LREC 2016 Workshop

CCURL 2016

Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity

23 May 2016

PROCEEDINGS

Editors

Claudia Soria, Laurette Pretorius, Thierry Declerck, Joseph Mariani, Kevin Scannell, Eveline Wandl-Vogt

Workshop Programme

Opening Session

09.15 – 09.30 Introduction

09.30 – 10.30 Jon French, *Oxford Global Languages: a Defining Project (Invited Talk)*

10.30 – 11.00 Coffee Break

Session 1

11.00 – 11.25 Antti Arppe, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N. Moshagen, *Basic Language Resource Kits for Endangered Languages: A Case Study of Plains Cree*

11.25 – 11.50 George Dueñas and Diego Gómez, *Building Bilingual Dictionaries for Minority and Endangered Languages with Mediawiki*

11.50 – 12.15 Dorothee Beermann, Tormod Haugland, Lars Hellan, Uwe Quasthoff, Thomas Eckart, and Christoph Kuras, *Quantitative and Qualitative Analysis in the Work with African Languages*

12.15 – 12.40 Nikki Adams and Michael Maxwell, *Somali Spelling Corrector and Morphological Analyzer*

12.40 – 14.00 Lunch Break

Session 2

14.00 – 14.25 Delyth Prys, Mared Roberts, and Gruffudd Prys, *Reprinting Scholarly Works as e-Books for Under-Resourced Languages*

14.25 – 14.50 Cat Kutay, *Supporting Language Teaching Online*

14.50 – 15.15 Maik Gibson, *Assessing Digital Vitality: Analytical and Activist Approaches*

15.15 – 15.40 Martin Benjamin, *Digital Language Diversity: Seeking the Value Proposition*

15.40 – 16.00 Discussion

16.05 – 16.30 Coffee Break

16.30 – 17.30 **Poster Session**

Sebastian Stüker, Gilles Adda, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Elodie Gauthier, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noel Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Markus Müller, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian, *Innovative Technologies for Under-Resourced Language Documentation: The BULB Project*

Dirk Goldhahn, Maciej Sumalvico, and Uwe Quasthoff, *Corpus Collection for Under-Resourced Languages with More than One Million Speakers*

Dewi Bryn Jones and Sarah Cooper, *Building Intelligent Digital Assistants for Speakers of a Lesser-Resourced Language*

Justina Mandravickaite and Michael Oakes, *Multiword Expressions for Capturing Stylistic Variation Between Genders in the Lithuanian Parliament*

Richard Littauer and Hugh Paterson III, *Open Source Code Serving Endangered Languages*

Uwe Quasthoff, Dirk Goldhahn, and Sonja Bosch, *Morphology Learning for Zulu*

17.30 – 18.00 Discussion and Conclusions

Workshop Organizers

Thierry Declerck	DFKI GmbH, Language Technology Lab, Germany
Joseph Mariani	LIMSI-CNRS & IMMI, France
Laurette Pretorius	University of South Africa, South Africa
Kevin Scannell	St. Louis University, USA
Claudia Soria	CNR-ILC, Italy
Eveline Wandl-Vogt	Austrian Academy of Sciences, ACDH, Austria

Workshop Programme Committee

Gilles Adda	LIMSI-CNRS & IMMI, France
Tunde Adegbola	African Languages Technology Initiative, Nigeria
Eduardo Avila	Rising Voices, Bolivia
Martin Benjamin	The Kamusi Project, Switzerland
Delphine Bernhard	LiLPa, Université de Strasbourg, LiLPA, France
Paul Bilbao Sarria	Euskararen Gizarte Erakundeen KONTSEILUA, Spain
Vicent Climent Ferrando	NPLD, Belgium
Daniel Cunliffe	Prifysgol De Cymru / University of South Wales, School of Computing and Mathematics, UK
Nicole Dolowy-Rybinska	Polska Akademia Nauk / Polish Academy of Sciences, Poland
Mikel Forcada	Universitat d'Alacant, Spain
Maik Gibson	SIL International, UK
Tjerd de Graaf	De Fryske Akademy, The Netherlands
Thibault Grouas	Délégation Générale à la langue française et aux langues de France, France
Auður Hauksdóttir	Vigdís Finnbogadóttir Institute of Foreign Languages, Iceland
Peter Juel Henrichsen	Copenhagen Business School, Denmark
Davyth Hicks	ELEN, France
Kristiina Jokinen	Helsingin Yliopisto / University of Helsinki, Finland
John Judge	ADAPT Centre, Dublin City University, Ireland
Steven Krauwer	CLARIN, The Netherlands
Steven Moran	Universität Zürich, Switzerland
Silvia Pareti	Google Inc., Switzerland
Daniel Pimienta	MAAYA
Steve Renals	University of Edinburgh, UK
Kepa Sarasola Gabiola	Euskal Herriko Unibertsitatea / University of the Basque Country, Spain
Felix Sasaki	DFKI GmbH and W3C fellow, Germany
Virach Sornlertlamvanich	Sirindhorn International Institute of Technology / Thammasat University, Thailand
Ferran Suay	Universitat de València, Spain
Jörg Tiedemann	Uppsala Universitet, Sweden
Francis M. Tyers	Norges Arktiske Universitet, Norway

Preface

The LREC 2016 Workshop on "Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity" (CCURL 2016) explores the relationship between language and the Internet, and specifically the web of documents and the web of data, as well as the emerging Internet of things, is a growing area of research, development, innovation and policy interest. The emerging picture is one where language profoundly affects a person's experience of the Internet by determining the amount of accessible information and the range of services that can be available, e.g. by shaping the results of a search engine, and the amount of everyday tasks that can be carried out virtually. The extent to which a language can be used over the Internet or in the Web not only affects a person's experience and choice of opportunities; it also affects the language itself.

If a language is poorly or not sufficiently supported to be used over digital devices, for instance if the keyboard of the device is not equipped with the characters and diacritics necessary to write in the language, or if there is no spell checker for a language, then its usability becomes severely affected, and it might never be used online. The language could become "digitally endangered", and its value and profile could be lessened, especially in the eyes of new generations. On the other hand, concerted efforts to develop a language technologically could contribute to the digital ascent and digital vitality of a language, and therefore to digital language diversity. These considerations call for a closer examination of a number of related issues.

First, the issue of "digital language diversity": the Internet appears to be far from linguistically diverse. With a handful of languages dominating the Web, there is a linguistic divide that parallels and reinforces the digital divide. The amount of information and services that are available in digitally less widely used languages are reduced, thus creating inequality in the digital opportunities and linguistic rights of citizens. This may ultimately lead to unequal digital dignity, i.e. uneven perception of a language importance as a function of its presence on digital media, and unequal opportunities for digital language survival.

Second, it is important to reflect on the conditions that make it possible for a language to be used over digital devices, and about what can be done in order to grant this possibility to languages other than the so-called "major" ones. Despite its increasing penetration in daily applications, language technology is still under development for these major languages, and with the current pace of technological development, there is a serious risk that some languages will be left wanting in terms of advanced technological solutions such as smart personal assistants, adaptive interfaces, or speech-to-speech translations. We refer to such languages as under-resourced. The notion of digital language diversity may therefore be interpreted as a digital universe that allows the comprehensive use of as many languages as possible.

All the papers accepted for the Workshop address at least one of these issues, thereby making a noteworthy contribution to the relevant scholarly literature and to the technological development of a wide variety of under-resourced languages. Each of the fifteen accepted papers was reviewed by at least three members of the Programme Committee, eight of which are presented as oral presentations and six as posters. We look forward to collaboratively and computationally building on this growing tradition of CCURL in the future for the continued benefit of all the under-resourced languages of the world!

C. Soria, L. Pretorius, T. Declerck, J. Mariani, K. Scannell, E. Wandl-Vogt May 2016

Table of Contents

Basic Language Resource Kits for Endangered Languages: A Case Study of Plains Cree

Antti Arppe, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N. Moshagen
University of Alberta & UIT Arctic University of Norway
Email: arppe@ualberta.ca, lachler@ualberta.ca, trond.trosterud@uit.no, lene.antonsen@uit.no, sjur.n.moshagen@uit.no

Abstract

Using Plains Cree as an example case, we describe and motivate the adaptation of the BLARK approach for endangered, less-resourced languages (resulting in an EL-BLARK), based on (1) what linguistic resources are most likely to be readily available, (2) which end-user applications would be of most practical benefit to these language communities, and (3) which computational linguistic technologies would provide the most reliable benefit with respect to the development efforts required.

Keywords: computational modeling, morphology, syntax, finite-state machines, (intelligent) electronic dictionaries, spell-checkers, grammar-checkers, (intelligent) computer-aided language learning, speech synthesis, optical character recognition, Plains Cree

1. Introduction to a BLARK

Our objective is to adapt the *Basic LAnguage Resource KIT* (BLARK) approach to the needs of under-resourced endangered language communities. As an example case, we will use Plains Cree (Algonquian, crk), an Indigenous language of central Canada. The approach advocated here stems from our collaboration with Miyo Wahkohtowin Education (Maskwacîs, Alberta, Canada) in the development of various technological resources for Plains Cree over the past several years, as well as two decades of fieldwork and language revitalization efforts with Indigenous communities across North America.

The BLARK is an approach proposed by Krauwer (2003) and Binnenpoorte et al. (2002) for establishing a roadmap for Human Language Technologies (HLT) for a given language. A BLARK aims to identify:

(1) What is minimally required to guarantee an adequate digital language infrastructure for that language?
(2) What is the current situation of HLT in that language?
(3) What needs to be done to guarantee that at least what is required be available?
(4) How can goal (3) be best achieved?
(5) How can we guarantee that once an adequate HLT infrastructure is available, it also remains so?

In defining a BLARK for a given language, Binnenpoorte et al. propose a three-way distinction between:

(1) Applications: end-user software applications that make use of HLT;
(2) Modules: the basic software components that are essential for developing HLT applications; and
(3) Data: data sets and electronic descriptions that are used to build, improve, or evaluate modules.

Moreover, the relationships between these three classes of resources can be presented as a matrix on:

(1) Which modules are required for which applications;
(2) Which data are required for which modules; and
(3) What the relative importance is of the modules and data.

2. A Core BLARK for an Endangered Language – EL-BLARK

For majority languages to which the BLARK approach has been primarily applied so far, there typically exist substantial written corpora of hundreds of millions of words, annotated spoken corpora, multiple comprehensive descriptions of the lexicon, morphology and syntax, thesauri, and other similar resources. Indigenous and endangered languages, on the other hand, are typically substantially less-resourced, with often only basic lexical and grammatical descriptions having been published, and little to no textual or spoken corpora available. Moreover, this rather dire situation represents the norm for most of the 7,000+ languages in the world today.

Therefore, in defining a core BLARK for these endangered languages – an EL-BLARK – the following two questions are of prime importance (Arppe et al. 2015):

(1) What types of relevant data resources are likely to be available?
(2) What HLT applications may be of most practical value in supporting the continued use and revitalization of these languages within their communities?

Together, (1) and (2) will determine the possible and necessary technological module components, the relationships of which are presented as a BLARK matrix in Table 1.

Modules	**Data**					**Applications**				
Language technology	Morphological description	Syntactic description	Bilingual dictionary	Written text collection (printed)	*Spoken recordings: (digitized)	Simple electronic dictionary	Intelligent electroni dictionary	*Spell-checker	**Grammar-checker	*Intelligent computer-aided language learning
Transcriptors				+		++	++	++	+	+
Morphological model: analyser/ paradigm generator	++		+				++	++	+	++
*Weighted morphological model	++		+	++			+	++		
Bilingual lexical database	+	+	++	++		++	++	++		++
*Written text corpus: electronic	+			++	++	++	++	+		
*Spoken text corpus: annotated					++	++	++			
*Speech synthesis					++					++
**Optical character recognition				++						
**Syntactic model	++	++		++	++				++	+

Table 1. *Overview of the importance of data for modules and the importance of modules for applications in EL-BLARK. Data/modules/applications for Plains Cree (**) not yet implemented, or (*) in very early stages of collection or development.*

As many Indigenous languages are severely endangered, with the number of fluent speakers often reduced to a mere handful, best practices in digital language planning lead us to focus on linguistic knowledge and resources which are already being collected by field and community linguists, as well as computational tools and applications that can be developed within a reasonable timeframe, with reasonable effort, and with tried and tested technologies. In particular, we argue that it is dubious to waste the already scarce time of fluent speakers to explore purely theoretically motivated research questions which offer little or no practical contribution to the continued sustainability of these languages.

Furthermore, in determining which applications will be of the most practical value, a wide range of community-specific variables need to be taken into account, including, but not limited to:

(1) How many (fluent/learner) speakers are there? While some Indigenous languages have many tens of thousands of speakers, most have many fewer, and far too many have already reached the stage where no first-language speakers remain.
(2) Where do they live? In some cases, community members still all reside in their same traditional territory. More often nowadays, though, many speakers and learners are found far from their traditional territory, having migrated to urban centres in search of better economic opportunities.
(3) To what extent does a standard and agreed-upon

written form of the language exist? While some Indigenous communities have been early adopters of literacy, most have only recently begun to develop and use a written version of their language, and orthographic standards are still very much in flux.

(4) What are the current domains of use for the language? A key indicator of language endangerment is both the loss of use in existing domains (social life, traditional religion, etc.), as well as a failure to expand into new domains (school, government, politics). While some communities aspire to use their language in all domains of modern life, others have more modest goals.

(5) What educational programs are offered in the language, and at what levels? Until only very recently, endangered languages were almost entirely excluded from mainstream educational programs, and where they were offered it was typically only for very young children. This has begun to change, and the demand for qualified teachers and government-approved curriculum for these languages is on the rise in many places.

(6) What infrastructure exists within the local communities for supporting language work? While the early years of linguistics often saw individual speakers working with individual field linguists, today many endangered language communities have language departments, language authorities and other bodies which support and regulate the work done in documenting and revitalizing their languages.

The above considerations are based on our collective experiences at the University of Alberta and that of our collaborators on field linguistic research conducted over the past several decades on North American Indigenous languages among the Algonquian (Plains Cree, East Cree, Innu, Ojibwe), Dene (Tsuut'ina, Dene Sųłiné), Siouan (Nakota), Iroquoian (Cherokee) and Keresan families, as well as isolates such as Haida. Together, these factors can be used to determine where HLT developers, field linguists and Indigenous language communities can create the most added value in language technology development, as well as to prioritize their documentary and analytical efforts in order to make various applications possible.

3. Reasonably Expectable Data Resources

In the case of endangered languages, language documentation work typically focuses on developing the following four sets of resources:

(1) descriptions of morphology and syntax, from basic sketches to comprehensive detailed grammars with explicit descriptions of inflectional paradigms and syntactic constructions;

(2) bilingual lexical descriptions with translations to a majority language, ranging from basic word lists to full-scale comprehensive lexical databases (including information on paradigm class and semantic restrictions);

(3) narrative text collections in either printed or electronic format (with or without accompanying spoken recordings); and

(4) recordings of spoken language, ranging from carefully pronounced individual words and sentences, multi-participant native speaker discourse, and narratives of various types, which may be annotated.

Other types of resources commonly found for majority languages, such as monolingual dictionaries, thesauri, and multimedia corpora, are almost entirely lacking for most endangered languages.

4. Relevant Language Technology Frameworks and Software Applications for Documentation

There are several types of software applications which have the greatest relevance for under-documented and endangered languages, and for which the underlying technological basis has become mature enough to produce results of genuine practical assistance. During the initial documentation phase, lexical database tools (e.g. Fieldwork Language Explorer [FLEX]), audio recording, editing and annotation tools (e.g. *Acrobat Audition*, *ELAN*, *Audacity*), and text corpus platforms (e.g. *Korp*, Borin et al. 2012, based on *Corpus Workbench*, Evert & Hardie 2011) are of primary usefulness. These are the basis on which actual language technology modules can be developed using a variety of frameworks, such as:

(1) Computational modelling formalisms for morphology and syntax (e.g. Finite-State Machines, e.g. Beesley & Karttunen 2003, Constraint Grammar, Karlsson et al. 1995). These are important because many of these languages are morphologically complex, and the patterns cannot be statistically learned through recourse to large-scale corpora. From experience we know that a general desideratum in selecting such computational formalisms is that they should well-known computational data structures, fast and efficient, as well as easily portable to different operating systems and platforms, and thus integratable as modules such as spell-checkers within other applications – all of which apply for e.g. finite-state machines. Moreover, weighted Finite-State Machines (Mohri 1997) are a recent, promising development from the perspective of endangered languages, as they allow for the modeling of the likelihood of the various

possible morpheme sequences even with relatively small amounts of usage data (Pirinen 2014).

(2) Optical character recognition programs (e.g. *tesseract*, Smith 2007). These are important because for many endangered languages there does exist a notable body of written texts – traditional narratives, Bible translations, writings by community members – dating from a time when the language was in broader use. In many cases, though, these resources exist only in printed form, and often in a pre-standardized orthography (cf. Hubert et al. 2016).

(3) Speech-synthesis development infrastructures (e.g. Simple4All, cf. http://simple4all.org/, Watts et al. 2013). These are important because most of these languages have very shallow literary traditions, and both speakers and learners may be more comfortable interacting with language technology through speech rather than text.

(4) Transcriptors (using e.g. regular expressions and rewrite rules within Finite-State Transducers) which allow automatic conversions between competing *de facto* orthographic standards across communities.

