.

Annotator Agreement Level	Count of Articles
3	222
2	514
1	264

Table 3: Count of articles per agreement level.

To analyze the topic assignments from the aspect of the annotators, Table 4 presents the pairwise inter-annotator agreement. As mentioned in the beginning of this section, 3 annotators are involved in the experiment. The column "Annotator 1" and the column "Annotator 2" contains the annotators. The column κ contains the inter-annotator agreement level of the two particular annotators. In order to present the results anonymously, the concrete annotators are denoted with A, B and C.

Annotator 1	Annotator 2	κ
A	B	0.466
A	C	0.414
B	C	0.423

Table 4: Pairwise inter-annotator agreement.

The Cohen's kappa coefficients are in the interval $(0.4, 0.6]$ indicating a moderate agreement of the annotators.

5. Conclusion

In this paper, we have presented a first initiative to establish a political *framing* data set for the Hungarian language. The end-goal with this data set is to develop an application that gives the reader the possibility to read about topic of interest with different political aspects.
As our annotator resources are limited, the MediaBubble Dataset has been set up as a crowdsourcing dataset. The dataset is available for download for research purposes with the restriction that if someone would like to have access to the dataset, a specific amount of annotation has to be conducted as a contribution to the research community. Having the annotation task completed, the dataset is available for download. Details on downloading and contribution can be found on the homepage (Grad-Gyenge, 2017) of the dataset.
Our plans for the future can be described as the following. At first, we would like to emphasize the visibility of our initiative. We hope that crowdsourcing will be a potential technique to emphasize the size and the quality of the dataset. At second, we would like to involve the dataset into the development of novel semantic representation techniques especially for the Hungarian language.

Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., and Hughes, M. (2013). Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 49–58, Atlanta, Georgia, June. Association for Computational Linguistics.

Baumer, E., Elovic, E., Qin, Y., Polletta, F., and Gay, G. (2015). Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482.

Nicoletta Calzolari, et al., editors. (2016). *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. ACL.

Card, D., Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. In *Proceedings of the*

53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers, pages 438–444. The Association for Computer Linguistics.

Fulgoni, D., Carpenter, J., Ungar, L. H., and Preotiuc-Pietro, D. (2016). An empirical exploration of moral foundations theory in partisan news sources. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.* European Language Resources Association (ELRA).

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience.* Harvard University Press.

Grad-Gyenge, L. (2017). The MediaBubble Dataset. Available at: `http://laszlo.grad-gyenge.com/#!/mediabubble`.

Scheufele, D. (1999). Framing as a theory of media effects. *Journal of Communication*, 49(1):103–122.

Tsur, O., Calacci, D., and Lazer, D. (2015). A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *ACL (1)*, pages 1629–1638.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm*, 10(1):178–185.

Williams, C. and Gulati, G. (2008). What is a social network worth? facebook and vote share in the 2008 presidential primaries. American Political Science Association.

Preservation of Original Orthography in the Construction of an Old Irish Corpus.

Adrian Doyle, John P. McCrae, Clodagh Downey
National University of Ireland Galway
a.doyle35@nuigalway.ie, john.mccrae@insight-centre.org, clodagh.downey@nuigalway.ie

Abstract

Irish was one of the earliest vernacular European languages to have been written using the Latin alphabet. Furthermore, there exists a relatively large corpus of Irish language text dating to this Old Irish period (c. 700 – c 950). Beginning around the turn of the twentieth century, a large amount of study into Old Irish revealed a highly standardised language with a rich morphology, and often creative orthography. While Modern Irish enjoys recognition from the Irish state as the first official language, and from the EU as a full official and working language, Old Irish is almost incomprehensible to most modern speakers, and remains extremely under-resourced. This paper will examine considerations which must be given to aspects of orthography and palaeography before the text of a historical manuscript can be represented in digital format. Based on these considerations the argument will be presented that digitising the text of the Würzburg glosses as it appears in *Thesaurus Palaeohibernicus* will enable the use of computational analysis to aid in current areas of linguistic research by preserving original orthographical information. The process of compiling the digital corpus, including considerations given to preservation of orthographic information during this process, will then be detailed.

Keywords: manuscripts, palaeography, orthography, digitisation, optical character recognition, Python, Unicode, morphology, Old Irish, historical languages

1. Introduction

Encoded within the original, handwritten text of the Würzburg glosses, the earliest large collection of glosses written in the Irish language, is a wealth of scribal knowledge. This paper argues for the preservation of this knowledge in the creation of a digital corpus of Old Irish text by faithfully representing the original orthographic features of the glosses.

An argument will be made in favour of focusing on material found only in manuscripts contemporary with the Old Irish period of c. 700 – c. 950. This paper will next outline why the text of the Würzburg glosses as it appears in *Thesaurus Palaeohibernicus* (Stokes and Strachan, 1901) is the best candidate for digitisation. Finally, the automated digitisation and proofing process of the corpus will be detailed.

2. Irish

Irish was one of the earliest vernacular European languages to have been written using the Latin alphabet. Thurneysen describes the language as "the earliest form of a Celtic language which can be more or less completely reconstructed from extant sources" (1946, p.1). There are many accepted stages in the development of the language from its earliest attested form to that which is spoken today. Of these stages, the earliest three, Primitive, Old and Middle Irish, are collectively referred to as Early Irish. Consisting, for the most part, of personal names, generally in the genitive case, engraved on standing stones utilising the Ogham alphabet, Primitive Irish is poorly attested compared to Old and Middle Irish.

2.1 Old Irish

A decent number of Old Irish texts survive into the modern period, among these the Old Irish Glosses; Würzburg, Milan and St. Gall. Despite the availability of textual source material, however, it is not necessarily useful to treat all texts written in Old Irish as equal. Stifter distinguishes, for example, between *Classical Old Irish* and *Late Old Irish* based on "linguistic variation [within the Old Irish glosses]" (2006, p.10). Before any text can be deemed suitable for inclusion in a digital corpus of Old Irish, it is first necessary that the term, Old Irish, be examined.

McCone notes that "some scholars have been wont to recognise four main phases [in the evolution of Irish]," (1997, p.163), whereby Old Irish can be understood as the language attested "from roughly the beginning of the 8th century to the middle of the 10th century A.D." Material preserved in manuscripts dated within this Old Irish period "is inevitably the corpus from which the norms of Old Irish grammar have been established in the first instance by modern scholarship" (McCone, 1997, p.164), and it is upon such material that Rudolf Thurneysen based his seminal work, *A Grammar of Old Irish* (1946). The surprisingly high degree of uniformity apparent in these texts suggests that Old Irish must have existed as "a literary language whose standard was taught to the Irish 'men of writing' in school, much as standardised Latin was taught to Continental pupils as a language of literary communication, long after Classical Latin had ceased to be a spoken language of the people" (Stifter, 2006, p.10). McCone asserts that "Old Irish can be defined linguistically in terms of a wide range of specific grammatical traits that together constitute a distinctive system" (1997, p.165).

If "essential conformity to the appropriate criteria… constitute grounds for describing … a text as Old Irish, regardless of the date of the manuscript in which it is preserved" (McCone, 1997, p.165), however, it follows that texts such as *Críth Gablach*, surviving in manuscripts dated later than the end of the Old Irish period, and even some written in the modern day, could be described as Old Irish provided they fall within the prescribed linguistic parameters. This notion, raises an issue regarding the potential inclusion of such texts in a digital corpus. A text composed later than the Old Irish period will be more reflective of a scribe's own understanding of an already archaic literary standard than it will be of the standard itself. Even text copied from earlier sources may be unreliable as "Middle-Irish transcribers have often modernised or corrupted these ancient documents" (Stokes and Strachan, 1901). While McCone lambasts "attempts at a more or less clear chronological definition of Old, Middle

and Modern Irish" (1997, p.165) citing "arbitrary transitional dates" as cause for concern, he concedes that material dating to within the Old Irish period "alone can be safely assumed to be free of the possible distortions of significantly later recopying" (McCone, 1997, p.164). A further issue with the inclusion of texts in a digital corpus of Old Irish based on their conformity to outlined linguistic criteria is that it begs the question, how much deviation from these criteria is too much? Despite the high degree of linguistic uniformity apparent in texts preserved in manuscripts dated earlier than the 10th century, McCone (1985) outlines several examples of deviation from the Old Irish norm already apparent in some of the earliest textual sources of Old Irish, including the Würzburg glosses. These deviations, McCone argues, are more consistent with linguistic developments associated with the subsequent Middle Irish period. Outlining hard linguistic criteria with which to justify a given text's inclusion in, or exclusion from a digital corpus is beyond the scope of this paper, and in any case, this practice may limit the utility of the corpus to researchers. Current research projects such as Chronologicon Hibernicum (Stifter, 2015), and LexiChron (Toner and Han, 2018), focus on linguistic features of select texts in order to establish reasonable means by which to linguistically date others. For these reasons this paper will focus on Old Irish text which is preserved only in manuscripts contemporary with the Old Irish period of the 8th to the middle of the 10th century, and will not exclude any such material based on linguistic criteria.

2.2 Old Irish Text and Resources

There are three large sources of Old Irish text which survive in manuscripts dated to within the Old Irish period. These are collectively known as the Old Irish Glosses. These consist of three large collections of interlinear and marginal glosses on Latin texts. The earliest of these, dated to the middle of the 8th century (Stifter, 2006), are the Würzburg glosses on the Pauline epistles. From the early 9th century come the Milan glosses on the psalms, and from the middle of the 9th century come the St. Gall glosses on the Priscian grammar of Latin. Projects undertaken by Dr. Aaron Griffith (2013) of the University of Vienna, and Dr. Pádraic Moran (2014) of the National University of Ireland, Galway have already collected, and published in digital format, the text of the Milan and St. Gall glosses respectively. While Kavanagh and Wodtko (2001) have produced a lexicon based on the Würzburg glosses, no collection has been published in digital format to date. For this reason, this project has been focused on the process of digitising the text of the Würzburg glosses.

Of the glosses which have been digitised, Moran (2014) suggests that St. Gall contains about 9,400 glosses, over a third of which are written in Old Irish. These do not equate to full sentences, as many glosses are fragmentary, or contain single words or phrases. Nonetheless, assuming a similar number are present in the Milan glosses, that brings the extant digital corpus of Old Irish to only about 6000 glosses. There currently exists no part-of-speech (POS) tagged corpus for any complete set of glosses. In fact, POMIC (Lash, 2014), a collection of fourteen Old and Middle Irish texts, contains the only currently available POS tagged text in Old Irish. While this provides an excellent resource for computer-based Early Irish research, texts which match this paper's definition of Old Irish are few, and those which have been POS tagged are fewer. As such, Old Irish remains highly under-resourced.

3. Old Irish Orthography and Palaeography

Having settled upon an appropriate Old Irish corpus, namely the Würzburg glosses, consideration must next be given to the source of text which will be drawn upon. As will be demonstrated in this section, drawing upon the original text as it appears on the folio would present many technical issues resulting from the original orthographic stylings of Old Irish scribes. In many cases, modern editors cannot preserve characters of the original manuscript script, and hence, must make emendations which alter the orthography of resultant modern editions.