Besides the above, (5) speech recognition, which would allow for (semi-)automatic transcription of the considerable amounts of recordings of Indigenous language interviews, narratives, elicitations, discussions and radio broadcasts spanning over many decades during the 20th century until the present day, would be of great value, as much of the available linguistic data for these languages is in the spoken form. Nevertheless, to the best of our knowledge the quality of current technological solutions for speech recognition that would be speaker independent and trainable with relatively small amounts of data is not yet at a level that would justify their use.

5. Practically Useful Applications

Based on our experience and conversations with community language activists, including both speakers and learners, there are several particular applications that are of primary and immediate value to endangered language communities. Moreover, in the modern, computerized world, technology plays an increasingly central role in *how* and *where* most people, in particular the younger generations – whether indigenous or not – use language, communicating constantly with laptops, smartphones or tablets.

For many communities, dictionaries are viewed as the most valuable language resource, because they serve as the repository for speakers' knowledge about the words of their language. As such, intelligent, web-accessible dictionaries (I-DICT), which pair lexical databases and computational morphological analysers and generators are given high priority in the development of language technology (Johnson et al. 2013). These dictionaries can recognize inflected forms and generate inflectional paradigms, allowing learners to access the structure of the language in ways that are not feasible with print dictionaries. They can be further enriched with recordings of individual words and sentences, and usage examples from text collections. Such dictionaries would typically be bilingual with the local majority language.

Second, computer-aided language learning (CALL) applications have an important role to play in extending opportunities for learners to improve their proficiency in the language. These can range from basic vocabulary practice to exercises with morphological alternations in context (intelligent CALL, or ICALL). Prompts and practice can include both the written and spoken forms of the language, especially with the aid of text-to-speech models. These low-cost applications are especially useful for Indigenous communities with a large diasporic population, many of whom may have little opportunity for in-person language learning with a fluent speaker/teacher, and little money to spend on expensive textbooks or CDs.

Third, writers' tools such as spell-checkers and grammar-checkers facilitate the use of the written form of the language, and support its spread into new domains of use. These are especially helpful in contexts where the orthographic traditions are new and still evolving, and where the majority of language users are second language learners, often having primary exposure to the spoken form of their heritage language.

Other types of language technology commonly found for majority languages may not be realistic or appropriate in endangered language communities. In particular, the availability of translation tools (beyond the level of an intelligent dictionary) may actually undercut learners' motivations to reach proficiency in the language.

6. Data resources for Plains Cree

Plains Cree is one of the best-resourced and most widely-spoken Indigenous languages in North America, with speakers found in communities across a vast stretch of central Canada. While most speakers live on reserves in rural areas, significant numbers of speakers are found in urban centres such as Edmonton, Saskatoon and Regina. Estimates of the number of speakers vary considerably, but 15,000-20,000 is a common range, with the majority of speakers being over the age of 30. Its use as a written language stretches back over a century, and it is taught as a subject in various elementary and secondary schools, as well as a handful of post-secondary institutions. Although certainly endangered, Plains Cree is in a much healthier state than most other Indigenous languages in Canada, and it is viewed as one of only a handful of such languages that is likely to survive into the next century (Cook and Flynn

2008).

The inventory of relevant data resources includes:

(1) At least four modern descriptions of the morphology and elements of the syntax (Wolfart 1973; Ahenakew 1987; Okimâsis 2004; Wolvengrey 2011);
(2) Two bilingual electronic dictionaries: the *Alberta Elders' Cree Dictionary* with 10,388 Cree-to-English and 22,970 English-to-Cree lexical entries (LeClaire and Cardinal 1998) and the *Maskwacîs Cree Dictionary* with 8985 lexical entries (Miyo Wahkohtowin Education), and one electronic lexical database, *nêhiyawêwin : itwêwina / Cree : Words* with 16,452+ lexical entries (Wolvengrey 2001);
(3) A variety of printed corpus materials, some of which are available in electronic form, including Bible translations and collections of traditional narratives: in particular (a) the Bloomfield texts (1930, 1934)[1]; (b) the *Cree Prayer Book* (Demers et al. 2010); and (c) the works of Freda Ahenakew and H. C. Wolfart: Ahenakew (2000); Bear et al. (1992); Kā-Nīpitēhtēw (1998); Masuskapoe (2010); Minde (1997); Vandall & Douquette (1987); and Whitecalf (1993); and
(4) A variety of spoken Cree collections, including some audiobooks, recorded narratives, and pedagogical materials.

While this is a robust collection in comparison to most other Indigenous languages, these resources are hindered by their small size (adding up to less than 1 million words), limited coverage, and lack of full standardization when compared to resources available for majority languages.

7. Language technology for Plains Cree

Based on the existing resources outlined in Section 6, we are currently developing a range of language technology for Plains Cree in partnership with Miyo Wahkohtowin Education, based in Maskwacîs, Alberta, First Nations University of Canada (Regina, Saskatoon)[2], the Faculty of Native Studies[3], the developers and editors of the current simple *Online Cree Dictionary / nehiyaw masinahikan* (http://www.creedictionary.com/) incorporating the three above-mentioned primary dictionaries [4], and the Cree Literacy Network (creeliteracy.org) [5]. These language technological modules include the following:

Foremost among these is a computational morphological model, based on FST formalism (Snoek et al. 2014; Harrigan et al. 2016). This model is informed by the grammatical descriptions of Wolfart (1973) and Wolvengrey (2001), and coupled with lexical data from Wolvengrey (2011). Currently it succeeds in recognizing and parsing 72% of the word form types in a small corpus consisting of the works of Freda Ahenakew and H. C. Wolfart (Ahenakew, 2000; Bear et al., 1992; Kā-Nīpitēhtēw, 1998; Masuskapoe, 2010; Minde, 1997; Vandall & Douquette, 1987; and Whitecalf, 1993).

Second is a speech-synthesiser based on the Simple4All framework. Our initial model is based on 10 minutes of audiobook data. This model is being augmented with recordings from dictionary sessions with speakers in Maskwacîs.

Third, we have transcriptors fully implemented for the three orthographic standards used in Plains Cree communities (vowel length marked with macrons, vowel length marked with circumflexes, vowel length unmarked), as well as conversion between Latin and Cree syllabic characters, all based on the FST model.

While we have found OCR indispensable in creating electronic text corpora based on historical printed materials for other Indigenous languages, most notably Northern Haida, this has been less a focus of our work on Plains Cree as we have been successful in gaining access to the electronic source files of many of the main text collections.

8. HLT applications for Plains Cree

In collaboration with our community partners, and building off of the language technologies described above, our team is currently developing a range of HLT applications for Plains Cree, with the goal of facilitating the use and acquisition of the language throughout the community.

At the forefront is an intelligent web-accessible bilingual dictionary, *itwêwina* (URL: http://itwewina.oahpa.no). This dictionary integrates elements from the lexical database underlying Wolvengrey (2001) and the Plains Cree FST. It allows users to search from Plains Cree to English, or the reverse. It accepts fully inflected forms of Plains Cree nouns and verbs, and returns a parse of that form, along with a link to a dynamically-generated full paradigm for that lemma. Current work on the dictionary includes augmenting the entries with audio files and example sentences, as well as linking the dictionary into the various written and spoken corpus materials we are collecting.

Second is an ICALL application, *nêhiyawêtân* (literally "Let's speak Cree", URL: http://oahpa.no/nehiyawetan/),

[1] Electronic version courtesy of Dr. Kevin Russell.

[2] Professor Arok Wolvengrey and Dr. Jean Okimasis.

[3] Cree instructor, M.Sc. Dorothy Thunder.

[4] Professor Earle Waugh, University of Alberta, and Managing Director Ahmad Jawad, Intellimedia.

[5] Director, M.A. Arden Ogg.

which integrates the Plains Cree FST and is based on pedagogical materials used by first-year students of Plains Cree at the University of Alberta. This is based on the *Oahpa* ICALL application developed for Sámi languages (Antonsen et. al. 2009). Learners are drilled on a wide range vocabulary keyed to the chapters in the textbook, as well as the production of targeted inflectional forms both in and out of conversational contexts. A demo version has been evaluated with five users as part of a pilot study (Bontogon 2016), and further improvements are currently being made, including the addition of audio files to the vocabulary drills and an expansion of the coverage of the application to include more advanced vocabulary and grammatical structures. Moreover, we intend to test the use of speech synthesis to provide oral prompts for the Cree sentence contexts provided in the morphological exercises, which are generated by combining exercise frames with the computational model, the number of which could not be prerecorded in practice, in particular when the vocabulary content and exercise types application will be extended.[6]

Third is a spell-checker integrated into an OS (Mac OS X 10.10 onwards), again based on the Plains Cree FST and the lexical database underlying *itwêwina*. The spell-checker handles both the standard roman orthography as well as syllabics. At present, there exists only an internal demo version, but plans are in place for eventual integration into LibreOffice and MS Office. This will then be piloted with university Plains Cree students, as well as our community partners.

A key element to make note of in the core applications that we have identified for EL-BLARK is that the computational morphological model is an integral component (I-DICT, ICALL, spell-checking, grammar-checking). Therefore, we are basing our own development work in the *giella* infrastructure, developed over the last decade firstly for creating similar modules and applications for the Indigenous Sámi languages of Northern Scandinavia by the Giellatekno and Divvun research teams at UIT Arctic University of Tromsø. (Trosterud, 2004, 2006). What is attractive in the *giella* infrastructure is that the development of the computational morphological model is seamlessly linked with the various applications, which have a quality ready for serious, extensive use by end-users in endangered communities practically out-of-the-box.

9. Concluding Notes

From the experiences of minority language speakers around the world (Rueter and Trosterud, 2012; First Languages Australia 2015, among many others), it is clear that digital language resources, and in particular Human Language Technology, are seen by both academics and community members as having an important role in supporting minority languages, both now and in the future. These resources are a benefit both for current native speakers who wish to continue using their language, as well as the generations of learners who are striving to recapture those languages and find a place for them in their modern lives.

We view the creation of EL-BLARKs as an essential component of the digital language planning strategies for endangered language communities. The EL-BLARK is a tool that can allow community language activists, policymakers, field linguists and other stakeholders to better understand the interconnectedness of various language activities – documentation, graphization, morphological and syntactic analysis, bilingual lexicography, development of pedagogical resources, etc. – and how these relate to the potential development of HLT for their languages, thus contributing to the maintenance if not expansion of digital diversity.

It is important to note, however, that what the EL-BLARK matrix fails to capture is that the development, deployment and assessment of these HLT applications require close collaboration between computational linguists, field linguists, native speakers, language teachers, second language learners and local community leaders, in ways which respect the intellectual property of the endangered language community and which prioritize projects which will have a tangible benefit to local language revitalization efforts.

Acknowledgements

Building a computational model of Plains Cree morphology and the subsequent applications is a task that relies on the knowledge, time and goodwill of many people. We thank the University of Alberta's Killam Research Fund Cornerstones Grant for providing initial support for this project, and the Social Sciences and Humanities Research Council (SSHRC Partnership Development Grant 890-2013-0047) and the University of Alberta's Kule Institute for Advanced Study (KIAS) Research Cluster grant for making the continuation of this project possible. We would like to acknowledge in particular the crucial advice, attention and effort of Jean Okimâsis and Arok Wolvengrey, and thank them for the resources they have contributed. We wish also to thank Jeff Muehlbauer for his time and materials. Further, it is important to acknowledge the helpfulness of Earle Waugh who at the very start of our project made his dictionary available to us, and who has been very supportive. Arden Ogg has worked tirelessly to build connections among researchers working on Cree, which has greatly promoted and facilitated our work. Ahmad Jawad and Intellimedia, Inc. who have for some time

[6] In the case of full-fledged ICALL application created for North Sámi by Giellatekno, the overall number of all possible sentential prompts/contexts is several hundred thousand (Antonsen et al. 2013).

provided the technological platform to make available a number of Plains Cree dictionaries through a web-based interface, have given us invaluable assistance in terms of resources and introductions. We would also especially like to thank the staff at Miyo Wahkohtowin Education for their wonderful enthusiasm, and for welcoming us into their community. Last but by no means least, we are indebted to innumerable Elders and native speakers of Plains Cree whose contributions have made possible all the dictionaries and text collections we are fortunate to have today.

References

Ahenakew, Freda. 1987. *Cree Language Structures: A Cree Approach.* Winnipeg: Pemmican Publications, Inc.

Antonsen, Lene. 2013. Constraints in free-input question-answering drills. Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013. *NEALT Proceedings Series 17 / Linköping Electronic Conference Proceedings* 86.11–26.

Antonsen, L., Huhmarniemi, S. and T. Trosterud 2009: Interactive pedagogical programs based on constraint grammar. Proceedings of the 17th Nordic Conference of Computational Linguistics. Nealt Proceedings Series 4.

Antonsen, Lene; Ryan Johnson; Trond Trosterud; and Heli Uibo. 2013. Generating modular grammar exercises with finite-state transducers. Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013. *NEALT Proceedings Series 17 / Linköping Electronic Conference Proceedings* 86.27–38.

Arppe, Antti, Lene Antonsen, Trond Trosterud, Sjur Moshagen, Dorothy Thunder, Conor Snoek, Timothy Mills, Juhani Järvikivi, and Jordan Lachler. 2015 Turning language documentation into reader's and writer's software tools. *4th International Conference on Language Documentation and Conservation (ICDLC)*. 26 February – 1 March 2015, Honolulu, HI

Beesley, Kenneth R. & Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford (CA).

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

Bontogon, Megan. 2016. *Evaluating nêhiyawêtân: A computer assisted language learning (CALL) application for Plains Cree.* Honours thesis, Department of Linguistics, University of Alberta.

Borin, Lars, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. Proceedings of LREC 2012. Istanbul: ELRA, 474-478.

Eung-Do Cook and Darin Flynn. 2008. Aboriginal languages of Canada. In: O'Grady, William and John Archibald (eds.) Contemporary Linguistic Analysis. Pearson, Toronto (ON).

Evert, Stefan and Hardie, Andrew (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.

First Languages Australia. 2015. Angkety map / Digital resource report. Available at: http://languageresources.com.au/files/fla-angkety-map.pdf. Accessed February 17, 2016.

Harrigan, Atticus, Lene Antonsen, Antti Arppe, Dustin Bowers, Katie Schirler, Trond Trosterud and Arok Wolvengrey. 2016. Learning from the computational modeling of Plains Cree verbs. Workshop on *Computational methods for descriptive and theoretical morphology*, 17th International Morphology Meeting. Vienna, February 18–21, 2016.

Hubert, Isabell, Antti Arppe and Jordan Lachler. 2016. Training & Quality Assessment of an Optical Character Recognition Model for Northern Haida. LREC 2016, Portorož, Slovenia, May 2016.

Johnson, Ryan, Lene Antonsen, Trond Trosterud. Using Finite State Transducers for Making Efficient Reading Comprehension Dictionaries. 2013. In: Stephan Oepen, Kristin Hagen, Janne Bondi Johannessen (eds.). *Proceedings of the 19th Nordic Conference of Computational Linguistics* (NODALIDA 2013), May 22–24, 2013, Oslo University, Norway. NEALT Proceedings Series 16, 59-71. url: http://www.ep.liu.se/ecp/085/010/ecp1385010.pdf

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Anttila, editors. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Natural Language Processing, No 4. Mouton de Gruyter, Berlin and New York.

Krauwer, Steven. 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. *Proceedings of the International Workshop "Speech and Computer"*, SPECOM 2003, Moscow.

LeClaire, Nancy & George Cardinal. 1998. *Alberta Elders' Cree Dictionary*. University of Alberta Press.

Jean Okimâsis. 2004. Cree: Language of the Plains, Volume 13 of University of Regina publications. University of Regina Press, Regina (SK).

Mohri, M. 1997. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23:269–311.

Pirinen, Tommi. 2014. Weighted Finite-State Methods for Spell-Checking and Correction. Ph.D dissertation. Department of Modern Languages, University of Helsinki.

Rueter, Jack and Trond Trosterud. 2012: How to help languages to survive during modern time by means of language technologies? . Available at http://giellatekno.uit.no/background/sykt.pdf (Accessed February 17, 2016)

Smith, R. 2007. An Overview of the tesseract OCR Engine. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2:629–633.

Snoek, Conor, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, & Trond Trosterud. 2014. Modeling the Noun Morphology of Plains Cree. *ComputEL: Workshop on the use of computational methods in the study of endangered languages*, 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, 26 June 2014.

Trosterud, T. 2004. Porting morphological analysis and disambiguation to new languages. In Carson-Berndsen, J. (Ed.), *First Steps in Language Documentation for Minority Languages: Proceedings of the SALTMIL Workshop at LREC 2004*, pp. 90-92.

Trosterud, T. 2006. Grammatically based language technology for minority languages. In Saxena., A., & Brin, L. (Eds.), *Lesser Known Languages of South Asia*, The Hague: Mouton de Gruyter, pp. 293-316.

Watts, O., A. Stan, R. Clark, Y. Mamiya, M. Giurgiu, J. Yamagishi, J., and S. King. 2013. Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis. In *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain. 121-126.

H. Christoph Wolfart. 1973. Plains Cree: A grammatical study. *Transactions of the American Philosophical Society*, No. 5.

Wolvengrey, Arok. 2011. *Semantic and Pragmatic Functions in Plains Cree Syntax*. Utrecht: LOT.

Wolvengrey, Arok. 2001. *Nēhiyawēwin : itwēwina* (*Cree: Words)*. University of Regina Press.

Language Resource References

itwêwina. 2016. *Intelligent Plains Cree – English electronic dictionary*. URL : http://itwewina.oahpa.no. Alberta Language Technology Lab, University of Alberta, First Nations University of Canada, and Giellatekno/Divvun, UIT Arctic University of Tromø.

nêhiyawêtân. 2016. *Intelligent Plains Cree – English electronic dictionary*. URL : http://oahpa.no/nehiyawetan. Alberta Language Technology Lab, University of Alberta, and Giellatekno, UIT Arctic University of Tromsø.

Original Text References

Ahenakew, Alice. 2000. *āh-āyītaw isi ē-kī-kiskēyihtahk maskihkiy / They Knew Both Sides of Medicine*. Told by Alice Ahenakew ; edited, translated, and with a glossary by H. C. Wolfart and Freda Ahenakew. Publications of the Algonquian Text Society / Collection de la Société d'édition des textes algonquiennes. Winnipeg : University of Manitoba Press.

Glecia Bear, Minnie Fraser, Irene Calliou, Mary Wells, Alpha Lafond, and Rosa Longneck. 1992. *kōhkominawak otācimowiniwāwa / Our Grandmothers' Lives As Told in Their Own Words.* Told by Glecia Bear, Minnie Fraser, Irene Calliou, Mary Wells, Alpha Lafond, and Rosa Longneck ; edited by Freda Ahenakew and H. C. Wolfart. Saskatoon : Fifth House Publishers.

Bloomfield, Leonard (compiler). 1930. Sacred stories of the Sweet Grass Cree. Ottawa: F.A. Acland.

Bloomfield, Leonard. 1934. *Plains Cree texts.* American Ethnological Society Publications 16. New York.

Patricia Demers, Naomi L. McIlwraith, Dorothy Thunder, Arok Wolvengrey and Patricia Demers. 2010. The Beginning of Print Culture in Athabasca Country. A Facsimile Edition & Translation of a Prayer Book in Cree Syllabics by Father Émile Grouard, OMI, Prepared and Printed at Lac La Biche in 1883 with an Introduction by Patricia Demers.

Kā-Nīpitēhtēw, Jim. 1998. *ana kā-pimwēwēhahk okakēskihkēmowina / The Counselling Speeches of Jim Kā-Nīpitēhtēw*. Told by Jim Kā-Nīpitēhtēw ; edited, translated, and with a glossary by Freda Ahenakew and H. C. Wolfart. Publications of the Algonquian Text Society / Collection de la Société d'édition des textes algonquiennes. Winnipeg : University of Manitoba Press.

Masuskapoe, Cecilia. 2010. *piko kīkway ē-nakacihtāt: kēkēk otācimowina ē-nēhiyawastēki mitoni ē-āh-itwēt māna Cecila Masuskapoe / There's Nothing She Can't Do : Kēkēk's Autobiography published in Cree*. Exactly as told by Cecilia Masuskapoe ; in a critical edition by H. C. Wolfart and Freda Ahenakew. Algonquian and Iroquoian Linguistics, Memoir 10.

Minde, Emma. 1997. *kwayask ē-kī-pē-kiskinowāpahtihicik / Their Example Showed Me They Way*. Told by Emma Minde ; edited, translated and with a glossary by Freda Ahenakew and H. C. Wolfart.

Vandall, Peter and Joe Douquette. 1987. *wāskahikaniwiyiniw-ācimowina / Stories of the House People*. Told by Peter Vandall and Joe Douquette ; edited, translated, and with a glossary by Freda Ahenakew. Publications of the Algonquian Text Society / Collection de la Société d'édition des textes algonquiennes. Winnipeg : University of Manitoba Press.

Whitecalf, Sarah. 1993. *kinēhiyāwiwininaw nēhiyawēwin / The Cree Language is Our Identity* : The Laronge Lectures by Sarah Whitecalf. Told by Sarah Whitecalf ; edited, translated, and with a glossary by H. C. Wolfart and Freda Ahenakew. Publications of the Algonquian Text Society / Collection de la Société d'édition des textes algonquiennes. Winnipeg : University of Manitoba Press.

Building Bilingual Dictionaries for Minority and Endangered Languages with Mediawiki

George Dueñas, Diego Gómez
Instituto Caro y Cuervo
Calle 10 # 4-69, Bogotá, Colombia
george.duenas@caroycuervo.gov, diego.gomez@caroycuervo.gov.co

Abstract

In this paper, we present the ongoing work of the indigenous languages team at Caro and Cuervo Institute in developing language resources and technologies to document and revitalize minority languages which are in some degree of endangerment. This work consists in creating not only electronic dictionaries, but also a space where linguistic and cultural information is stored about the language as personal names and toponyms, among others. In order to do this, we have created different templates for placing the information. The main software that we have used to create these dictionaries is MediaWiki and the Semantic Mediawiki extension (free software open source). The Mediawiki software, adapted to lexicography needs, has become an important tool in this project. When the information has been stored in variables, we can display it through queries in the Semantic Mediawiki syntax. All of these tools have enabled us to show and recover information from each lexicographic entry. Two bilingual (uni- and bidirectional) dictionaries of Saliba and Carijona indigenous languages of Colombia were built with these tools. These dictionaries will increase the amount of content available for such languages in Internet and reduce a little the lack of places for using their languages over digital media.

Keywords: Semantic Mediawiki, Electronic Lexicography, Dictionary, Minority Languages

1. Introduction and Background

1.1. Basic Issues about Dictionaries

We assume the readers are somewhat familiar with three aspects about dictionaries: first, from its definition, a dictionary is "a description of the vocabulary used by members of a speech community" (Atkins and Rundell, 2008, p. 2). Furthermore, it is also a methodical, orderly and concise description. Second, the types of dictionaries are monolingual, bilingual (unidirectional or bidirectional), and multilingual. And the third concerns with the means of disclosure. These can be printed, electronic, or web-based (Atkins and Rundell, 2008, p. 24). So, the dictionary has been used to describe the vocabulary of many languages of the world both extinct and living, as well as to describe the inventory of words from different areas of knowledge.

In The Americas, the creation of bilingual dictionaries began with the arrival of the Spaniards and the Catholic missionaries that undertook the evangelization of the Indians, creating some kind of grammars called "artes" and bilingual vocabularies between Spanish and american languages with which they had contact. These texts were copied by hand or printed in the best case to be distributed among missionaries who needed them to learn the languages in order to carry out the Catholic evangelization; for instance, the saliba manuscript[1], written in Spanish, Latin and Saliba, and dated in 1790, was developed to describe the grammar and usage of Saliba language by and for missionaries to understand the language and evangelize.

An example of typewritten dictionaries (including some drawings) were created by the Summer Institute of Linguistics (SIL)[2] in Colombia since 1962. This work was carried out for several languages, such as Wayuunaiki, Koreguaje, Muinane, and so on. These dictionaries are composed of two parts: first part is an explanation about grammar, alphabet, and sounds of speech. The second part is a list of words in indigenous language. Each word has the part of speech (e.g. noun, verb, affix), one or more meanings in Spanish, sometimes has an example with its translation into Spanish, and some others have a representative image. At present, the SIL has digitized dictionaries which were handwritten or typewritten. This kind of format presents disadvantages to perform some computational tasks as fast searches or build word lists according to part of speech (POS).

Nowadays, different programs are used in the production of dictionaries (design, use and application) to perform tasks as: corpus query, real-time preview, full customize, cross-references track and auto-update, automatic lemma reverse, automatically number and sort, multi-user support, and so on (Granger, 2012). For some languages as English, French, German, Spanish, web-based high quality dictionaries have been developed by their own speakers. Unfortunately, for indigenous languages (Nahuatl, Quechua, Aymara or Guarani) or more accurately, indigenous minority languages (Nasa, Wayuu, Embera in the Colombian case) remain unavailable. These people are interested in the use of web-based tools as a way to revitalize and document their own languages, but their participation is absent or very low, in part because of the low internet quality in the indigenous settlements, but mainly because of the speakers that do not have enough computer literacy required to create or to handle these tools. Certainly, a greater participation of the native speakers is essential to develop web-based dictionaries and to strengthen their languages.

[1] http://www.bibliotecanacional.gov.co/content/arte-de-la-lengua-saliba

[2] http://www-01.sil.org/americas/colombia/show_subject.asp?pubs=online&code=Dictionaries&Lang=eng

1.2. Some Basic Issues about Electronic Lexicography

The lexicography is a field which concerns with dictionaries. On the practical side, it is associated with creating a dictionary (planning, resourcing, compiling, writing, and editing). On the theoretical side, it is associated with developing models on the structural and semantic relationships between words within the dictionary. A branch of increasing interest is the electronic lexicography (Atkins and Rundell, 2008; Granger, 2012). It became an interesting research subject when Urdang (1966) showed the first approach to create the first edition of *the Random House Dictionary of the English Language*[3] by computer technology at the time. This seminal work broke down the information in the dictionary into seven categories in order to tackle the problem of getting the information into the computer: 1) illustrations, 2) pronunciations, inflected forms, and part of speech, 3) definitions, 4) variants, 5) etymologies, 6) words that are not defined because their meanings are self-evident, and 7) synonyms, word lists and usage notes.

1.3. Semantic MediaWiki Issues

MediaWiki[4] (MW) is a free and powerful wiki engine to process and display stored data in an easy and fast way. The versatility of this engine in using text (Wikitext), in integrating multimedia content as audio, images, and video, in easing the content managing between different versions, in being more efficient the access through hyperlinks, in customizing to the user needs, in combinating one or more types of dictionaries, in providing a collaborative framework, and not requiring prior knowledge by the users in HTML or CSS to insert information, makes it an outstanding tool to create a new kind of web-based dictionary (Granger, 2012, p. 3). Moreover, an useful extension to generate automatically lists and perform more efficiently searches in a wiki-dictionary is Semantic MediaWiki[5] (SMW). This extension allow us to query the dictionary as long as users tag the wiki's contents with explicit, machine-readable information, i.e. the main prerequisite of exploiting semantic technologies is the availability of suitably structured (semantic) data (Krötzsch et al., 2007). Furthermore, results are automatically updated in the page as long as they meet search patterns. For example, it is possible to list all verbs with their translations on a page. Thus, it is not necessary repeatedly perform the same searches.

Recently, there is a lot of interest relating to the development of dictionaries, vocabularies, glossaries and transcriptions to MW and SMW engines. We sort these studies in two fields according to their applications: *Medical Sciences* covering biology (Muljadi et al., 2006), medicine (Dueñas and Gómez, 2012), forensics (Baldatti and Griechisch, 2013), neuroscience (Larson and Martone, 2013); and *Human Sciences* concerning multilingual and multicultural disciplines (Khelifa et al., 2010), languages: medieval Latin (Bon and Nowak, 2013), Finno-Ugric (Laxström and Kanner, 2015), Muisca (Gómez, 2012), Uw cuwa (Gómez and Headland, 2013), Saliba (Dueñas and Gómez, 2015), Carihona (Gómez et al., 2015), and transcriptions of colonial manuscripts about indigenous languages: collection of National Library of Colombia (Gómez et al., 2012a) and Mutis collection (Gómez et al., 2012b). [GD: In cases of dictionaries of native languages of Colombia, the aim has been to create these dictionaries in a collaborative-non-intrusive way between experts and non-experts. We say in a collaborative way, because we provide the basic platform and native speakers can decide how and what information should be on it.

The purpose of this study is to show the implementation of MW and SMW in the elaboration of dictionaries or vocabularies. To accomplish this purpose, we will show the procedure for one dictionary and one vocabulary of Colombian languages: Saliba and Carijona, respectively. The paper is organized as follow: Section 2 describes two dictionaries of under-resourced languages created with MW and SMW. Section 3 presents how to visualize information of the dictionaries through semantic queries. Section 4 provides some discussion about the social implications of the use of technology by indigenous speakers. Finally, in Section 5 we provide some conclusions and perspectives for future work.

2. Two Lexicographical Colombian Examples of Endangered Language

In Colombia, besides Spanish, there are 64 or more indigenous languages classified into 13 linguistic families, two creoles, romani language, and Colombian sign language. In this section, we consider some applications of MW and SMW for two of them, Saliba and Carijona languages. The two examples are described in the sections 2.1 and 2.2. The Saliba people is located in Colombia (Vichada, Meta, and Casanare department) and Venezuela. There are an estimated between 1488 and 3035 indigenous in Colombia (González, 2011), but only 2 % of these indigenous speak the language in everyday life (Ramírez, 2010). Similarly, the Carijona people of Colombia is located in the departments of Guaviare, Vaupes, and Caqueta. There are an estimated between 30 and 425 speakers in Colombia (González, 2011), who do not use the language in everyday life. For these reasons, according to the Atlas of the World's Languages in Danger, Saliba is considered as "severely endangered", which means that "language is spoken by grandparents and older generations; while the parent generation may understand it, they do not speak it to children or among themselves", and Carijona is considered as"critically endangered", which means that "the youngest speakers are grandparents and older, and they speak the language partially and infrequently" (Moseley, 2010).

This is a worried situation, because if the few speakers of these languages disappear, components of diversity of our world disappear too. Dictionaries are sources of knowledge about languages. In our case, they are a wellspring of information about the vocabulary of the indigenous languages. They embody the sounds of languages of oral tradition.

[3] This dictionary was not published in a magnetic tape. In 1978, it appears the first dictionary which contains a lexical database, the *Longman Dictionary of Contemporary English* (Granger, 2012, p. 1).

[4] http://www.mediawiki.org/wiki/Manual:What_is_MediaWiki

[5] http://semantic-mediawiki.org

Variable	Name	Description
eti	Etimology	Name of the language if the word in *Lang1* is a loanword.
loc*	Locution	*Lang1* term
fon*	Pronunciation	Phonetic transcription of the locution.
cat_gra*	Part of speech	Grammatical category of the locution.
equ*	equivalence	Translate the locution into *Lang2*.
var_d	Dialect variant	Regional variations of locution written in *Lang1*.
ej_1*	Example	*Lang1* sentence where locution is used.
tr_1*	Translation	Example translated into *Lang2*.
sab_1*	Knowledgeable	Name of the person who provided the example and translation.
obs_gra	Grammatical observation	Grammatical rules that explain the behavior of locution in the language.
obs_cul	Cultural observation	Ethnographic and cultural data of locution.

Table 1: Variables of *acep* template for the Lang1-Lang2 dictionary

Nowadays, an amount of online resources concerning these two languages can be accessed, among others, by native young members of each community. We show the steps to create a bilingual dictionary using MW and SMW.

2.1. MW-SMW-Based on Saliba Dictionary[6]

The terms of this dictionary were collected by Hortensia Estrada during her field trips between 1993 and 2002. To insert the Saliba-Spanish and Spanish-Saliba dictionary words (named entry or headword), it is necessary to add additional information through two different templates. The *Lang1-Lang2* and the *acep* templates. In the former the order of the abbreviations indicates the direction of the dictionary. Thus, in this dictionary *Lang1* is Saliba and *Lang2* is Spanish. The order of the descriptions of templates is described in accordance with its creation priority.

2.1.1. Saliba-Spanish Template

`{{SAL-ESP}}`

This code consists of two left curly brackets {{, the abbreviation for the word Saliba (SAL), the abbreviation for the word Spanish (ESP), and two right curly brackets }}. As we said before, the order of the abbreviations indicates the direction of the dictionary. In this case it is a bilingual bidirectional dictionary, in which the entries will be contained and listed in the Saliba-Spanish dictionary.

[6]http://saliba.caroycuervo.gov.co

2.1.2. acep Saliba-Spanish Template

The another requirement for an entry belongs to Saliba-Spanish dictionary is to include lexicographical information through the *acep* template. This template lists the fields that an entry has. It is made up for two left curly brackets {{, the word *acep*, the abbreviation for the words of the Table 1 (variables marked with asterisk (*) are required), and ends with two right curly brackets }}. The variables of Table 1 are the same for both Saliba-Spanish and Spanish-Saliba direction, but the information stored in each variable of each dictionary is different. Each variable must be preceded by a vertical bar |. For example, the Saliba-Spanish entries have the *acep* template and the Spanish-Saliba entries have the *acep_es* template. The former template contains the variable *eti*, which stored the name of the language when it is known that the word in Saliba is a loanword, *loc* is the term in Saliba, *fon* stores the phonetic transcription of the Saliba word, *cat_gra* has to be selected from the options in the cat_gramatical template (see: section 2.1.3), *equ* stores the translation of the Saliba term into Spanish, *var_d* stores the regional variations of the Saliba term if it has, *ej_1* stores the expression or sentence where the locution is used, *tr_1* contains the example translated into Spanish, *sab_1* stores the name of the person who provided the example and translation, *obs_gra* contains the grammatical rules that explain the changes of the term in the language, and *obs_cul* stores ethnographic and cultural data of the term. An example of the previously described code is shown below for an entry in the Saliba-Spanish dictionary:

```
{{SAL-ESP}}
{{acep
  |eti=
  |loc=bae kelegiaja
  |cat_gra=s.
  |equ=muchas gracias
  |fon=bae kelegiaha
  |ej_1=Bae kelegiaja ortensia
  |tr_1=Muchas gracias Hortensia
  |sab_1=Belarmino Pónare Guacarapare
  |obs_gra=''Bae kelegiaja'' 'muchas
gracias' término conformado por
''bae'' 'bien' y ''kelegiaja'' 'hacer'.
Literalmente 'hacer el bien'
}}
```

2.1.3. cat_gramatical Saliba-Spanish Template

The purpose of this template is to list different grammatical categories in each dictionary. The code for this template consists of two <*includeonly*> markups at the beginning and end of the code, two left curly brackets {{, the *#switch* function that compares an input value (inserted between the left-right curly brackets {{{ . }}}) against different listed cases, and two right curly brackets }}. When a match is found, this function returns the value associated with that case and it is stored in the *cat_gra* field. Each case is composed of a vertical bar, the variable name, and the assigned value to the variable. An example of the previously described code is shown below for some grammatical categories in the Saliba-Spanish dictionary (Ramírez, 2010; Ramírez, 2011; Ramírez, 2015).

```
<includeonly>
{{#switch: {{{1}}}
  | num. an. = Numeral animado
  | num. an. pl. = Numeral animado
plural
  | num. in. = Numeral inanimado
  | num. in. pl. = Numeral inanimado
plural
}}
<includeonly>
```

Next, we explain briefly the Spanish-Saliba dictionary in a similar way as we proceed before highlighting only the possible differences in both cases.