The insular script employed by Irish scribes utilises a selection of variations on Latin alphabetical symbols. Many of its distinctive letters, diacritics and symbols, such as "ɔ", "Ṡ", and "Ḟ", are supported by Unicode. As such, much of the orthography of Old Irish text can be represented digitally. Nevertheless, a variety of contraction markers which remain unsupported by Unicode, and which are used throughout the Old Irish glosses, prevent them from being perfectly represented by Unicode characters alone. These abbreviating contractions come in many forms, and are used in place of the plene spelling of a word. One common example is the suspension stroke which can be used in combination with a variety of different letters to produce various differing sounds. Combination with the letter "b", for example, could produce the sound "bar". Hence, words like "Conchobar", could be written out in full, or contracted, "ɔchob" with a suspension stroke over the "b". As such, the use of contractions in Old Irish text saves valuable space on vellum. Editors compiling modern editions may opt to represent such contractions by supplying the full plene spelling in their place. Such is the case with the two-volume collection of Early Irish texts, *Thesaurus Palaeohibernicus* (TPH) (Stokes and Strachan, 1901; Stokes and Strachan, 1903) Importantly for the purpose of this project, the editors of TPH retain many diacritics and symbols such as those outlined earlier. Moreover, where orthographic features could not be retained, the editors identify plene text which they have supplied. Therefore, by drawing upon the text as it appears in TPH, it is possible to digitise the contents of the Würzburg glosses extremely faithfully, without sacrificing important elements of the source material's orthography. This, in turn, will allow for statistical linguistic analysis to be carried out on a digital corpus of Old Irish text which represents the language in as close a manner as possible to its original format.

4. Digitisation of *Thesaurus Palaeohibernicus*

Both volumes of TPH were initially captured using a Kirtas (Kirtas, 2015) scanner with APT manager software, and edited with Book Scan Editor software. At this point, ABBYY FineReader (ABBYY, 2018) OCR software was utilised to recognise the text in the captured image files. The output of this process was a machine-readable PDF file containing both the image, and digital text of the entire two volumes of TPH.

The character recognition, while generally successful on the English language content, apparently had difficulty

with the Latin, and particularly with the Irish text. Footnote markers were regularly mistaken for a variety of characters not present in the hard copy, including "®", "*", and "^". Diacritic markers present in the Irish text posed a similar problem. Often, acute accents were represented mistakenly as umlauts, for example "domsa höre" for "domsa hóre". Even in the English text where character recognition had been generally better, the regularity with which characters were mistaken warrants strict proofing of each line. Examples of such mistakes include, "…because 1 believe…", and "he vvho shall believe". Information originally appearing in marginalia, such as folio and line numbers, were often combined erroneously with linguistic text.

4.1 Automated Analysis of OCR Success

A number of Python scripts were written to measure the general success of the OCR process. Initial efforts were focussed on measuring the success of character recognition in page headers, as these contain TPH page numbers. The page range of the Würzburg glosses in TPH spans from 499 to 712. The first script written checked to see if all of these page numbers exist, in sequence, within the digital text. This found that 28 of 214 page numbers, roughly 13%, had been incorrectly digitised. Once identified, these missing page numbers were manually corrected.

The next concern was to discover how many page headers had their textual content correctly digitised. Another Python script identified 27 page headers, about 12.6%, which had been incorrectly digitised. Again, these were manually corrected.

4.2 Automatic Approach to Proofing

With all page headers and numbers now correctly in place, a script was written to count each line of text per page and represent it as a sequentially increasing decimal following the relevant page number. For example, the title line on the first page of the Würzburg glosses would be indexed at 499.1. This new indexing system allows for quick comparison between the often difficult to recognise digitised text and its equivalent text in TPH. This has increased the speed with which the digital text can be proofed by eliminating excess time spent attempting to recognise a given line of text.

At this point the text was still interspersed with arbitrary characters where diacritics and footnotes had been incorrectly assigned. These characters made manual proofing a cumbersome task. A script was written to replace any such unexpected characters with a single underscore, this would serve as a clear signal to a proof reader that something had been removed, and hence, a given section of text would require particular attention.

4.3 Preservation of Information

TPH contains a variety of metadata related to the text on a given page. Page headers, mentioned earlier, contain not only TPH page numbers, but also information pertaining to the content of the text on the page. Even numbered page headers read "Biblical Glosses and Scholia", a common theme throughout the first volume of TPH. Between pages 499 and 712, odd numbered page headers read "Glosses on the Pauline Epistles." followed by information on the specific letters referenced on the page, beginning "Rom. I." on page 499 and working through to "Heb. V, VI." on page 711. It is a simple matter, therefore, to create a Python dictionary into which page number and content data can be automatically collected as keys and values respectively. Similar information is contained in secondary titles on pages where a new set of letters begin. A primary section title, "CODEX PAULINUS WIRZIBURGENSIS", is given on page 499. Such titles, when encountered during proofing, are surrounded by square bracket tags, [H2][/H2] and [H1][/H1] respectively, in order to enable automatic identification of them at a later stage.

The text presented on a typical TPH page is split into three sections. The first, presented at the top of each page, is the Latin text of the Pauline Epistles. Only lines containing glosses are included, and the point within a line of text to which a gloss corresponds is marked with a superscript number. These footnote-style numbers caused particular trouble during the OCR process. No instance of these was correctly digitised. In proofing these are replaced with the same number, enclosed within square brackets. The section itself is also enclosed within tags which identify it as the original Latin text, [Lat][/Lat].

The second section, positioned in the middle of each page, contains the text of the glosses which relate to the Latin text above. Each gloss is numbered in accordance with the superscript numbers of the above Latin section. These glosses are written in a combination of Latin and Old Irish, with code switching occurring regularly. The editors of TPH distinguish between the two languages by printing Irish content in italics, while leaving Latin text unaltered. Where part of an Irish word has been supplied by editors in place of a manuscript contraction, this is identified by the editors by returning to roman type. Such supplements are surrounded by contraction tags, [Con][/Con], during the proofing process to preserve metadata relating to breaks from original orthography. Similarly, letters supplied by the editors but omitted in the manuscript are identified in TPH by square brackets. In proofing these are replaced by supplement tags, [Sup][/Sup].