2.1.4. Spanish-Saliba Dictionary

The main template `{{ESP-SAL}}` shows the direction for which the entries will be contained and listed into the Spanish-Saliba dictionary. As this is a bilingual bidirectional dictionary, we name the *acep* template for the ESP-SAL dictionary as *acep_es*. From this template, we do not consider the *eti* variable because we focus on indigenous language information. Of course the labels *Lang1* refers to Spanish and *Lang2* refers to Saliba and all the information of the Table 1 must be provided.

For instante, now the variable *loc* stores the term in Spanish, *fon* stores the phonetic transcription of the Saliba word, *cat_gra* have to be selected from the options in the *cat_gramatical_es* template, *equ* stores the translation of the Spanish term into Saliba, *var_d* stores the regional variations of the Saliba term if it has, *ej_1* stores the expression or sentence where the locution is used, *tr_1* contains the example translated into Saliba as we show below:

```
{{ESP-SAL}}
{{acep_es
  |loc=muchas gracias
  |cat_gra=s. pl.
  |equ=bae kelegiaja
  |fon=bae ke-le-gi-a-ha
  |ej_1=Muchas gracias Hortensia
  |tr_1=Bae kelegiaja ortensia
  |sab_1=Belarmino Pónare Guacarapare
  |obs_gra=''Bae kelegiaja'' 'muchas
gracias' término conformado por
''bae'' 'bien' y ''kelegiaja'' 'hacer'.
Literalmente 'hacer el bien'
}}
```

We name the *cat_gramatical* template for the ESP-SAL dictionary as *cat_gramatical_es*, because it uses only grammatical categories of this language (Spanish). An example of this template is shown below:

```
<includeonly>
{{#switch: {{{1}}}
  | s. m. = Sustantivo masculino (es)
  | s. f. = Sustantivo femenino (es)
  | s. m. pl. = Sustantivo masculino
plural (es)
  | s. f. pl. = Sustantivo femenino
plural (es)
}}
<includeonly>
```

2.2. MW-SMW-Based Carijona Dictionary[7]

The term list of this vocabulary can be viewed in (Robayo, 2000). Unlike the Saliba dictionary, which is bilingual bidirectional, the Carijona dictionary is bilingual unidirectional. To create Carijona-Spanish dictionary we used the above mentioned templates. The main template `{{CAR-ESP}}` shows the direction for which the entries will be contained and listed into the Carijona-Spanish dictionary.

2.2.1. acep Carijona-Spanish template

The description of the variables in Table 2 are the same as those in Section 2.1.2, but in this case, they are few. For instance, this dictionary does not have examples, because the main objective was to collect the Swadesh list. This dictionary does not have the *locution* field filled, because this language still does not have its unified spelling system. The speakers developed a highly oral tradition[8]. The *cat_gra* variable has to be selected from the options in the *cat_gramatical* template (see: section 2.2.2). An example of the previously described code is shown below for an entry in the Carijona-Spanish dictionary:

```
{{acep
  |loc=
  |cat_gra=s.
  |equ=cielo
  |fon=kahu
}}
```

Variable	Name	Description
loc	Locution	Carjona word
cat_gra	Part of speech	Grammatical category of the pronunciation
equ	equivalence	Translate the pronunciation into Spanish
fon	Pronunciation	Phonetic transcription

Table 2: Variables of *acep* template Carijona Dictionary

2.2.2. cat_gramatical Carijona template

This code works in the same way as the Saliba code, but the difference consists in the number and the type of grammatical categories. As this is a small vocabulary, we have used the most common grammatical categories. An example is shown below for some grammatical categories in the Carijona-Spanish dictionary:

```
<includeonly>
{{#switch: {{{1}}}
    | s. = Sustantivo
```

[7] http://carijona.caroycuervo.gov.co

[8] A phonological analysis of the consonant and vowel timbre was developed by (Robayo, 1983a; Robayo, 1983b).

```
    | adj. = Adjetivo
    | v. = Verbo
    | adv. = Adverbio
}}
<includeonly>
```

In summary, the fields are easy to complete. However some of them can confuse the user when entering information. Users must understand what kind of information is going in each field, if it is a equivalence, an example or a translation.

3. Displaying Information from Dictionaries

We have created dictionaries using templates, because they permit ordering the information in a easy way. The information stored in these templates is raw text. Another way of entering the information is by means of a Semantic Form. The information of any variable can be queried by making use of the "Semantic search"page or the parser function *#ask*. In this paper, we show an example using the latter. One advantage is that the query results can be presented in a suitable order as long as the pages have the property. If at least a property is listed in a query, SMW will search only results in pages that have at least one value for this property. The following code lists all the verbs in Saliba dictionary with their locution, equivalence in Spanish, and part of speech.

In this case, the parser function *#ask* is followed for a category called verb, and the latter is followed by three properties. First, SMW will search for all entries matching the category, then it will list the locution, the equivalence, and the part of speech of each entry. The function of the vertical bar or the pipe symbol is to separate property conditions to display. The function of the question mark followed by the property name is to display all the values assigned to a certain property as is shown in the example below:

```
{{#ask:
[[Categoría:Verbo]]
    | ? locucion
    | ? equivalencia
    | ? categoria_gramatical
}}
```

4. Contribution

Our results confirm that bilingual dictionaries can be built when MediaWiki and Semantic MediaWiki are implemented. Furthermore, we have shown how to display information from dictionaries by means of queries written in MediaWiki syntax. Unlike handwritten or typewritten dictionaries which are digitized, creating dictionaries using virtual tools as MW and SMW allow us to display any kind of information that is in the dictionary as long as it has been tagged. We show how to create templates to store information in order to be able to retrieve it for further work; for example, supporting human translators or machine translation. This is the main difference with respect to the most of cited papers, which only show the systems running, not how they created such systems.

We suggest some directions that can configure a country to use language technologies such as software tools to document, to preserve and to spread widely the native languages would be a breakthrough in the struggle against inequality of linguistic rights and digital opportunities for all languages and for their speakers. However, there are two scenarios which allow us to understand why to get this goal is so difficult. On the one hand, there are communities that have interacted little or nothing with these tools, because they do not know them, or they even know them but the advantages (and disadvantages) have not been shown in favor of communities, or simply because of cultural rejections. On the other hand, there are established laws and actions made by the Government, which works to improve the conditions of these languages, but its efforts are insufficient. For example, the Colombian Government approved the law 1381 in 2010, which sets rules about recognition, promotion, protection, use, preservation and strengthening of the languages of ethnic groups and their linguistic rights. This law commands Ministry of Culture, through the ICC and other entities, will encourage collection, preservation and dissemination of written materials, audio and audiovisual representation of native languages and oral traditions produced in these languages. Also, the Ministry will provide access to new technological media and communication to speakers of native languages using documents in native languages and encouraging the creation of Internet portals for this purpose. As for research programs and training, Colciencias[9] will support researches and documentations about indigenous languages, and will ensure that the results are known to the communities (de Colombia, 2010).

Despite of what the law states, the current reality is another. The support for researches in Human Sciences is very small since resources are allocated to the field of exact sciences in an absurd disproportion. For example, the support for researches framed in the 1381 law is limited. The ICC provided only one stimulus for research in linguistics (Spanish, indigenous languages, Afro-Colombian, Rom or sign of Colombia) by around 8000 dollars in 2015. Meanwhile, for the same year, "Direccion de Poblaciones" office[10] under control of the Ministry of Culture, offered twelve stimuli for strengthening indigenous languages at risk of Colombia, each one also by around 8000 dollars. The latter stimuli are no longer offered in 2016. As for Colciencias, there is high distrust in this institution by researchers in Human Sciences. The researchers argue that Colciencias considers the Human Sciences do not contribute to the development of Colombia. Hence, to encourage the use of languages technologies and to benefit the Colombian context, first the Ministry of Information Technologies and Communications of Colombia must ensure stable internet access in places where native speakers are located. Subsequently, both the Ministry of Education and Culture should support and finance the projects in which communities wish to incorporate new technologies. Logically, these projects must be accompanied by topic related professionals. In addition, and not least, these projects should be coordinated with projects of education, health, etc., of the communities, so that

[9] Colciencias is the Administrative Department of Science, Technology and Innovation in Colombia.

[10] The purpose of this office is to guide and implement policies, programs and projects that advance the understanding of culture as an integral part of the development of Colombia.

in this way the projects could have continuity.

In view of the foregoing, this kind of work is useful for a country like Colombia or similar, because it is a country where the most of the minority language populations have not virtual resources to interact with juridical institutions, public administration, health services, education system, and so, to demand their rights. It is because of the Internet platforms of the above institutions and even more government documents are written only in Spanish, and not any translation to these languages is provided. Although there are laws which promote the inclusion and disclosure of the information on these languages, there have had difficulties to implement them in the practice. For instance, there is not appropriate virtual media and enough staff to communicate with speakers of minorities in these institutions.

Recently, on the one hand, few tools are being created to overcome this situation, mainly by people outside the communities. However, the progress of these tools is delayed, because these people do not know deeply how each language works. On the other hand, the current knowledge based on or registered in indigenous languages is insufficient; to wit, researchers should perform several field trips to collect quality data. Even they must spend more time when the place where communities are located is far or complicated to arrive. Hence, the native speakers have no access to reliable information in their languages. It promotes social inequalities to these communities limiting the benefits they can obtained by public policy. The above situation occurs by the absence of synergy between native speakers and government to build and implement public policies. This scenario can be improved promoting research and training programs on these issues aimed to native speakers.

One possible solution to the problem of not having access to reliable information in their languages is to build electronic dictionaries. Thus, they contribute to struggle against the inequality of digital opportunities as for all minority language groups as for all their inhabitants, because when native speakers create and continue dictionaries, they acquire and improve their knowledge about both their languages and computer topics (if the latter are used to create them). An additional contribution of these resources is to motivate state entities responsible for technological development to implement mechanisms that allow access to technological developments; for example, establish internet access in remote areas where there are minority communities.

We expect that additional work by the communities improve considerably the information and the use of the dictionary. In order to control and guarantee the quality of the entered information by Saliba people, the ICC has assigned an expert researcher in the language to review the consistency of itself for which the use of a web-based platform (Mediawiki) instead of programs for personal computers is less demanding. In the case of Carijona platform, there is a research group at the Linguistics Department of the National University of Colombia which reviews the consistency of itself.

5. Conclusion

In this paper we have shown it is possible to create a bilingual dictionary for indigenous languages by means of collaborative system as MediaWiki. We have used MW and SMW software to create two dictionaries for Colombian indigenous under-resourced languages. Saliba language dictionary is a bilingual bidirectional dictionary and Carijona language dictionary is a bilingual unidirectional dictionary. This system can be opened to anyone, but is expected to be used by native people with knowledge in both language and computer skills.

Each template is configured to store and list the terms in accordance with the direction of the dictionary; that is, if we register a new entry, we must first decide whether it is in the dictionary of the native language or is in foreign dictionary. For this we use the {{Language1-Language2}} template which indicate the direction of the dictionary. Then, we have to create the *acep* template, which list the fields that the entries will contain. At the same time, we have to create the *cat_gra* template, which list the grammatical categories that the term will be assigned.

We consider that main contribution of this tool is to display real-time information, i.e. this tool allows native speakers residing or not in their native communities to enter information related to their language. In the case they do not reside in their communities, perhaps, by displacement (forced or not) or other situations, this tool allows them to enter information that can be viewed by other speakers who are in the community or elsewhere. Another advantage is to integrate developed extensions for different tasks; for example, integrate a Semantic Maps extension for georeferencing. We hope this platform promotes more involvements of the native speakers and other researchers with the language technologies. In future work, we plan to incorporate sound and images to each entry to know how it sounds and how it looks according to the worldview of native speakers. Also, we intend to create an algorithm for automatic conjugation of verbs.

6. Acknowledgements

The authors acknowledge the support and the sponsorship of the Instituto Caro y Cuervo under the project "Diccionario electrónico sáliba-español: una propuesta de documentación de la lengua y la cultura sálibas". We would like to thank Hortensia Estrada and Camilo Robayo for gathering, studying, and sharing their data with us, and J. G. Dueñas for discussions and revising the manuscript.

7. Bibliographical References

Atkins, S. and Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Baldatti, I. and Griechisch, E. (2013). Joint glossary of forensic document examination and pattern recognition. *Proceedings of the 2nd International Workshop on Automated Forensic Handwriting Analysis: A Satellite Workshop of International Conference on Document Analysis and Recognition (ICDAR 2013)*, pages 51–55.

Bon, B. and Nowak, K. (2013). Wiki lexicographica. linking medieval latin dictionaries with semantic mediawiki. In *Proceedings of the eLex 2013 conference*, pages 407–420.

de Colombia, C. (2010). Ley 1381 "ley de lenguas".

Dueñas, G. and Gómez, D. (2012). Diccionario académico de la medicina. http://dic.idiomamedico.net/.

Dueñas, G. and Gómez, D. (2015). Diccionario electrónico sáliba-ẽspanol: una herramienta interactiva para la documentación de la lengua y de la cultura sálibas. *Forma y función*, 28(2):49–61.

Gómez, D. and Headland, E. (2013). Diccionario uw cuwa-español. http://uwa.cubun.org/.

Gómez, D., Dueñas, G., Torres, J., and Pérez, Y. (2012a). Colección de lingüística misionera colonial de la biblioteca nacional de colombia. http://coleccionmutis.cubun.org/BNC.

Gómez, D., Torres, J., Giraldo, D., Sierra, O., and Soler, C. (2012b). Colección mutis. http://coleccionmutis.cubun.org/.

Gómez, D., Dueñas, G., and Robayo, C. (2015). Diccionario karihona-español. http://carijona.caroycuervo.gov.co/.

Gómez, D. (2012). Diccionario muisca-español. http://muysca.cubun.org/Categoría: Dictionary.

González, M. S. (2011). *Manual de divulgación de las lenguas indígenas de Colombia*. Publicaciones del Instituto Caro y Cuervo: Series minor.

Granger, S., (2012). *Electronic Lexicography*, chapter Introduction: Electronic lexicography-from challenge to opportunity, pages 1–11. Oxford University Press.

Khelifa, L., Lammari, N., Fadili, H., and Akoka, J. (2010). A wiki-oriented on-line dictionary for human and social sciences. In *Fifth Workshop on Semantic Wikis Linking Data and People (SemWiki2010)*, pages 79–88.

Krötzsch, M., Vrandečić, D., Völkel, M., Haller, H., and Studer, R. (2007). Semantic wikipedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(4):251–261.

Larson, S. D. and Martone, M. E. (2013). Neurolex.org: an online framework for neuroscience knowledge. *Frontiers in neuroinformatics*, 7.

Laxström, N. and Kanner, A. (2015). Multilingual semantic mediawiki for finno-ugric dictionaries. In *Septentrio Conference Series*, number 2, pages 75–86.

Moseley, C. (2010). *Atlas of the World's Languages in Danger*. UNESCO, 3rd edition. http://www.unesco.org/culture/en/endangeredlanguages/atlas.

Muljadi, H., Takeda, H., Kawamoto, S., Kobayashi, S., and Fujiyama, A. (2006). Towards a semantic wiki-based japanese biodictionary. In *Proceedings of the 1st Workshop on Semantic Wikis (ESWC2006)*, pages 202–206.

Ramírez, H. E. (2010). La modalidad epistémica en la lengua sáliba. *UniverSOS Revista de Lenguas indígenas y universos culturales*, (7):107–118.

Ramírez, H. E. (2011). Categorías léxicas del adjetivo en la lengua sáliba. *UniverSOS Revista de lenguas indígenas y universos culturales*, (8):79–94.

Ramírez, H. E. (2015). El sistema de clases nominales en la lengua sáliba. *Revista Brasileira de Linguística Antropológica*, 6(1):137–164.

Robayo, C. (1983a). *Analisis fonológico de timbres vocálicos de la lengua carijona*. Instituto Caro y Cuervo.

Robayo, C. (1983b). *Cuadro fonológico de timbres consonanticos de la lengua carijona*. Instituto Caro y Cuervo.

Robayo, C. (2000). Avance sobre morfología carijona. *Lenguas indígenas de Colombia: una visión descriptiva*, pages 171–180.

Urdang, L. (1966). The systems designs and devices used to process the random house dictionary of the english language. *Computers and the Humanities*, 1(2):31–33.

Quantitative and Qualitative Analysis in the work with African Languages

Dorothee Beermann[1], Tormod Haugland[1], Lars Hellan[1]
[1] Norwegian University of Science and Technology, Norway
Email: dorothee.beermann@ntnu.no, tormod.haugland@gmail.com, lars.hellan@ntnu.no

Uwe Quasthoff[23], Thomas Eckart[2], Christoph Kuras[2]
[2] NLP Group, Dept. Computer Science, University Leipzig, Germany
[3] Dept. of African Languages, University of South Africa
Email: {quasthoff, teckart, ckuras}@informatik.uni-leipzig.de

Abstract

We discuss the development of a combined search environment for the Leipzig Corpora Collection (LCC) and the TypeCraft Interlinear Glossed Text Repository (TC). This digital infrastructure facilitates corpus methodologies for under-resourced languages. By providing multiple-site accessibility of all data, we hope to give a new impetus to linguists and language experts to employ digital services for data analytics. The definition of export and import APIs using Web Services are shown to be useful for a collaboration between two different projects, and to extend and combine existing linguistic material. In this way we also increase the access to data from under-resourced languages.

Keywords: corpus methodologies, quantitative analysis, qualitative analysis, African languages

1. Introduction

This paper describes a digital infrastructure project. Its purpose is to develop an integrated search environment for the Leipzig Corpora Collection (LCC) and TypeCraft (TC v2.3) which allows for an easy and controlled access to our online corpora and linguistic tools. The project thereby aims to promote quantitative and qualitative approaches to language studies in research and teaching. We explicitly address work with under-resourced languages where the need for corpus creation is as important as the need to ensure that researchers and educators are able to access these resources. To this end, powerful corpus search and corpus creation facilities need to be integrated seamlessly. Especially for the work with under-resourced languages it is important that resources are usable in a collaborative setting and for community-driven language development. The LCC-TC resources have in common that they present monolingual corpora in a multilingual setting. While the LCC offers monolingual corpora of standard sizes from different sources such as the web, newspapers and the Wikipedia, TypeCraft is specialised on Interlinear Glossed Text, that is, corpora of manually morpheme-to-morpheme annotated natural language text. The combination of these different resources introduces new possibilities for both sides.

The question we will focus on in the following is how our combined resources can best be accessed and queried. To this end, we will discuss the prototype of an LCC-TC environment which allows its users to work with self-defined corpora from already existing external or LCC-TC internal resources. LCC-TC is not unlike Sketch Engine in that it facilitates the creation of custom-made corpora. The difference is that LCC-TC also allows its users to add further annotations to text, using the various facilities that the LCC-TC environment can offer. In this way the LCC-TC brings us one step closer to a suitable research management system for textual data which is able to support the workflow of linguistic research projects.

The paper is organised as follows: the first three sections will describe the participating research groups with a focus on their strengths and primary areas of expertise, followed by more detailed information about the technical prerequisites for the presented collaboration in section 4. Section 5 will elaborate on the general approach for two independent but interlinked query interfaces including data analytics and visualisation. Section 6 will discuss data availability in the context of languages spoken in Central, West and East Africa. The paper closes with some remarks on future work.

2. The Leipzig Corpora Collection

The Leipzig Corpora Collection (LCC) provides monolingual corpora in more than 200 languages. The corpora are collected and processed with the same processing chain, the data are available online both for querying and download. The following different genres are collected: Newspaper texts (using a newspaper directory like abyznewslinks.com), Wikipedia and randomly crawled Web pages.

For the processing of the LCC corpora, there is a language independent processing chain (Goldhahn et al., 2012) developed during the last years: Crawling using the Heritrix Web crawler (Mohr et al., 2004), text extraction from HTML files, sentence segmentation, language identification, pattern-based quality checking and

deduplication. This pattern-based cleaning uses language-independent rules described by regular expressions. If necessary for a certain language, these rules can be overridden. After word tokenization, an inverse list connects words with its occurrences in sentences and the URLs as sources. Moreover, word co-occurrences are calculated. All data are stored in a MySQL database.

For a deeper analysis, the NoSketch Engine (Rychlý, 2007) can be used: Any additional layer of annotation (often POS tagging, but also more elaborated annotation as for the language Luganda below) can be queried using the Corpus Query language CQL (Christ, 1994).

Figure 1: KWIC with several annotation layers in the NoSketch Engine for a Luganda text corpus.

3. TypeCraft Annotation Tool

The TypeCraft (TC)[1] application (Beermann & Mihaylov, 2014) is an Open Infrastructure that allows for the creation and retrieval of Interlinear Glossed Texts (IGT) - the standard data format in linguistics. TC is a user-driven database which offers functionalities needed for the management of textual data. Its main function is to enable the sharing of linguistic data sets, such as transcribed and annotated oral narrations, annotated small texts, and linguistic collections exposing phenomena of special interest to linguists, such as *multi-verb constructions*, *valence frames*, t*ense-aspect systems, infinitival* and other *hypotactic construction* types (to just name some). At present TC hosts 2137 texts from 146 languages. An overview of the size of the database is given in Table 1.

Data type	*Data count*
Text count	2145
Phrase count	316,604
Word count	5,297,405
Morpheme count	4,527,478
Part-of-speech tagged words	4,851,807
Gloss-tagged morphemes	330,714
Sense-tagged morphemes	1,173

Table 1: TypeCraft database in terms of stored data and annotations assigned.

TC works with closed vocabularies. The size of the word level tag sets is shown in Table 2.[2]

Data type	*Data count*
Gloss tags	360
Part-of-speech tags	101
Sense tags	53

Table 2: Size of the main TypeCraft annotation sets.

The application itself consists of the following modules:

TypeCraft Importer: The importer is a lightweight web-application. At the time of writing, it allows for the import of Toolbox (ELAN) and TypeCraft-XML data into the database.

TypeCraft Editor: The TypeCraft editor is the primary tool of the TypeCraft system, and is used for manual annotation of linguistic data. Users work with several predefined layers of annotation and controlled vocabularies.

TypeCraft Exporter: An internal system handling exportation queries to the TypeCraft database. The exporter is capable of passing on phrases and text in a variety of formats, e.g. the word list-export.

TypeCraft Mediawiki: The TypeCraft wiki is powered by a Mediawiki instance with several extensions. The Wiki is maintained by the TypeCraft users and the content mainly addresses lesser-described languages.

Phrase Renderer: The dialogue displayed in the TypeCraft Editor when a user opens a phrase.

The TypeCraft Search Interface: TC has a graphical menu-based interface presented as part of its Mediawiki. It allows for basic aggregations but at the moment no direct searches of the database. The TC search facilities are a topic in section 5.

4. Connecting resources and tools

Since 2005 the NLP group at the University of Leipzig provides their resources via SOAP Web services to interested communities (Büchler & Heyer, 2009). Since then around a billion requests were handled by these APIs[3]. With the participation of the group in the German branch of the CLARIN research infrastructure (Wittenburg et al., 2008) increasing effort was put in replacing the SOAP-based services with RESTful equivalents[4]. Via this new JSON API, language material for more than 100 different languages is accessible and

1 http://typecraft.org

2 a list of TC gloss and pos tags can be found at: http://typecraft.org/tc2wiki/Special:TypeCraft/GlossTags/ and http://typecraft.org/tc2wiki/Special:TypeCraft/POSTags/ respectively.

3 http://wortschatz.uni-leipzig.de/Webservices/

4 http://wortschatzwebservices.informatik.uni-leipzig.de

the number of supported services, languages and the extent of provided material is steadily increased.

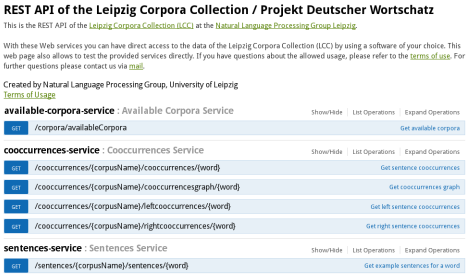

Figure 2: Documentation of the LCC REST-API.

In the context of LCC-TC this API is used to request relevant material directly from the TC user interface. This material includes reference sentences for relevant terms, which can be directly imported and annotated in the Typecraft annotation tool. Because of a lack of available tools for linguistic annotation for the considered languages (like parsers or part-of-speech taggers) it is expected that mostly plain text enriched with statistical annotations will be imported into TC.

TypeCraft (v2.3) exposes its database to search and data collection through a search interface and a simple RESTful API. As already mentioned it is intended to extend the search interface to directly query LCC data for a subsequent import. The TypeCraft system includes an import tool that is capable of importing annotated text in several formats. For the purpose of this project it will be extended to allow the end-user an import directly from the LCC. For the import of TC data into the LCC system the TypeCraft REST API will be used as primary interface.

5. General approach

In the following we discuss the requirements for a query environment, which (a) is suitable for work with under-resourced languages and (b) adequate for linguistic mining of different types of corpora. The general idea is, that we would like our users to be able to leverage the advantages of a combined quantitative (word frequencies, co-occurrence patterns, semantic maps) and qualitative analysis of grammatical features and their distribution. Needless to say, all standard requirements for a well-defined corpus interface also apply here (Soehn et al., 2008). This concerns the search itself, the visualisation of the data and the export of the query results. Describing ongoing development, we look at which information can be requested by the user (word, phrases, annotations etc.). We also are concerned with the input format, and with the display of search results ('show or seek' display); we finally describe the handling of the search results. We first will look at each query interface separately:

5.1. LCC

The LCC corpora are stored in a MySQL database and accessible via several interfaces including a RESTful API and several web interfaces with focus on varying user interests. Some of the queries are predefined and directly accessible. Provided material contains

- words, their frequency. Some multiword units (taken from Wikipedia)
- string-similar words
- if provided by a POS tagger: POS and lemma form for words
- sample sentences, ordered by GDEX (Rychlý et al., 2008)
- word co-occurrences (within a sentence or with immediate next neighbours)
- semantically similar words (sharing many word co-occurrences)
- many predefined corpus statistics including length distribution for words and sentences, distribution of sources, and much more.

The number of queries is ever increasing, and customized queries are possible. Moreover, the POS tagged corpus in the NoSketch Engine allows for the full power of CQL queries by using regular expressions for words, lemmas and POS tags. The optional use of the UD17 tag set (Petrov et al., 2012) provides tagger-independent querying with a standardized tag set, and frequency distributions are provided for all queries. Furthermore the search can be restricted to user-defined subcorpora.

5.2. TypeCraft Queries

A TC search operates on phrases, which means that the result of a query is a phrase level representation. Search results can be represented in a 'show' display where we list sentences. We highlight the search term and do some basic aggregations counting numbers of phrases and tokens that satisfy the search. The user can select a narrow view of sentences which gives a detailed view of the interlinear glosses, and allows the linguist access to sentence internal information (seek-display). These search results can be browsed as HTML files for further data exploration using general browser functionality. TypeCraft allows multi-layered search in text-fields and via drop down menus; word or morpheme queries can relatively freely be combined with a search for specific glosses or combinations of glosses, co-occurring either in a phrase, or in a word. The latter distinction is useful when we want to compare tokens containing for example tense-aspect markings, or a single verb, and distinguish these from those spread over the phrase in the form of a periphrastic construction.

5.3. Queries in the LCC-TC corpora

Ultimately, the API exposing the LCC will be used conjointly with TypeCraft's search and import features to allow cross-project work. TypeCraft (v2.3) exposes its database to search through an interface as described

above, and a simple RESTful API. At the project's end, the TC search facilities and the handling of search results will be extended to allow for queries directly to the LCC. This means that sentences containing relevant terms can be imported to the TypeCraft Editor for further annotation. The amount of added annotations depends on the task at hand and may include the tagging of word and phrase level lexical, morpho-functional, syntactic and semantic information. To illustrate the usage of our corpora in the work with under-resourced languages, we would like to briefly discuss our Luganda resources. Luganda (ISO 639-3 'lug') is a major language in Uganda, spoken by around 5 million people. As part of our LCC-TC corpora, Luganda is one of the smaller languages. In the LCC it is represented with 13,000 sentences, consisting of 190,000 word tokens. The material consists mainly of newspaper articles crawled in 2011 and 2013 enriched with material from Wikipedia. The TypeCraft corpus consists of 2511 in-depth annotated sentences (IGT data) corresponding to 5609 words. The material has been annotated by native speaker graduate students as part of linguistic graduate work. The most frequent POS category for the Luganda corpus is by far verb, followed by nouns, conjunctions and proper nouns, as shown in Figure 3.

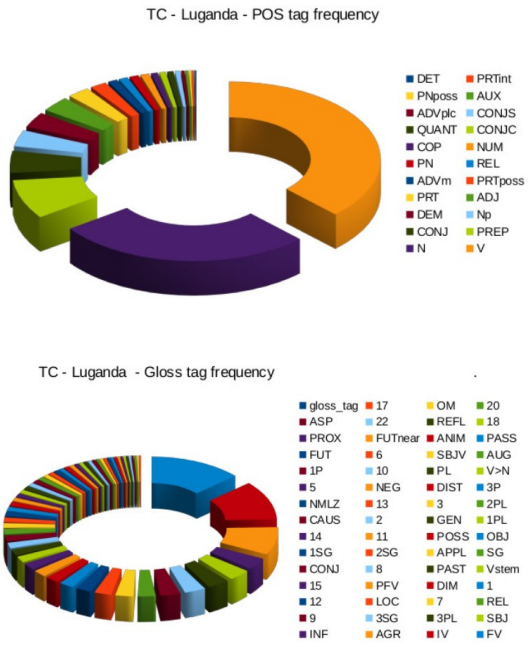

Figure 3: Frequency of POS and Gloss tags in the TC Luganda corpus

The most frequently assigned gloss tag is the final vowel (FV) which is an obligatory verbal suffix. Its form is conditioned by the tense/aspect/mood specification as well as by the polarity of the verb. The initial vowel (IV) on the other hand occurs on nouns, and conveys information about definiteness/specificity. The high frequency of the latter glosses correlates with the frequency of verbs and nouns in our corpus. Already now our corpus allows for an analysis of relative frequencies such as those of verbs and nouns, which is a linguistic parameter of some interest as has been shown in Bickel et al., 2013. Specific to Bantu languages is that the relation between IVs and nouns is one indicator of referential density (in the sense of (Stoll & Bickel, 2009)). At present our combined Luganda corpus is still small, but already now word search in the Luganda LCC corpus is fully possible. The word **abantu** means people and 'nti' , one of its most frequent left neighbours is a demonstrative meaning *these*. A frequent right neighbour is the word *bangi* which means *many*. The profiling of the Luganda noun phrase structure will improve as the size of our corpus increases.

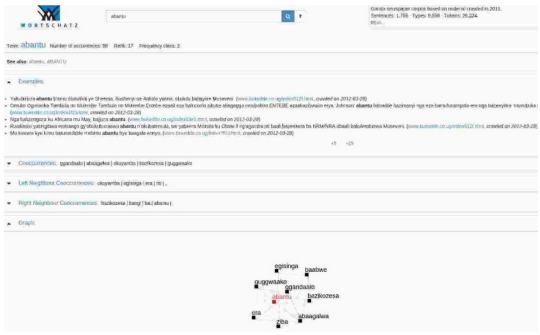

Figure 4: LCC search result for "abantu" in a Luganda newspaper corpus from 2011.

6. Future work

The approach described in this paper can be generalized to other languages. Both the LCC and TypeCraft contain material for many more languages spoken in the considered regions. Among those, many lack any linguistically annotated material or even a base stock of raw text. For many years the LCC has performed several text acquisition methods to gather material in as many languages as possible. Many of these approaches (like exploiting standard Web search engines or bulk crawling of complete top level domains) have proven to be problematic when dealing with under-resourced languages. It is expected that the deeper knowledge about these languages and a direct feedback of the TypeCraft project will enhance future text acquisition approaches and will lead to more complete images of the present stock of online available material.

As a major benefit of the cooperation it is planned to utilize both kinds of expertise in an improvement and quality assurance loop with the ultimate aim of larger and more systematic annotated textual material for these under-resourced languages. Figure 5 visualizes this

process of constant feedback and new enhanced text acquisition procedures.

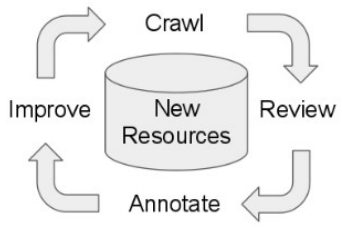

Figure 5: Intended improvement process in the LCC-TC cooperation.

Even without new input material the current stock in both projects allows the adaption of the workflow for more languages, as a short overview of the LCC and Typecraft resources for some other Central and East African languages shows (Table 3 and 4). Whereas the LCC data is mostly newspaper or Web crawled material, the listed TC material is IGT data created by graduate students enrolled in linguistics at NTNU[5] between the years 2007 – 2015. Table 4 focuses on TypeCraft's African languages with more than 10,000 word tokens.

Language	*Spoken in*	*Sentences*	*Word Tokens*
Akan	Ghana	505	7208
Ewe	Ghana, Togo	1000	12,476
Ganda	Uganda	13,183	189,825
Lingala	Dem. Rep. Congo	2783	40,048
Tigrigna	Ethiopia, Eritrea	1852	23,667
Yoruba	Nigeria	12,886	203,592

Table 3: Current LCC data stock of under-resourced languages of the considered region (excerpt).

Language	*Word Tokens*
Akan	79,428
Mandinka	24,134
Runyankore-Rukiga	103,314
Ga	16,799

Table 4: Current TC languages with more than 10,000 word tokens.

With a more pronounced support of corpus methodologies under development, new representations of data, for example in terms of expressions of higher level formal languages, might be of value in the future. The generation of attribute value matrices displaying functional information or syntactic and dependency trees come to mind as desirable extensions to the general workflow. TypeCraft annotations go beyond the IGT format, as they allow the markup of dependence and syntactic information. Especially with these annotations in place the TC resources are sufficiently rich to allow for the (semi)-automatic creation of additional linguistic visualisations of datasets. One way to achieve this is by mapping lexical resources onto formats understood by constraint-based parsers (Hellan & Beermann, 2014). In this way it becomes possible to project LFG and HPSG style representations.