Like the Latin section above, the glosses are surrounded with [SG][/SG] tags, identifying the section as a whole. However, Latin content within the section is separated from Irish content by means of separate Latin tags, [GLat][/GLat], which surround uninterrupted strings of Latin text, as well as individual instances of Latin abbreviations such as ".i." and "Ł" which frequently appear in TPH. The Latin tags used within the glosses' section are distinct from those used earlier to ensure that the Latin content of each section can be automatically identified as separate. Within this section footnotes are marked out by means of superscript alphabetical letters. These are matched in a footnote section at the bottom of each page. As with the superscript numbers of the Latin section, these markers caused difficulty for the OCR software and none were correctly identified. In proofing these are replaced by the same letter enclosed within square brackets. In instances where the footnote suggests that the editors have emended a manuscript form, or supplied a form not present in the manuscript, the word is surrounded by opening and closing tags bearing the letter of the relevant footnote, for example, [a][/a]. This will allow the original manuscript orthography to be automatically restored. The tag-set utilises square brackets so that single-letter tags such as these will not be mistaken for html tags identifying elements such as hyperlinks, bold text, or paragraphs.

The third section, towards the bottom of a page, placed just above the footnotes, provides an English translation of the

Irish gloss content. Where present the Latin gloss content is left untranslated, however, much of it is simply removed. Footnotes continue into this section from the preceding section of glosses, and are treated in the same manner. The section is enclosed within tags, [Eng][/Eng], which identify its content as the translation of the glosses above.

Information regarding the location of a given page's text within the original manuscript is given in the outer marginalia of each page, to the left or right of the block of text to which they refer. Information supplied here includes the folio number, and a letter corresponding to the column on that folio, from which the text was taken. This folio information, regularly combined mistakenly with the main body of text during the OCR process, is removed during proofing and replaced with folio tags which surround the relevant blocks of text, for example, [f. 1a][/f. 1a].

The preservation of this metadata by means of a specialised tag-set creates a number of possibilities for researchers (Petrova, et al., 2009). The original text can be drawn upon as easily as the text which appears within the pages of TPH. Moreover, the identification of original orthographic details by means of tags allows for statistical analysis of variant spellings and word choices which may be useful to researchers in the identification, by computational means, of different scribal hands, linguistic registers, and dialect within the glosses.

5. Further Use of the Digital Corpus

As this paper is being written, proofing of the text content is ongoing. Once this process has been completed, focus will shift to POS and dependency tagging of the glosses, after which the corpus will be made available online. Ultimately, it is expected that this corpus will aid researchers in the field of Early Irish by allowing automation of a variety of research tasks, a possibility first proposed by Teresa Lynn (2012).

6. Conclusion

In creating a digital corpus for a historic language, preservation of the original orthographical content enables significant forms of text analysis to be performed on the resultant digital corpus. This paper advocates careful selection of source material, such as *Thesaurus Palaeohibernicus* (Stokes and Strachan, 1901; Stokes and Strachan, 1903), which, where possible, carefully preserves distinct orthographic diacritics and symbols where present in the original manuscript. A method is outlined for the preservation of metadata relating to original orthographical features of manuscripts where editors have been unable to preserve the features themselves in their edition.

In the case of Old Irish, it is envisioned that the production of this digital corpus will aid in research tasks which rely on the study of orthographical features by allowing the automation of tasks dependent on these features. Such tasks may include identification of different scribal hands, identification of linguistic register or dialect, and linguistic dating, where such tasks may be based on the frequency or location of orthographical features within a text.

This paper has shown that the speed with which Old Irish text can be digitised can be significantly increased by the combined use of OCR software with a variety of techniques intended to improve the proofing process.

A tag-set has been created which will be used to identify features within the digital corpus including original folio information, points of scribal contractions, text supplied by editors, code switching between Irish and Latin, editorial emendations of provided manuscript forms, as well as headers, sections and footnotes present in the source material. As the text requires proof reading, implementation of this new tag-set will be carried out in tandem with this process. Therefore, time taken to produce this digital corpus will not be significantly increased by its introduction.

7. Acknowledgements

This research is supported by the National University of Ireland, Galway's DAH (Digital Arts and Humanities) scholarship and by Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

8. Bibliographical References

Griffith, A. (2013). A Dictionary of the Old-Irish Glosses. http://www.univie.ac.at/indogermanistik/milan_glosses.htm (Accessed: 10/01/2018).

Kavanagh, S. & Wodtko, D.S. (2001). A Lexicon of the Old Irish Glosses in the Würzburg Manuscript of the Epistles of St. Paul. Verlag der Österreichischen Akademie der Wissenschaften, Vienna.

Lash, E. (2014). The Parsed Old and Middle Irish Corpus (POMIC). Version 0.1. https://www.dias.ie/celt/celt-publications-2/celt-the-parsed-old-and-middle-irish-corpus-pomic/ (Accessed: 10/01/2018).

Lynn, T. (2012). Medieval Irish and Computational Linguistics. *Australian Celtic Journal*, 10:13-28.

McCone, K. (1985). The Würzburg and Milan Glosses: Our Earliest Sources of 'Middle Irish'. *Ériu*, 36:85-106.

McCone, K. (1997). The Early Irish Verb. An Sagart, Maynooth, 2nd edition.

Moran, P. (2014). St Gall Priscian Glosses. http://www.stgallpriscian.ie/ (Accessed: 10/01/2018).

Petrova, S., Solf, M., Ritz, J., Chiarcos, C. & Zeldes, A. (2009). Building and using a Richly Annotated Interlinear Diachronic Corpus: The Case of Old High German Tatian. Traitement automatique des langues, 50(2), 47–71.

Stifter, D. (2006). Sengoidelc. Syracuse University Press, New York.

Stokes, W. & Strachan, J. (Eds.). (1901). Thesaurus Palaeohibernicus Volume I. The Dublin Institute for Advanced Studies, Dublin, 3rd edition.

Stokes, W. & Strachan, J. (Eds.). (1903). Thesaurus Palaeohibernicus Volume II. The Dublin Institute for Advanced Studies, Dublin, 3rd edition.

Thurneysen, Rudolf. (1946). A Grammar of Old Irish. The Dublin Institute for Advanced Studies, Dublin.

9. Language Resource References

ABBYY. (2018). FineReader. https://www.abbyy.com/en-eu/finereader/ (Accessed: 10/01/2018).