5 Norwegian University of Science and Technology

7. Conclusion

We have presented a collaboration between the Leipzig Corpora Collection (LCC) and the TypeCraft Interlinear Glossed Text Repository (TC). Both partners offer linguistic services, and together we can present a digital infrastructure that allows the work with quantitative and qualitative corpus methods. We have argued that our facilities are particularly suited to introduce online corpus methodologies more actively into the work with under-resourced languages. By providing multiple-site data access to under-resourced languages, we hope to give a new impetus to linguists and language experts to employ digital services for data analytics. We further have argued that the export and import APIs using Web Services are useful for a collaboration between projects. Our set-up will allow us to extend our resources more efficiently than otherwise, and to prepare them for linguistic use where quantity of data counts as much as a certain depth of annotation. But most importantly our infrastructure allows us to give linguists and language experts direct working access to data from under-resourced languages.

8. Acknowledgements

This research is supported by the German-Norwegian collaborative research support scheme of DAAD (German Academic Exchange Service) and the Research Council of Norway (Norges Forskningsråd).

9. Bibliographical References

Beermann, D., Mihaylov, P. (2014). TypeCraft collaborative databasing and resource sharing for linguists. Lang. Resour. Eval. 48, 2 (June 2014), 203-225. http://dx.doi.org/10.1007/s10579-013-9257-9.

Bickel, B., Strunk, J., Seifart, F., Pakendorf, B., Witzlack-Makarevich, A., Danielsen, S., Wichmann, S., Zakharko, T. (2013). Noun-to-verb ratio and grammar. Paper presented at the international workshop "The relative frequencies of nouns, pronouns, and verbs in discourse", Leipzig, August 12-13, 2013.

Büchler, M., Heyer, G. (2009). Leipzig Linguistic Services - A 4 Years Summary of Providing Linguistic Web Services . In: Proceeding of TMS 2009

conference, Augustusplatz 10/11, 04109 Leipzig, Germany, 2009.

Oliver, C. (1994). A modular and flexible architecture for an integrated corpus query system. In Papers in Computational Lexicography (COMPLEX '94), pages 22–32, Budapest, Hungary.

Goldhahn, D., Eckart, T., Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012.

Hellan, L., Beermann, D. (2014). Inducing grammars from IGT. In Z. Vetulani and J. Mariani (eds.) Human Language Technologies as a Challenge for Computer Science and Linguistics. Springer.

Mohr, G., Kimpton, M., Stack, M., Ranitovic, I. (2004). "Introduction to Heritrix, an archival quality web crawler" (PDF). Proceedings of the 4th International Web Archiving Workshop (IWAW'04).

Petrov, S., Das, D., McDonald, R. (2012). A Universal Part-of-Speech Tagset. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey.

Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno : Masaryk University, 2007. p. 65-70. ISBN 978-80-210-4471-5.

Rychlý, P., Huask, M., Kilgariff, A., Rundell, M., McAdam, K. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In Proceedings of the XIII EURALEX International Congress. 1. vyd. Barcelona: Institut Universitari de Lingüística Aplicada, 2008. p. 425-432, 7 pp. ISBN 9788496742673.

Soehn, J-P, Zinsmeister, H., Rehm, G. (2008). Requirements of a User-Friendly, General-Purpose Corpus Query Interface. Proc. of the LREC 2008 Workshop "Sustainability of Language Resources and Tools for Natural Language Processing", Andreas Witt, Georg Rehm, Thomas Schmidt, Khalid Choukri, Lou Burnard (eds.).

Stoll, S., Bickel, B. (2009). How deep are differences in referential density? In Lieven, E., J. Guo, N. Budwig, S. Ervin-Tripp, K. Nakamura, & Ş. Özçalişkan (eds.) Crosslinguistic Approaches to the Psychology of Language: Research in the Traditions of Dan Slobin, 543 – 555. London: Psychology Press.

Wittenburg, P., Wynne, M., Váradi, T., Krauwer, S., Koskenniemi, K. (2008). CLARIN: Common Language Resources and Technology Infrastructure. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair) Khalid Choukri, editor, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, 2008. European Language Resources Association (ELRA).

Somali Spelling Corrector and Morphological Analyzer

Nikki Adams, Michael Maxwell
University of Maryland, College Park MD 20742 USA
nadams@umd.edu, mmaxwell@umd.edu

Abstract

For any language, a basic requirement for a natural language processing capability is the availability of an electronic dictionary in a form which other NLP tools can use. For all but isolating (or nearly isolating) languages, another basic requirement is the capability to both generate and analyze all inflected forms. This second requirement is usually fulfilled by a finite state transducer that uses the morphological (and perhaps phonological) rules of the written language, together with the dictionary. A third need, for languages where there is significant variation in spelling, is a spell corrector, which can also be implemented as a finite state transducer. These three resources are mutually supportive: the morphological transducer requires the dictionary, and because of the properties of finite state grammars, it is simple to couple finite state transducers together, giving inflectional lookup of misspelled forms. And testing the parser and spell corrector on a web corpus can supply new words for the dictionary, completing the cycle.

This paper reports on the cleaning of an electronic dictionary for Somali, the construction of a Somali morphological analyzer and a spelling corrector, and their resulting composed form. Somali is an Afro-Asiatic language in the Cushitic sub-family with complex morphology, complex morpho-phonological rules, and an orthography which, though officially standardized, is often not used consistently among speakers of the language.

The electronic dictionary is showing signs of age; in particular, we believe there is need for expansion of its vocabulary to cover modern forms. While we have not as yet implemented dictionary expansion for Somali, we describe similar work in Yemeni and Sudanese Arabic, which could be extended to Somali.

Keywords: morphology, morphophonology, finite-state machines, Somali

1. Introduction

Somali is a language in the Cushitic branch of the Afro-Asiatic family. It is spoken in Somalia as the national language, but also by Somalis in Kenya and other countries to which they have immigrated, with an estimated total number of speakers of just under 15,000,000 according to the Ethnologue (Lewis et al., 2015). It has been written in an official Latin-based alphabet since 1972.

Somali is a tonal language, with pitch accent marking both lexical and grammatical distinctions; however, tone is not written, so it does not play a role in what follows.

Somali has a suffixing (with the exception of a few verbs that also take prefixes) agglutinative morphology. It is complicated enough that it cannot be discussed in detail here; see (Green et al., in preparation) for a book-length discussion. The major word classes are nouns and verbs.

Nouns are inherently of either masculine grammatical gender or feminine grammatical gender, and most can be morphologically marked for singular or plural number, and for definiteness. When singular, they take one set of definiteness markers, and when pluralized, a different set; the plural set for masculine nouns resembles the singular set for feminine nouns, and vice versa. This has been analyzed as gender polarity. It is not clear that this is the correct analysis, but we will ignore this theoretical question here.

Nouns fall into a number of declension classes ((Green et al., in preparation)) based on how they are pluralized and (to a lesser extent) on the forms of the definite suffixes that they take. The realization of plurality ranges from a simple suffix, to reduplication, to tone shifts (unmarked in the orthography, as mentioned above), to other more complicated forms (including plurals borrowed from Arabic). Indefiniteness is unmarked; definite determiners come in a variety of forms, depending on other morphosyntactic features they bear (such as interrogative, possessive, and demonstrative features). Some determiners are suffixes, while others can be linguistically analyzed as independent words; nevertheless, all are written as if they were suffixes. The choice of the determiners is also dependent on the gender and number of the noun.

Verbs stems are derived from verbal roots by zero to two (rarely more) derivational suffixes; these suffixes give rise to causative, middle voice, passive, experiencer, inchoative, factitive and reciprocal forms. Verb stems are further inflected by a suffix marking agreement with the person, number and gender of their subject, plus an additional suffix marking tense, aspect, mood and polarity. Processes of epenthesis, consonant "weakening", and syllable reduction affect the shape of affixes. A small set of verbs undergoes prefixation and changes to the root vowels, and there is a further small set which is simply irregular.

Though lacking computational resources compared to the larger Indo-European languages, Somali is the most well-resourced of the Cushitic languages, having textbooks, dictionaries, and grammars available, as well as a fairly large internet presence. Nevertheless, Somali lacks many computational resources needed to become a well supported language in computational terms. This paper describes an effort to build some of those resources. In particular, we describe the cleaning of an electronic dictionary, and the building of a morphological transducer and a spell corrector, and how they work together. While we have not done vocabulary expansion (updating) for the Somali dictionary, we describe similar efforts for Yemeni and Sudanese Arabic, and show how those efforts could be applied to Somali.

The original purpose for which our resources were developed was to aid English speaking students of Somali, in

particular with respect to dictionary look-up and morphology learning. As students encounter new words, they may not know how to spell them nor be able to parse them, and our tools were built to address these issues. However, these resources can also form the cornerstone to any future development of natural language processing as part of a Basic Language Resource Kit (BLARK) for Somali (Krauwer, 2003; Maegaard et al., 2006). For example, given the morphological complexity of Somali, a morphological analyzer is a prerequisite to machine translation or any kind of syntactic parsing, and would probably also be needed in order to build a named entity tagger. And while our spelling corrector was directed towards correcting the kinds of errors English language learners are apt to make, it appears that native speakers make similar errors, and the methods we use to choose the most likely corrections to English speakers' misspellings could be easily adopted to native speakers' spelling errors.

There are three dialects of Somali (Lewis et al., 2015), with the Northern dialect considered more or less standard (Saeed, 1999). Since the spelling of Somali is somewhat in flux, this dialectal variation affects the way words are written by native speakers. The morphological analyzer discussed here is based on the Northern dialect, but the spelling corrector takes special account of pronunciation variation across dialects.

2. Dictionary Cleaning

A common requirement for most NLP processes is an electronic dictionary of the target language. For some applications, this can be a monolingual dictionary, but a bilingual dictionary in which the second language is a widely used language is needed for many applications (such as machine translation or cross-language information retrieval).

For Somali, there are several print dictionaries, but so far as we know, only one substantial electronic dictionary, (Zorc and Osman, 1993). This was made available to us and to at least one other researcher in XML form for research purposes by the by AECOM (`www.aecom.com`; it contains about 35 thousand entries. A typical entry looks like this:

```
<entry>
  <keyForm type='word' lang='som'
           reg='modern'>
    <term scr='la' orth='native'>
           araaree</term>
  </keyForm>
  <pos>verb</pos>
  <note type='grammar'>v2b=</note>
  <sense>
    <gloss>mediate between fighting
           groups</gloss>
    <gloss>stop a fight</gloss>
  </sense>
</entry>
```

If all the lexical entries had been digitized as consistently as this one, that would have been the end of our dictionary work. They were not. In fact we made over 3500 corrections to this dictionary–that is, we corrected about one in ten of the lexical entries.

Had we needed to make all those corrections by hand (not to mention finding all the errors behind those corrections), we might still be working on this dictionary. Fortunately, this is but one of many electronic dictionaries that we have repaired over the last decade, and we have a substantial infrastructure devoted to finding anomalies (potential errors) and diagnosing and repairing them. This work is described elsewhere (Zajic et al., 2011; Zajic et al., 2012; Zajic et al., 2013; Zajic et al., 2015; Bloodgood and Strauss, 2016); here we provide a synopsis of how the infrastructure has been applied to the task of dictionary repair.

We distinguish "anomalies" from errors: an anomaly is an unusual structure that has been noticed, either by a computational process or by a human scanning through the dictionary; an error is a structure which a human expert has determined is incorrect. Errors range from typos to information which has been mis-tagged (meaning that it is in the wrong field) to fields that are missing or out of order. An example of information that was found in the wrong field of a particular lexical entry of a different dictionary was the word "rare"; this had been mistagged as belonging to a *usage* field, when in fact it belonged in the *gloss* field. (Probably this was the result of an earlier tagging effort which had capitalized on the fact that one of the usage codes was in fact the word "rare"–but at least one instance of this word was part of a definition.) Missing fields have included errors ranging from missing glosses to entire missing lexical pages of entries (which can often be restored from the print dictionary).

Not all anomalies are errors, obviously; in terms of accuracy measures, an anomaly which is not an error would count against a precision measure for error detection. Likewise, some errors are not anomalies, in the sense that the error represents a substantial class of similar ill-formed but incorrect structures. For example, suppose the desired form of a part of speech label is "verb". Alongside this correct label, there may be dozens or even hundreds of labels "Verb" or "V" or "verbal", or any number of other (probably) incorrect labels. But precisely because these incorrect labels may appear frequently, they cannot be classed as anomalies; they are too frequent.

Since in a typical dictionary there may be thousands of errors (the Somali dictionary was by no means at the high end of errorfulness), it is labor intensive to find them with manual search methods; indeed, the class of "error" is open-ended, so that it becomes impossible to search for particular errors. Rather the dictionary editor would be faced with reading the lexical entries one by one, looking for oddities. Our group has therefore implemented a computer program that searches for anomalies and brings them to the attention of a human expert; we call this program "ADALT", for Automatic Detection of Anomalies in Lexical Text. This is an unsupervised machine learning program that learns the normal structures of a particular electronic dictionary, and then finds anomalies–deviations for statistical normality. ADALT employs several heuristics for this; for example, it compiles a list of all sequences of XML tags up to a certain length, and then reports sequences which are rare.[1]

[1] ADALT does not know in any sense what the tags mean. A

It is also possible to detect content anomalies. An example is a headword which contains parentheses in a dictionary in which headwords usually consist of alphabetic characters and spaces. In the particular dictionary in which we found this sort of anomaly, the parentheses served as a space saving measure in the original print dictionary–much as if the American English spelling "color" and the British spelling "colour" had been conflated to "colo(u)r". While not an error in a print dictionary, this indeed constituted an error in the electronic dictionary, since it effectively prevented lookup of either "color" or "colour".

Another technique we are exploring is to create models of the transiteration between two fields, for instance a field representing the print form of a word and a field representing its pronunciation. Such a model could be as simple as an average length ratio plus the variance of that length ratio, or as complicated as a learned mapping between the two fields–the latter using tools such as Phonetisaurus (Novak et al., 2012) to automatically induce the mapping. Again, ADALT reports anomalies–outliers in the statistical sense.

ADALT creates a list of anomalies ordered by a measure of their abnormality.[2] The usual workflow is to start at the top of such a list (i.e. with the most anomalous cases) and work down the list until the precision (the ratio of errors to anomalies) decreases enough to make it unlikely that many true errors remain. This is of course a subjective method, but given that the real constraint on dictionary repair is the time available to human editors, it is a practical method.

Our team has also created a special purpose GUI-based editor for editing dictionaries: VELMA (a visual environment for exploring and editing electronic dictionaries, (Zajic et al., 2013)). This editor has a number of advantages over ordinary programmers' editors or XML editors. First, it is coupled to ADALT, so that the lists output by ADALT generate a workflow for people. Second, while editing actions may be performed by dragging and dropping, copy-paste, or typing in information, under the hood the editor records the user's actions as steps in a specialized programming language designed to manipulate XML nodes, which we have dubbed "Dictionary Manipulation Language" (DML). The commands in this language are visible to the user, who can add comments to them so that the record preserves not only *what* was done, but *why* it was done. Furthermore, most actions are independent, in the sense that editing one node in the XML does not affect nearby nodes. This in turn means that a set of actions performed on a node can be undone months later without affecting subsequent actions. We find this to be a much easier way of rolling back erroneous transactions than trying to use a version control system to undo changes.

As mentioned, because ADALT looks for anomalies, it is good at finding rare errors; it is not good at finding frequent errors. Of course as a result of initial mistagging, or subsequent batch editing with traditional tools (such as perl), some errors are frequent. But humans tend to be good at finding these frequent errors, precisely because it is not usually necessary to look through a huge number of entries to find them. But even if ADALT is not good at finding frequent errors, we would like to be able to use VELMA to correct them. In current work, we are adding this capability to VELMA, and some initial capabilities were available when we repaired the Somali dictionary. VELMA already allows the user to correct frequent errors one by one. What it has lacked is the capability to aid the user in creating a template of DML code to capture a set of errors having a common structure, and execute that code on all XML nodes found in such a structure. The method we are developing allows the user to select a set of nodes in an example error as an XPath, to remove unnecessary parts of that XPath (frequently the text content), visualize the set of nodes throughout the dictionary that the template will match and adjust the XPath if necessary, and then apply the change to all matching nodes.

We have employed ADALT and VELMA to edit dictionaries that were "born" as print dictionaries and subsequently digitized, where the digitization process introduced errors, and where subsequent attempts to repair errors with traditional tools such as perl and text editors have often introduced additional errors.[3] While some tools for building digital dictionaries from the ground up, such as Tshwanelex (`http://tshwanedje.com/tshwanelex/`) and SIL's FieldWorks Language Explorer (FLEx, `http://fieldworks.sil.org/flex/`) are intended to make such errors less likely (particularly as compared with earlier tools used by field linguists, such as SIL's earlier Shoebox and Toolbox), errors can still happen with these sophisticated tools through inattention, or through inconsistency during the life of a dictionary project. In fact, the longer a dictionary project, and the more lexicographers are involved, the more likely such errors are. We believe that ADALT and VELMA may be useful in such projects as well.

Finally, we suspect there is a role for tools like ADALT and VELMA in crowd-built dictionaries, such as the Wiktionaries (`https://en.wiktionary.org/`), particulary since the multiple contributers are unlikely to have extensive experience or training in lexicography.

3. Morphological Transducer

While machine learning has replaced hand crafted language technologies in most areas, for languages with significant morphological complexity machine learned grammars are still not as accurate as good hand crafted morphological parsers.[4] Machine learning sufficient to provide full morphological analysis of a language with a complex morphol-

human lexicographer might instead hypothesize what the tags in a particular dictionary mean in terms of typical lexical entries in a variety of other dictionaries, and use that knowledge to find errors. The equivalent in AI terms would be supervised machine learning; we have not explored that.

[2]Of course not all classes of anomalies are comparable, because the metrics used for different heuristics are not necessarily commensurate. Some art lies in deciding how to weight the results of different heuristics, or display the results of different heuristics to the user in different lists.

[3]The misplacement of "rare" as a value of a usage field, instead of as a gloss, is such an example.

[4]Technically, we are talking about morphological transducers, that is software (typically implemented as a finite state transducer) which can both analyze a given surface form into its constitutent

ogy like Somali would require substantial tagged training data, and as is often the case for low-resource languages, there is no Somali training corpus available for this purpose. The difficulty in acquiring training data is exacerbated by the morphological complexity, in particular by the partly agglutinating morphology of Somali, since languages with significant morphology have more word forms and thus the data needed for machine learning is greater than for langauges with typical fusional morphology.

At the same time, hand written parsers can be difficult to debug, and can become obsolete with changes to the underlying technology. For some languages, dialectal variability and the different ways of writing found in social media can also be problematic. For all these reasons, the grammar used in a morphological transducer must be well documented, so it can be modified later. An approach which we used in building the Somali morphological parser is to write a descriptive grammar (Green et al., in preparation) in tandem with the parser, so that the human-readable description of the grammar can directly inform the writing of the computer-readable former grammar rules.

In fact, our approach (described in Maxwell and David (2008; David and Maxwell (2008; Maxwell (2013)) uses Literate Programming to embed particular formal grammar rules (written in a declarative XML formalism) alongside the human readable description of the grammar processes. This makes it easy to compare the two formats and detect differences. Furthermore, the same examples which are intended to explicate the grammar to the human reader can be automatically extracted and used as test cases for the parser. In this way, even morphological constructions which would be rare in corpora, but which are described in the grammar, can be verified.

We write our morphological grammar in a declarative XML-based representation, which allows us to use a format very similar to a traditional linguistic format (Maxwell, 2012). In particular, the format allows for both "ordinary" morphological affixation, and for morphological processes such as reduplication. The format also allows for (morpho-)phonological rules, which can be used to change underlying forms into surface forms by phonological processes such as assimilation, dissimilation, epenthesis and deletion, etc. These rules are treated as applying in series, so that each rule successively modifies (often vacuously) the form output by the previous rule. Our linguists find this way of modeling grammars, which is very much like a traditional linguistic view of morphology and phonology, to be very easy to work with. Notice that a linguist using this system can model allomorphy in either of two ways: as different listed allomorphs, either of affixes or of lexemes; or by using a single underlying form, and deriving the allomorphs by phonological rules.

This declarative XML-based representation of the grammar is automatically converted to the programming language of the Stuttgart Finite State Toolkit (SFST, Schmid (2005)),[5] and SFST is then used to "compile" the result into an operating morphological transducer.

During the development of the Somali parser, we also developed a debugger tool (Maxwell, 2015). This tool proved useful when (not if!) the reason for a non-parse of a Somali word was not readily apparent. In particular, the debugging mode was used to visualize the operation of the many phonological rules which our analysis of Somali employed. The user proposes an expected parse–an underlying form; the debugger first informs the linguist if the proposed underlying form violates some constraint, e.g. if two suffixes have conflicting morphosyntactic features. Assuming the underlying form is allowable, the debugger then displays the result of applying each phonological rule (of which the Somali grammar has about 20) in sequence. The user can then see the rule in the derivation which results in an incorrect surface form, whether by inappropriate application of a phonological rule, or by non-application of a phonological rule which the user expected would apply. Such errors may be the result of an inappropriate formulation of the rule (often the phonological environment in which the rule will apply). Errors may also arise due to incorrect ordering of the rule with respect to some other rule, so that the input to the rule was not what the linguist expected.

Somali is primarily suffixing, but also has prefixes. Both are easily modeled in our XML formalism. Morphological rules are of course always specified as being associated with one or more parts of speech, while the general phonological rules were by default broadly applied, but can also be restricted by part of speech or other lexical features.[6]

In addition to a grammar, a morphological parser requires a lexicon. Converting a dictionary intended for human users into one that will serve as an input to a morphological parser is a task in itself. The complete lexicon as represented in the Zorc and Osman (1993) dictionary was broken into sublexicons by part of speech and converted into the SFST format.

Included in the lexicons for the major parts of speech were feature specifications for each word, where the features were taken from the electronic dictionary's entry. For ex-

morphemes, and create surface forms from an appropriate set of morphemes. Nevertheless, we will frequently refer in this paper to such software as a morphological parser or a morphological analyzer.

[5] A reviewer asked why we chose SFST. While a full answer would require more space than we have available, the short answer is that we standardized on SFST because at the time it was the only open source FST tool that provided serial rules. (Our XML-based grammars provide serial rules, which is what rule-based morphology and phonology accounts have relied on for decades. It would be very hard to convert serial rules in our XML format into the two-level rules provided by some other FST tools. The Xerox xfst/lexc product, which also provides serial rules, was also available, but only under a commercial license. Since that time, Foma ((Hulden, 2009), which provides most of the xfst/lexc functionality, has become available, and we believe it would be possible to target Foma with the our converter from XML, should there be a compelling reason to do so.

[6] An example from a familiar language is diphthongization of the stem vowel in languages like Spanish, e.g. the Spanish verb *poder* "to be able" has allomorphs *pod* and *pued*. Some dictionaries provide both forms, but an alternative is to apply a phonological rule which converts /o/ to /ue/ in the relevant context, where the context includes the lexical rule feature that allows this rule to apply. This rule-based approach to Spanish diphthongization follows the analysis in Harris (1977).

ample, the noun class specification was taken from the dictionary entry and represented as a feature associated with that noun head-word. This is an area which is particularly tied to dictionary repair, as the information for nominal declension classes was provided in such a variable way that it was difficult to extract it for correct morphological generation until we regularized the notation in the process of cleaning the dictionary. (A side benefit of this dictionary repair was to make the electronic dictionary easier for human users to understand.)

Additionally, underlying forms of stems were represented where we could reconstruct that information. For example, Somali has words whose underlying forms end with "m", but Somali has a general phonological rule that changes all syllable-final "m"s to "n"; the "m" is only realized when a vowel-initial affix is added. This was represented as a surface (upper) "n" mapping to an underlying (lower) "m". A phonological rule that changes all syllable-final "m"s to "n" applies. An example is shown below, where the headword is also represented as being underlyingly geminate ("mm"; whitespace added to fit margins).

```
nishaabtan:m<>:m<InflClass>:<>
<NClass2>:<><Gender>:<> <masculine>:<>
```

To handle morphophonology, allomorphs were listed according to their environments (including a left and/or right context). In some cases, natural classes were proposed that grouped phonemes (graphemes, really) according to some shared phonological feature they represented (e.g. voiceless fricatives) to more efficiently make multiple references to an environment that frequently determined affix allomorphy. Reduplication was also treated, as Somali uses this process both for the pluralization of certain nouns and for emphasizing the relative value of an attribute in attributive adjectives.

Where the same form could represent multiple feature values, but where Somali would make a distinction in those features elsewhere in the language, affixes were listed separately. For example, a morpheme may be identical in the first person singular and the third person singular masculine, but as Somali does make this distinction elsewhere (in its pronouns, for example), these string-identical affixes were represented separately and each marked with their own set of features. Morphosyntactic features, in turn, could be used to enforce feature agreement across inflected forms that contained multiple affixes. For example, the affix for the possessive marker used for a first person singular possessor of a third person singular feminine possessum was marked with the features third person singular feminine. This thus prevents this suffix from attaching to nouns marked with any feature that disagrees with that specification, i.e. nouns marked as masculine, as plural, or with person features other than third person.

Affixes were then combined by grouping them into slots and then specifying the order in which slots can combine. An example is shown below, illustrating an ambiguous parse of the stem + affixes *keen+ay++ey* using a morphological template of headword+Aspect+Person+Tense (white space has been added to allow this example to fit):

```
keenayey
[keen]<Verb><-PAST.PROG><-3.SG.M>
<-PAST.3.SG>
[keen]<Verb><-PAST.PROG><-1.SG>
<-PAST.1.SG>
```

As mentioned, phonological rules were also included which reduced the amount of allomorphy that needed to be handled as suppletive allomorphs. Phonological rules, by default, apply broadly and were primarily used to account for phenomena that happened either across multiple parts of speech or which applied to stems as opposed to affixes; where the phonology was restricted to affixes with a well-defined environment, this was represented by designating allomorphs of the same affix. An example of a parse where a phonological rule applied is shown below, where the underlying string of affixes are *gacam++eed*.[7] A phonological rule deleting the middle vowel in a sequence of three syllables is then applied to derive the surface form.

```
gacmeed
[gacan]<Noun><F.SG><ATTRIB.ASSOC>
```

A reviewer pointed out that another morphological analyzer has been developed by Giellatekno, the Center for Saami Language Technology, and is available as open source from `https://victorio.uit.no/langtech/trunk/langs/som`. We were not aware of that tool until the reviewer brought it to our attention, indeed it is not listed on the Giellatekno website (`http://giellatekno.uit.no`. This coincidentally brings out the importance of conferences like LREC–we imagine the developers at Giellatekno likewise did not know about our Somali parser. At any rate, we plan to download their parser and compare its performance with ours, but have not had the time to do so as this paper goes to print. We do note in passing that their noun dictionary is smaller than ours (which, as noted above, is an electronic form of the the published dictionary (Zorc and Osman, 1993). Oddly, their verb dictionary has more entries than ours; we speculate that this is because their dictionary treats alternative verbal stems as separate lexical entries.

4. Spelling Corrector

The spelling corrector was implemented in OpenFST as a weighted finite-state transducer. Its output can be filtered on the electronic Somali-English dictionary (the lexicon filter) such that query results map to headwords in the dictionary. However, given that we have a morphological analyzer implemented as a finite state transducer, we chose instead to compose the spell checker with the morphological analyzer. This is easily done by dumping the morphological analyzer (built with the Stuttgart Finite State Transducer) to a table format, and recompiling that into OpenFST for composition with the spelling corrector.

Allowable edit operations were determined by examining spelling variation in textbooks and on the internet, as well

[7] A null affix labeled for gender was introduced so that a parse always represents the gender of a lexeme. This affix agrees with the features specified on the headword.

as with consultation from students and a teacher of Somali. The weights were assigned by a linguist familiar with these spelling variations, who had an approximate sense of their frequency, and who understood basic phonological principles. In particular, the latter was important because the spelling corrector was designed to accommodate not only frequent misspellings, but also likely mis-hearings, with particular focus on native English-speaking students of Somali.

Allowable edits, in the form of input-output pairs and the cost (weight) for performing that edit, were created to address several likely sources of mismatch between a query and the headword. First, frequent spelling variations were addressed. As mentioned, though the orthography has been officially standardized since 1972 ((Biber and Hared, 1991)), due to instability in the country and the disruption of the educational system, this standard is often unknown to speakers of the language. Common spelling variations for Somali involve using "e" for "a" and vice versa, particularly in certain affixes, variation in the representation of consonants as single or geminate (e.g. "d" versus "dd"), and variation in vowel length representation (e.g. "a" versus "aa").

Secondly, spelling variations due to dialectal differences were also accounted for. These included "dh" and "r" mappings, as well as "kh" and "q" mappings, as there are words where which member of the pair is used differs by dialect.

Finally, but critically, the edits also accounted for mishearings, which can be a source of misspelling for students who have heard a word and are trying to find it in a dictionary. Some of these input-output edit pairs overlap with those in the first two conditions mentioned. For example, long and short vowels are commonly confused in the spelling of Somali words, and it is also the case that English-speaking students of Somali may have a hard time hearing the distinction between long and short vowels, as this is not a distinction English makes. Thus, the inability to hear the distinction may cause the student to misspell their query. Other edit pairs, however, are more representative of the inability of a non-native hearer of Somali to distinguish between certain sounds, such as "c" (a pharyngeal fricative) and "kh" (a voiceless uvular fricative). Finally, certain sounds are simply phonetically similar, so that even where English makes the distinction, it is still possible for a native English speaker to mis-hear. An example of this would be an "s" /s/ and "f" /f/ pair.

Spelling variations due to differences in word boundaries were not accounted for here.

5. Future Work

The results reported here represent work in progress. Clearly there is a need for further evaluation, both corpus-based, and comparisons with other tools. Some corpus-based evaluation is mentioned above, in particular our preliminary results regarding variations in spelling by native speakers. By way of comparison with other tools, we mentioned above that a reviewer point out the Giellatekno parser, which we intend to obtain and try out.

One can also find a Hunspell dictionary for Somali (e.g. `http://extensions.services.openoffice.org/en/project/somali-language-spell-checker`), although it does not appear to exist in the main Hunspell repository (`https://hunspell.github.io`). Since one of the goals of our parser is spell correction, it should be informative to compare performance of our spell corrector with that of Hunspell, and incorporate any improvements such a comparison might suggest. Comparing spell correctors is of course not a trivial task; since finding the most likely corrections to a misspelled word, and ordering those corrections in some logical way, is a statistical matter, one cannot simply say that the top-most correction in system A ought to be the top-most correction in system B. Ideally, one would use a corpus of spelling errors where one knows for each misspelled word what the correct spelling is. To our knowledge, such a corpus does not exist for Somali. Alternatively, one might compare the ranked set of corrections (or the top N of those ranked corrections) between the two systems, using some rank-sensitive score (such as inverse rank). Significant differences (for example, a suggested correction in the top 3 of system A which does not appear at all in the top 10 of system B) must then be compared for plausability on a case-by-case basis. This is not an easy task.

Other work which we would like to do includes dictionary augmentation. The dictionary we used for the morphological parser (Zorc and Osman, 1993) was compiled over twenty years ago. The country of Somalia has been in turmoil for much of that time, and many words have entered the language as loans (particularly from Arabic) and perhaps as coinings. An informal sampling of words from a 600,000 word corpus of news articles which do not parse, even with spell correction, shows that our coverage would be better if we expanded the dictionary to include new vocabulary, and also if we did back-off search to Arabic. We report here on vocabulary expansion work that we have done for Yemeni and Sudanese varieties of Arabic; very preliminary work has shown encouraging results for similar backoff from Somali to Arabic. In addition, we briefly report here on cross-language detection of loan words, using as our example French vocabulary found in North African Arabic.

The Sudanese and Yemeni vocabulary expansion was used to add vocabulary found in modern Sudanese and Yemeni, but which was not already found in our dictionaries of those two languages. The method we employed relied on being able to find Sudanese- and Yemeni-specific websites. For this, we simply searched for websites containing vocabulary which we already knew to be specific to these two languages from previous work, based on published dictionaries of these two varieties (Tamis and Persson, 2013; Qafisheh, 1999), and work by Peter Behnstedt on the dialects of Yemen. Once we had such sites, we explored them for vocabulary that could not be accounted for on the basis of a morphological parser and our known lexemes.

From this unknown vocabulary, we selected the more common words (after stripping known clitics and affixes), and presented these to native speakers of these varieties for evaluation. For each word, the native speakers were asked to decide whether it was a misspelling of some already

known word, or a transliteration of some foreign term (often a place name or personal name), or a previously unattested word. For the latter, they were asked to pick its part of speech and certain other grammatical information, including irregular forms (such as "broken" plurals), and provide a gloss or definition in English. Finally, we asked the native speakers to choose one or two example sentences from among the examples our method had culled, and translate these into English.

A linguist who knew Arabic, but was not a Yemeni or Sudanese speaker, took over from there, editing the glosses, definitions and translations of example sentences. The result was a supplemental lexicon of dialectal vocabulary which had not appeared in earlier dictionaries, and which was common on-line, at least in the websites we sampled.

A similar method could be employed to create a supplementary lexicon for Somali vocabulary.

In addition, we already know that many of the words we see in Somali texts which we cannot parse to lexical entries in our existing Somali dictionary are loans from Arabic. The use of loanwords is of course common in developing languages (not to mention many developed languages). Loanwords are often under-represented in dictionaries despite their wide usage because they are felt not to be "real" words of the language, or because the lexicographer feels it would be better to use native vocabulary. In other work, we have developed experimental techniques for find French loans in north African Arabic, particulary Moroccan Arabic. These loans are generally not distinguished in written Arabic, that is they are written in Arabic script, and usually take Arabic (not French) prefixes and suffixes. This presents a problem for Arabic language learners who do not happen to know French, as well as an impediment to NLP in north African Arabic. The technique we have developed takes a French dictionary and transliterates the headwords (which are the forms to which Arabic affixes are attached) into Arabic script, based on previously seen examples, and then uses these French words written as if they were Arabic stems as part of the vocabulary of a Moroccan Arabic morphological transducer.

A similar technology could be used to find Arabic loanwords in Somali. Arabic and Somali are written in different scripts, so a transliterator must be developed from the Arabic script to the Latin script based on known examples (of which we already have a sampling, including but not limited to words tagged in Zorc and Osman (1993) as having Arabic etymology. The resulting latinized Arabic words would then be supplied as additional input vocabulary to the morphological rules we have constructed for Somali. The result would be a back-off morphological analyzer for Somali, which would offer possible analyses of Arabic loanwords in Somali.