Kirtas. (2015). https://www.kirtas.com/ (Accessed: 10/01/2018).

Stifter, D. (2015). Chronologicon Hibernicum. http://dhprojects.maynoothuniversity.ie/chronhib/ (Accessed: 10/01/2018).

Toner, G. & Han, X. (2018). LexiChron. https://www.qub.ac.uk/schools/ael/Research/Languages/LexiChronProject/ (Accessed: 10/01/2018).

Krik: First Steps into Crowdsourcing POS tags for Kréyòl Gwadloupéyen

Alice Millour, Karën Fort
Sorbonne Université, STIH - EA 4509, Paris, France
alice.millour@etu.sorbonne-universite.fr, karen.fort@sorbonne-universite.fr

Abstract

This article presents the adaptation to Guadeloupean Creole of a project of crowdsourcing part-of-speech (POS) tags initially designed for a French regional language, Alsatian. We do not detail here the specifically developed crowdsourcing platform and methodology, but rather focus on the construction of the required elements for a language to be a candidate for this task: i) an open-source raw corpus, ii) a tokenizer, iii) adapted annotation guidelines, iv) a minimal reference, and, preferentially, v) one or two baseline tagger(s). After describing the preliminary work we have carried out for Guadeloupean Creole to comply with these prerequisites, we present the first results on crowdsourcing POS tags through the platform specifically developed for this task: `Krik`.

Keywords: Guadeloupean Creole, crowdsourcing, POS tagging, less-resourced languages

1. Introduction

Despite the progress made in unsupervised learning, manually annotated corpora are still necessary both to develop and to evaluate natural language processing (NLP) tools. However, building such corpora is notoriously expensive (see, for example, Böhmová et al. (2001)). For less-resourced languages, the (lack of) availability of language experts represents yet another obstacle to overcome. However, *a priori* non-expert speakers can be solicited online to share their linguistic knowledge and thus participate in the creation of resources for their language. To take advantage of this potential, we have designed a lightweight crowdsourcing platform enabling both the training of the participants to the task of part-of-speech (POS) tagging and the collaborative annotation of open-source corpora.

We led our first work on crowdsourcing POS tags on a Germanic French regional language: Alsatian (Millour and Fort, 2018).[1] The results obtained being promising, we tested the portability of our approach by adapting it on another less-resourced French regional language: Guadeloupean Creole (GC). This adaptation requires the availability of: i) a freely available raw corpus, ii) a tokenizer, iii) adapted annotation guidelines, iv) a minimal reference and, if possible, v) at least one baseline tagger for the language considered.

After presenting the existing resources for GC, we describe the five steps of the preparatory work we had to perform for GC to be a candidate language for the crowdsourcing task. Finally, we present the very preliminary results of the latter and discuss the perspectives.

2. Related Work

2.1. Guadeloupean Creole

Guadeloupean Creole (GC) is a French-based Creole spoken in the French department and archipelago of the West Indies: Guadeloupe. GC accounts for around 600,000 speakers (400,000 in Guadeloupe, and approximately 200,000 elsewhere (Colot and Ludwig, 2013)). GC is very close to the other main variety of Antillean Creole: Martinicain Creole (MC). Yet some lexical and morphological features distinguish them (see for instance the personal pronouns "*man*"/"*an*" in MC, "*moin*"/"*mwen*" in GC for the first person singular pronoun "I", or the possessive pronouns "*fidji**'w***" in MC, "*figi **a'w***" in GC ("**your** face")). What is more, GC presents a greater linguistic variation as a result of its less compact geography (Observatoire des pratiques linguistiques, 2005). Additionally, no spelling standard is recognized as the legitimate norm among speakers. Two main spelling systems coexist: one has been developed by the GEREC-F[2] (Ludwig et al., 1990), and later modified by Bernabé (2001), the other has been introduced by Hazaël-Massieux (2000). In particular, no agreement has been reached regarding the positioning towards French orthography when it can be invoked. For instance, both the forms "*chien*" (French for "dog") and "*chyen*" can be found in GC. Similarly, "*latè*" is the agglutination of the French determiner "*la*" ("the") and proper noun "*Terre*" ("Earth"). It is generally perceived as a unique entity, meaning "Earth" as a whole, and is consequently written as such. Still, we have found occurrences of the separated form, which is considered as erroneous by creolists. Generally speaking, we have encountered in our yet relatively small corpus a great variety of spelling alternatives, regardless of the conventions suggested by the two main standards. For instance, the use of the hyphen between nouns and postponed determiners (e.g. "*tifi-la*" or "*tifi la*" ("the young girl")), or the suppression of the space between adjectives and nouns in some cases (e.g. "*jenn fi*", "*jenn-fi*, "*jennfi*" ("young girl") (Delumeau, 2006)), are not consistent across the corpus.

The case of "*a pa*"/"*apa*" also exemplifies the poor penetration of the standards among the speakers. While Ludwig et al. (1990) introduced a graphic convention to distinguish "*a pa*" (negative existential) found in context such as "*A pa pas ou ni lajan* [...]"[3] ("Not because you have money [...]"), from "*apa*" ("apart (from)"), two out of three GC speakers we have been working with had never encountered the separated form.

Furthermore, GC presents a reduced inflectional and derivational morphology: the plural is indicated only by

[1] See: `http://bisame.paris-sorbonne.fr`.

[2] The GEREC-F (Groupe d'Études et de Recherches en Espace Créolophone et Francophone) is the investigation group for Creole and French speaking areas.

[3] This example was taken from (Delumeau, 2006).

the particle "*sé*" ("*timoun-la*" ("the child"), "*sé timoun-la*" ("the children")), the verbal lexeme is mainly invariable, and tenses and aspects are marked by combinations of particles. It makes it impossible to identify the part-of-speech of some words independently of the context: for instance, "*manjé*" can both mean "eat" (in its infinitive and conjugated form) and "food".

2.2. Existing Resources for GC

Some work can be found regarding GC processing. For instance, Delumeau (2006) introduces a linguistic description for GC in a natural language generation perspective, Carrión Gonzalez and Cartier (2012) detail the existing lexical resources for various French-based Creoles, Schang (2013) presents a metagrammar for GC, and Schang et al. (2017) describes the result of the annotation of coreference relations of a transcribed spoken GC corpus (the same we use here).
Yet, to our knowledge, no POS tagged corpus or tagger was available until now.

3. Methodology

3.1. Raw Corpus

To ensure the further availability of the annotated resources produced through the platform, we have focused on gathering a freely available corpus, which can be described as "opportunistic" (McEnery and Hardie, 2011), thus introducing a bias in term of content. In fact, our corpus is made of texts gathered from two sources:

- The COCOON[4] database, which contains 11 transcripts[5] of conversations led in GC (we actually used 10 out of them, for the 11^{th} contained too many French utterances), available under the CC BY-NC-SA license.[6]
- Wikipedia: we collected the proverbs found on the French page for GC[7], and the 17 articles from the Wikimedia incubator for a Guadeloupean Creole encyclopedia.[8] This corpus, C_{Wiki}, contains 74 sentences adding up to 873 tokens.

3.2. Tokenizer and Annotations Guidelines

For the sake of adaptability, we chose to work with the universal POS tagset presented in (Petrov et al., 2012), which synthesizes the tagsets of 22 languages and can be adapted to the specificities of each language.[9] Initially, the only modification we made was to have the X category ("Others", a catch-all category hard to interpret) to match only the cases of code-switching, which can not be analyzed as loan words. Eventually, just as for Alsatian, we had to further enrich this tagset with four additional categories described hereafter. The refinement of the tagset, the adaptation of the tokenizer, the elaboration of the guidelines and the building of the reference are simultaneous processes including back and forth adjustments.
The tokenizer, initially developed for Alsatian, has been provided by D. Bernhard (LiLPa, Université de Strasbourg) and adapted to the specific needs of Creole. Two kinds of operations were added to the classic tokenization process:

- Merging: decided when space-separated tokens matched a unique morphological entity. For instance, the sequence "*ki jan*" (meaning "how", literally "which kind") only appears in its separated form in our corpus. Although one could be tempted to annotate "*ki*/PRON *jan*/NOUN", this goes against the intuition of native speakers. We thus created the token "*ki_jan*/ADV". The same operation was led on the equally more intuitive "*ki_tan*/ADV" ("when", literally "which time"), "*ki_koté*/ADV" ("where", literally "which side"), etc. Prepositional locutions such as "*a fòs*" ("by dint of") were also merged for annotation consistency reasons.
- Splitting: applied when punctuation-separated tokens matched a sequence of two morphological entities understandable as such on their own. This case is exemplified by the cases of postponed determiners "*-la*" (definite article) and "*-lasa*" (demonstrative determiner), which are stick to the noun they determine in their usual form (e.g. "*Egliz-lasa*", "this Church").

Note that we did not split the tokens containing an apostrophe, indicating a contraction, but which refer to a sole interpretation for native speakers. This is the case for the tokens such as "*k'ay*/PART+VERB", contraction of "*ka*" (particle for the present tense) and "*ay*" (3^{rd} person singular for the verb "have"), for which the tokenization "*k' ay*" makes the reading and understanding confusing. For the same reason, tokens involving pronouns such as "*ba'y*/ADP+PRON" ("for him/her"), "*trapé'y*/VERB+PRON" ("catch him/her"), or "*sa'w*/PRON+PRON" (contraction of "*sa*" ("this") and "*ou*" ("you")), were not split.
These considerations resulted in the addition of 4 new categories to the universal tagset: PRON+PRON, PART+VERB, ADP+PRON and VERB+PRON.
The tagset we present here matches the needs encountered in our reference corpus. It should then not be considered as definitive, as the corpus we managed to gather is far from representative of all spelling habits and variants existing in Guadeloupe.
The annotation guidelines, inspired from the TCOF-POS (Benzitoun et al., 2012) guidelines, were developed to accompany both the expert annotators and the non-expert participants of the crowdsourcing project. For that reason, we followed the methodology set up for the crowdsourcing experiment on Alsatian and opted for a description of

[4] *COllection de COrpus Oraux Numériques* (Collection of digital oral corpora), see: https://cocoon.huma-num.fr/.
[5] See for instance: https://cocoon.huma-num.fr/exist/crdo/meta/crdo-GCF_1022. The full list of transcripts can be accessed by clicking on "Guadeloupean Creole French" in the "Langue(s)" section.
[6] See: https://creativecommons.org/licenses/by-nc-sa/3.0/.
[7] See: https://fr.wikipedia.org/wiki/Creole_guadeloupeen.
[8] See: https://incubator.wikimedia.org/wiki/Wp/gcf.
[9] See: http://universaldependencies.org/u/pos/all.html.

ADJ	ADV	ADP	ADP +PRON	AUX	CCONJ	DET	INTJ	NOUN	NUM
5%	7%	6%	0.1%	1%	2%	6%	0.1%	14%	0.2%
PART	PART +VERB	PRON	PRON +PRON	PROPN	PUNCT	SCONJ	VERB	VERB +PRON	X
7%	0.3%	17%	0.2%	3%	10%	3%	17%	0.2%	1%

Table 1: Tag distribution in the reference corpus.

the categories through illustrations in context. We enriched these lists of examples with "Watch out!" sections intended to prevent possible mix-ups and explain ambiguous cases.

3.3. Reference Corpus

We extracted 100 sentences (1,623 tokens) from both C_{Speech} and C_{Wiki} to build the reference corpus to be annotated by experts: C_{Ref}. It contains a sample of declarative, interrogative, imperative, either simple or complex, sentences of different sizes, and of direct and indirect speech.

While the sentences taken from the C_{Wiki} corpus can be immediately used for annotation purposes, we had to carry out some pre-processing on C_{Speech} to obtain grammatically correct sequences of ready to annotate tokens. In fact, the speech dysfluencies are fully transcribed as raw text. As a result, the "speech fragments" very seldom match an understandable utterance, let alone a full grammatical proposition, when taken out of context. As a consequence, we were forced to alter the original corpus in two main ways:

- cleaning up some of the dysfluencies such as the ellipses which resulted in some token being arbitrarily split. In the following example "*gwoka*" meaning literally "big drum", a Guadeloupean music genre:

  1. "*sé pou sa jodijou nou ka respékté gwo...*" ("This is why we respect big...")
  2. "*ka*" ("drum")

  In fact, and although "*gwo ka*" can be found in its space separated form[10], the first utterance is grammatically incomplete. Not to mention that the separated token "*ka*" is ambiguous and could be annotated either NOUN or PART, if presented without any context.

- bringing together the "speech fragments", such as:

  1. "*Lagwadloup dévlopé pli*" ("Guadeloupe has developed more")
  2. "*vit sé on grand tè*" ("fast, it is a big land")

We further split C_{Ref} into two groups, each annotated independently by two annotators (either a GC speaker or expert of the annotation task). The 100 sentences were then manually adjudicated. Table 1 gives the tag distribution across our 100 sentences reference corpus.

3.4. Baseline Taggers

The existing crowdsourcing platform enables participants to correct pre-annotations, thus easing and fastening the annotation process (Fort and Sagot, 2010). Two pre-annotation tools were used for this.

The first pre-annotation tool we developed relies on the weak morphological complexity of GC and uses the 100 most frequent unambiguous tokens of our corpus. This list is undoubtedly not representative of the most frequent words in GC, some common nouns being for instance repeated several times in our corpus and therefore overrepresented. Nonetheless, the most frequent words being also frequent in absolute (for instance, the particle "*ka*" represents 4.6% of the corpus, the pronoun "*an*" (meaning "I") 3.6%, the verb "*sé*" ("be") 2.8% etc.), the basic associative python script we created from this list enabled to annotate 37% of our raw corpus.

Our second pre-annocation tool is the `MElt` tagger (Denis and Sagot, 2012), used without an additional lexicon. To overcome the evaluation bias due to the very small size of our corpus, we split C_{Ref} into ten sets of two subcorpora: $C_{Training,1..10}$ (85 sentences randomly extracted from C_{Ref}) and $C_{Test,1..10}$ (containing the 15 remaining sentences). They were used respectively to train and to evaluate the tagger. We trained `MElt` on the 10 sets and obtained an average accuracy of 82%. The $MElt_{Init}$ tagger was chosen among them.

4. Krik

4.1. The Crowdsourcing Platform

The five requirements having been fulfilled, we provided the dedicated crowdsourcing platform: `Krik`[11] with the required elements. After a training phase of 4 sentences taken from C_{Ref}, which must be entirely properly annotated, the participants access the production phase in which they annotate full sentences extracted from C_{Raw}. This phase is illustrated on Figure 1. Whenever the two pre-annotation tools agree on the annotation for a given token, the consensual tag is suggested to the participant who can either validate or reject it (see on Figure 1 the case "*ou*" ("you") and the suggested tag PRON). When the pre-annotation tools disagree, the two discordant tags are suggested (see on Figure 1 the case "*la*" (postponed determiner) and the suggested tags ADP and DET). In either cases, the full list of tags is available.

4.2. Results

So far, 35 persons created an account on the platform, 17 completed the training phase, and 11 actually produced a total of 1,205 annotations during a period of 9 days. This is far from enough, both in terms of participation and of production.

Still, the annotation on `Krik` resulted in a new, freely available, collaboratively annotated corpus of 74 sentences (698 tokens).

The annotation platform does not compel the participants to annotate every token in the production phase. Thus, we filled the gaps with the $MElt_{Init}$ annotations to obtain consistent tag sequences. This resulted in a new corpus of 933 tokens: C_{Krik}. The addition of this corpus to $C_{Training}$ in the training of the tagger leads to a drop in

[10] This is not the convention used in the corpus we gathered.

[11] See: `http://krik.paris-sorbonne.fr`.

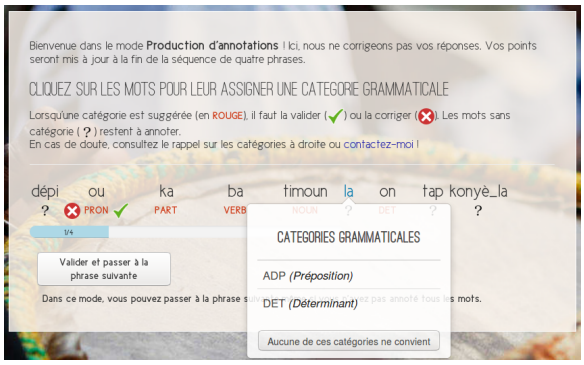


Figure 1: Screen shot of the annotation production phase on the Krik platform.

performance, even though the size of the corpus increases of 62%. This reflects the poor quality of the annotations crowdsourced so far. This is consistent with the low confidence score we calculated for the participants[12], and the difficulties they expressed.
To understand the cause of the errors we manually inspected and corrected the crowdsourced annotations. Among the 124 tokens (nearly 13% of C_{Krik}) that we corrected we identified two main difficulties:

- issues related to the nature of the raw corpus: the C_{Speech} corpus has not been entirely corrected, as described in Section 3.3. This caused the presence in C_{Raw} of unintelligible, hence discouraging, sentences.

  What is more, some additional tokenization problems have been brought to our attention. This explains the thin difference in number of tokens before and after correction. Most of these problems concerned tokens that had to be manually split, but we also encountered tokens as "*anba la*" meaning "down" which must be merged when they are not followed by a noun. We also noticed some spelling mistakes, such as the missing capital letter of "*étazini*" ("United States"), that led the participants to erroneously annotate the token as NOUN instead of PROPN.

- guidelines flaws: the guidelines we initially proposed could not prevent certain mix-ups such as the confusion for "*té*" between the verb and the particle expressing the past. Besides, the case of code-switching remains challenging, especially given the proportion of loan words in GC. During this first experiment, "French words" were alternatively annotated with either the category X or their corresponding category in French (which does not necessarily match the expected tag in GC).

Table 4.2. shows the results of the training of MElt on the corpora described above. The best results are obtained

[12]We do not detail our methodology of evaluation for the users here, for more information, see (Millour and Fort, 2018).

Training corpus	Size (tokens)	Accuracy
$C_{Training}$	1,501	82%
C_{Krik}	933	76%
C_{Krik}+$C_{Training}$	2,434	81%
$C_{KrikCorrected}$+$C_{Training}$	2,439	**84%**

Table 2: Accuracy of the trained MElt taggers.

with the manually corrected corpus $C_{KrikCorrected}$, which reaches a 84% accuracy on C_{Test}.
These results highlight two points:

- The pre-processing of the raw corpus and the annotation guidelines can and must be improved. In fact, these enhancements are compulsory as our experiment shows that a drop in the annotation quality may degrade the performance of the tagger and could remain unnoticed.

- The performance of the tagger could easily be enhanced if more annotations were to be crowdsourced. As already stated by Guillaume et al. (2016) and confirmed by our own experience on Alsatian (Millour and Fort, 2018), quality rises with participation. As a comparison, we have collected, thanks to the Alsatian platform, 18,917 annotations in 73 days, reaching a 93% accuracy for manual annotation. The annotation campaign we led resulted in a newly POS tagged corpus of 6,878 tokens. This is why efforts on advertising about the platform should be carried on.

5. Conclusion and Perspectives

We have described the steps for preparing the necessary elements for a language to benefit from the POS tags crowdsourcing platform developed in our previous work.
This process resulted in the development of new resources for GC, among which a corpus of 2,439 tokens annotated with POS tags (Millour and Fort, 2018) and the first dedicated POS tagger reaching 84% accuracy. They are both freely available under the CC BY-NC-SA license.[13]
Although the data crowdsourced so far is not satisfactory, due to the low participation, the methodology has been validated for Alsatian, and we intend to follow our efforts on advertising about the crowdsourcing platform. The code of the platform is freely available on GitHub[14] under the CeCILL v2.1 license[15], and is ready to be adapted to any language fulfilling the prerequisites we presented here.

6. Acknowledgements

We wish to thank G. Feler and A. Thibault (Sorbonne Université) for participating in the building of the reference, E. Schang (LLL, Université d'Orléans) for his advice, as well as the participants of Krik for their contribution.

[13]See: https://krik.paris-sorbonne.fr/corpora.
[14]See: https://github.com/alicemillour/Bisame.
[15]See: http://www.cecill.info/.

References

Benzitoun, C., Fort, K., and Sagot, B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe (TCOF-POS : A freely available pos-tagged corpus of spoken french) [in french]. In *Proc. of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 99–112, Grenoble, France, June. ATALA/AFCP.

Bernabé, J. (2001). *La graphie créole*. Guides du CAPES de Créole, Ibis Rouge edition.

Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2001). The prague dependency treebank: Three-level annotation scenario. In *Treebanks: Building and Using Syntactically Annotated Corpora*, pages 103–127. Kluwer Academic Publishers.

Carrión Gonzalez, P. and Cartier, E. (2012). Technological tools for dictionary and corpora building for minority languages: example of the French-based Creoles. In *Proc. of LREC'2012 (Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012))*, pages 47–53, Istanbul, Turkey, May.

Colot, S. and Ludwig, R. (2013). Guadeloupean and Martinican Creole. In Susanne Maria Michaelis, et al., editors, *The survey of pidgin and creole languages. Volume 2: Portuguese-based, Spanish-based, and French-based Languages*. Oxford University Press.

Delumeau, F. (2006). *Une description linguistique du créole guadeloupéen dans la perspective de la génération automatique d'énoncés*. Ph.D. thesis, Université de Nanterre - Paris X.

Denis, P. and Sagot, B. (2012). Coupling an Annotated Corpus and a Lexicon for State-of-the-art POS Tagging. *Language Resources and Evaluation*, 46(4):721–736.

Fort, K. and Sagot, B. (2010). Influence of pre-annotation on POS-tagged corpus development. In *The Fourth ACL Linguistic Annotation Workshop*, pages 56–63, Uppsala, Suède, July.

Guillaume, B., Fort, K., and Lefebvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proc. of International Conference on Computational Linguistics (COLING)*, pages 3041–3052, Osaka, Japan, December.

Hazaël-Massieux, M.-C. (2000). *Ecrire en créole : Oralité et écriture aux Antilles*. L'Harmattan.

Ludwig, R., Montbrand, D., Poullet, H., and Telchid, S. (1990). Abrégé de grammaire du créole guadeloupéen. In *Dictionnaire créole français (Guadeloupe), avec un abrégé de grammaire créole et un lexique français-créole*, pages 17–38. SERVEDIT.

McEnery, T. and Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Millour, A. and Fort, K. (2018). Toward a Lightweight Solution for Less-resourced Languages: Creating a POS Tagger for Alsatian Using Voluntary Crowdsourcing. In *Proc. of Language Resources and Evaluation Conference (LREC'2018)*, Miyazaki, Japan, May.

Observatoire des pratiques linguistiques. (2005). *Les créoles à base française*, volume 5. DGLFLF.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proc. of Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Schang, E., Antoine, J.-Y., and Lefeuvre-Halftermeyer, A. (2017). Les chaînes coréférentielles en créole de la Guadeloupe. In *Proc. of TALN'2017 (DILITAL workshop)*, pages 54–61, Orléans, France, June.

Schang, E. (2013). Extended Projections in a Guadeloupean TAG Grammar. In *Proc. of ESSLLI 2013 (HMGE workshop)*, pages 55–67, Düsseldorf, Germany, June.

6.1. Language Resource References

Millour, Alice and Fort, Karën. (2018). *POS Tagged Corpus of Guadeloupean Creole*.