Of course loanwords can suffer semantic drift. But knowing the etymology is often an important clue–indeed, perhaps the only clue–to the meaning of loanwords, particular recent ones.

6. Conclusion

We have described technology which we developed to aid English speakers learning Somali, but which has clear use for natural language processing of Somali. In addition, we describe techniques for vocabulary expansion and loanword discovery which we have developed for varieties of Arabic, but which could also be employed for Somali. Indeed, we feel that the importance of the techniques we have described here lies not only in their relevance to Somali and Arabic dialects, but that they will form a necessary part of the toolkit for many low density languages of the world. In that sense, the lesson is not that we did this for Somali, but that anyone can do this for any less resourced language.

7. Bibliographical References

Biber, D. and Hared, M. (1991). Literacy in somali: Linguistic consequences. *Annual Review of Applied Linguistics*, 12:260–282, 3.

Bloodgood, M. and Strauss, B. (2016). Data cleaning for XML electronic dictionaries via statistical anomaly detection. In *Proceedings of the Tenth IEEE International Conference on Semantic Computing (ICSC 2016)*, Laguna Hills, California.

David, A. and Maxwell, M. (2008). Joint grammar development by linguists and computer scientists. In *IJCNLP*, pages 27–34. The Association for Computer Linguistics.

Green, C. R., Morrison, M. E., and Adams, N. B. (in preparation). *A Grammar of Common Somali*. de Gruyter, Berlin.

Harris, J. W. (1977). Remarks on diphthongization in spanish. *Lingua*, 41:261–305.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, page 2932. Association for Computational Linguistics.

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" (SPECOM 2003)*.

Lewis, M. P., Simons, G. F., and Fennig, C. D. (2015). *Ethnologue: Languages of the World*. SIL International, Dallas, eighteenth edition.

Maegaard, B., Krauwer, S., Choukri, K., and Jørgensen, L. (2006). The BLARK concept and BLARK for Arabic. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 773–778. European Language Resources Distribution Agency.

Maxwell, M. and David, A. (2008). Interoperable grammars. In Jonathan Webster, et al., editors, *First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, pages 155–162, Hong Kong.

Maxwell, M. (2012). Electronic grammars and reproducible research. In Sebastian Nordoff et al., editors, *Electronic Grammaticography*, pages 207–235. University of Hawaii Press.

Maxwell, M. (2013). A system for archivable grammar documentation. In Cerstin Mahlow et al., editors, *Systems and Frameworks for Computational Morphology: Proceedings of the Third International Workshop on Systems and Frameworks for Computational Morphology*, pages 72–91. Springer.

Maxwell, M. (2015). Grammar debugging. In Cerstin Mahlow et al., editors, *Systems and Frameworks for Computational Morphology: Proceedings of the Fourth International Workshop on Systems and Frameworks for Computational Morphology*, pages 166–183. Springer.

Novak, J. R., Minematsu, N., and Hirose, K. (2012). WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 45–49, DonostiaSan Sebastian.

Qafisheh, H. A. (1999). *NTC's Yemeni Arabic-English Dictionary: A compact dictionary of the contemporary Arabic of Yemen*. McGraw-Hill, Columbus, Ohio.

Saeed, J. (1999). *Somali*. Number 10 in [London Oriental and African Language Library. John Benjamins.

Schmid, H. (2005). A programming language for finite state transducers. In Anssi Yli-Jyrä, et al., editors, *Proceedings of the 5th International Workshop on Finite State Methods in Natural Language Processing (FSMNLP 2005)*.

Tamis, R. and Persson, J. (2013). *Sudanese Arabic–English / English–Sudanese Arabic: A Concise Dictionary*. SIL International Publications in Linguistics. SIL International, Dallas.

Zajic, D. M., Maxwell, M., Doermann, D. S., Rodrigues, P., and Bloodgood, M. (2011). Correcting errors in digital lexicographic resources using a dictionary manipulation language. In *Proceedings of Electronic Lexicography in the 21st Century (eLex)*, volume abs/1410.7787.

Zajic, D. M., Doermann, D. S., Bloodgood, M., Rodrigues, P., Ye, P., Foley, D., and Zotkina, E. (2012). A hybrid system for error detection in electronic dictionaries. Technical report, Center for Advanced Study of Language, University of Maryland.

Zajic, D. M., Doermann, D. S., Rodrigues, P., Ye, P., and Zotkina, E. (2013). Faster, more accurate repair of electronic dictionaries. Technical report, Center for Advanced Study of Language, University of Maryland.

Zajic, D. M., Bloodgood, M., Strauss, B., and Zotkina, E. (2015). Faster, more thorough error detection in electronic dictionaries. Technical report, Center for Advanced Study of Language, University of Maryland.

Zorc, R. D. and Osman, M. M. (1993). *Somali-English dictionary with English index*. Dunwoody Press, Kensington, MD.

Reprinting scholarly works as e-books for less-resourced languages

Delyth Prys, Mared Roberts, Gruffudd Prys
Language Technologies Unit, Bangor University, Bangor, Wales, UK
E-mail: {d.prys, mared.roberts, g.prys}@bangor.ac.uk

Abstract

This paper on the DECHE project for Digitization, E-publishing, and Electronic Corpus reports on a project undertaken for the Coleg Cymraeg Cenedlaethol, the virtual Welsh-medium College for universities in Wales. The DECHE project's aim is to digitize out of print scholarly works across multiple disciplines and help create a library of e-books available to Welsh-speaking academics and students. The context of e-book publication in Wales, and the digitization and e-book production process is described, together with the software tools used. The criteria for selecting a shortlist of books for inclusion are given, as are the types of books chosen. Attitudes and take-up of the students surveyed are also discussed, as are the dissemination of the resulting e-books, and statistics of use. This takes place in the context of the increasing popularity of e-books for education, including at university level, and their value for less-resourced language communities because of the lower production and distribution costs, and their contribution to raising the image and status of those languages as fit for purpose in a digital age.

Keywords: E-books, Digitization, Less-resourced languages, Welsh

1. Background

Producing commercially viable books in languages that do not have the readership numbers of larger languages can be challenging. According to the 2011 Census, the number of Welsh speaker in Wales was 562,016, or 19% of the population (StatsWales, 2011), thus marking Welsh as a small language in terms of economic returns for the publishing industry. In some markets, subsidies and grants are used to help publish titles that would not otherwise be commercially sound ventures. For example, in Ireland, approximately one-third of the estimated total yearly income of Irish-language publishers, of less than US $2 million, came in the form of grants (Finkelstein and McCleery, 2005). A similar position exists in Wales, where the Welsh Book Council, funded by the Welsh Government, is responsible for providing financial subsidies and various support services to publishers publishing non-academic books in Wales in English or Welsh.

These subsidies and grants however only usually cover a book's first print-run. A common problem for less-resourced languages is that it is rare for a book that becomes out of print to receive a further grant to cover reprinting costs. This became an acute problem for Welsh-medium academic books in Wales with the establishment of the Coleg Cymraeg Cenedlaethol in 2011. The Coleg is a virtual college with branches in universities across Wales, dedicated to furthering teaching and research through the medium of Welsh (Andrews and Prys, 2016). By driving a rapid expansion of Welsh-medium university education across all subject areas, the establishment of the Coleg has substantially increased the demand for scholarly and academic books in Welsh at university level. There are currently 1,000 Welsh-medium university courses on offer, with over 100 new Welsh-medium lecturers appointed in diverse subject areas, including the Arts, Health, Social Sciences and Science (Coleg Cymraeg Cenedlaethol, 2016). In order to address the need for teaching materials in Welsh at university level, the Coleg has developed a grant programme to fund the creation of academic resources in Welsh. This has led to the establishment of a comprehensive portfolio of Welsh language resources in the various subject areas, contributed to and shared between the participating universities. Most of these resources are available on-line at the Coleg's e-learning platform, Y Porth, and its digital library (Llyfrgell Adnoddau, n.d.), and include videos, lectures, presentations and other teaching materials as well as original e-publications. Despite the wealth of new resources appearing in Welsh, it was also recognised that there was a real need for existing out of print Welsh-language academic books to be re-published as these were often relevant and sometimes seminal works in a number of disciplines. These titles' lack of availability resulted in waiting lists for library copies and the circulating of photocopied chapters amongst students and academics – a situation that was far from ideal.

2. E-publishing as a Solution

A solution to this problem was offered in 2012 with the suggestion that relevant out-of-print books be digitized by the Coleg and re-published in electronic format. A grant was sought by the Language Technologies Unit, Bangor University for this purpose and awarded by the Coleg, leading to the establishment of the DECHE project (DECHE being an acronym for "Digido, E-gyhoeddi a Chorpws Electronig" or Digitization, E-publishing and Electronic Corpus) at Bangor University. One factor in the application's success was that its aims were well aligned with Welsh Government policy, whose Welsh Language Strategy for 2012-2017, *A living language: a language for living*, stated that reading Welsh should be encouraged and that "In doing so, it is essential that the Welsh language keeps up with current developments, such as by ensuring that a wide range of e-books are available across all contemporary devices" (Welsh Government, 2012).

The LTU had already undertaken a survey for the Welsh Book Council on e-publishing in Welsh (Prys, Prys, Jones & Chan, 2011), including sections on the international e-book market and a questionnaire on the use of e-books and

e-book reading devices by the Welsh-speaking public. Separate Technical Guidelines were also produced (Language Technologies Unit, 2012) intended to guide small publishing houses in Wales on the process of producing e-books from their authors' manuscripts. This was published on the Welsh Book Council's website as part of Welsh Book Trade Info section, intended as a repository of useful information for the book trade in Wales.

A stipulation of the DECHE project receiving funding was that the resulting e-books, as with all other Coleg resources within the Llyfrgell Adnoddau digital library website, were to be available to all through open access for no charge. As a result, the e-books published to date are available to all for free, as downloads from the Coleg's digital library (Prosiect Digido n.d.). However, for these titles to be distributed for free, a number of copyright issues would first have to be resolved.

3. Copyright and Digitization

Copyright issues are one of the major challenges in digitizing out of print works that are still within the copyright period. In the DECHE project, a small fee was offered for permission to publish as e-books and include the text in the DECHE corpus of academic Welsh, and some publishers were also paid to help in dealing with matters such as tracing copyright holders. Where publishing houses had ceased to exist, Coleg staff undertook this work themselves. Clearance was obtained before any digitization tasks were undertaken. However, in a number of cases, this process of clearing the copyright, although essential, took more time than the digitization.

The work of scanning the paper originals and preparing the page images was undertaken at the National Library of Wales, which has extensive expertise in digitizing content, having written its first digitization strategy in 2008, and subsequently updated it in 2012 (Digital Preservation Strategy, 2012).

Informed by the Welsh Book Council survey, it was decided that the project would create three electronic versions of each book to be published: EPUB (which can be used by most e-reading devices), MOBI (which is the format needed by the popular Kindle e-readers), and a PDF version for printing copies for personal use.

The digitized files were received from the NLW as images in TIFF files. These images were then converted from images to electronic text files at the Language Technologies Unit. Initially the Tesseract-ocr (n.d.) open source optical character recognition engine was used for character recognition, having first been trained specifically for Welsh. The original images, although excellent reproductions, were only as good as the books being scanned, and if the original books had been printed on poor quality paper, or the font was difficult to read, these issues were also present in the scanned images. Because of this, character error rates varied widely from one or two errors per paragraph in the books with the best print quality to almost no sentence without at least one error in the books with the worst print quality. The most common errors were confusion between letters such as 'rn' instead of 'm'; confusion between 'c', 'o' and 'e' and the numeral '0'; and confusion between 'i', 'l', and the numeral '1'. Surprisingly, accented characters were not as great an issue as anticipated, and common errors such as the numeral '6' instead of 'ô' were dealt with at an early stage by means of an additional piece of code within Tesseract. Infrequently, words were run together, with the spaces between them being omitted.

Once the image files had been converted to electronic text files, proofreading with a spellchecker was then carried out, using the Welsh Microsoft Office spellchecker. For books with a large number of character errors, there were too many errors to proceed, and staff therefore had to break the texts up into smaller chunks, sometimes of as few as three pages at a time, for processing.

The staff then manually proofread and corrected the text document, adding html tags where paragraph and font styles were required, before transferring the file to Sigil (n.d.), an open source multi-platform e-book editor designed to edit books in EPUB format.

The proprietary OCR program, OmniPage Ultimate (n.d.), was purchased in 2014, initially for its ability to output files in EPUB format. An unexpected benefit was that its OCR feature, although notionally language independent, gave better results than Tesseract for character recognition accuracy, despite Tesseract having been trained on Welsh. The Welsh Microsoft Office spellchecking dictionary was then integrated with OmniPage, and this was found to work well, with the ability to process larger chunks, up to 150 pages each, at a time. However, accented characters proved to be less accurately displayed than with Tesseract, with characters either showing with their accents missing, or not showing at all. In Welsh there are many word forms where only an accented character denotes a difference of meaning, for example 'mor' is the comparative form 'as', whereas 'môr' is the Welsh word for 'sea'; and 'a' means 'and', whereas 'â' means 'with'. Problems with accented characters were therefore left until the proofreading stage in Sigil to be rectified. The other advantages of OmniPage over Tesseract however meant that the project switched from using Tesseract to OmniPage, with a modest increase in processing speed.

Quotations in other languages, e.g. English, French or Latin, or even Middle Welsh (the medieval form of the language), were more problematic for both programs, but was dealt with by careful human proofreading as it was only a feature of a small subset of texts, and did not occur often enough to merit a technical solution.

After processing the TIFF files in OmniPage, the EPUB files were cleaned with an additional script, as OmniPage's html markup creates many redundant CSS classes. The e-books were then manually processed within Sigil and this included editing their structure, markup, table of content and additional proofreading. The Hunspell Welsh language spellchecker, previously produced by the LTU, was integrated within Sigil to provide an additional layer of automatic proofreading by highlighting further errors in the texts. It is estimated that this helped identify an additional 0.5 or 1 minor errors per thousand words. The next stage, whether Tesseract or OmniPage had been used, was to add

the publication's metadata details in Sigil, including author, publication date, and original publisher, along with a new, specially designed book cover, table of contents, images, and any other additional content.

Following this, a draft e-book was produced for testing on different reading devices and platforms, e.g. iPad, Kindle, Adobe Digital Edition, before being provided to a second human proof-reader for additional proofreading. Very occasionally, this proofreading phase would identify a word which although not a misspelling was not what was in the original text. Only careful reading of the text's meaning would highlight such discrepancy, e.g. 'hyd' (length) instead of 'hyn' (this one), or 'dau' (two) and 'dan' (under). During this stage the layout was also checked on various devices, and here problems with spacing, tables, diagrams and illustrations were sometimes identified and corrected.

Once those changes had been incorporated, the e-book was deemed to be ready for publication. Three versions of the final EPUB were created, each with its own unique ISBN. Using Calibre (Goyal et al, 2006-), an open source e-book management program, one of these EPUBs was converted to the MOBI format and another to the PDF format, so that each format possessed a unique ISBN. The three versions were then published online by the Coleg Cymraeg Cenedlaethol.

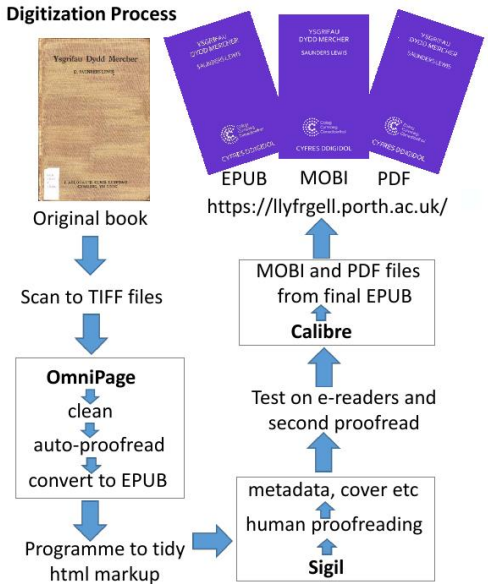

Figure 1: Digitization Process

4. Corpus Generation as a Secondary Benefit

Although the main purpose of this project was to digitize out-of-print books for publication as e-books, the availability of the books in well-formatted, high quality digital editions also enabled the creation of a corpus of Welsh late nineteenth and twentieth century scholarly works. Copyright clearance for the titles' inclusion in the corpus was sought at the same time as clearance was sought for their digitization, thus saving on what can be a time-consuming and costly process. As all the books shared similar characteristics in that they were intended for an academic audience, written in a scholarly style or register, and belonging to the same time period, they formed a cohesive corpus, useful for many types of linguistic research.

Other corpora had previously been created by the Language Technologies Unit, originally for internal use, such as terminology research, the development of machine translation, and the building of new language models for Welsh. The DECHE project provided the opportunity to make the DECHE corpus, as well as other LTU corpora, publically available and searchable on-line through a National Corpus Portal (Prys, Prys & Jones, 2014).

An easy to use interface was created in-house to simplify the transfer of the digital text into the corpus, together with associated metadata. The interface accesses EPUB file that has been published online, validates the importing of the e-book and updates the on-line DECHE corpus accordingly. To date, the corpus contains nearly 2 million words, from around 50 digitized books.

5. Selecting and Prioritizing Books for E-publication

It was acknowledged from the outset that it would not be possible to digitize and e-publish all the candidate books that were deemed worthy of inclusion in this project. A selection process was established to identify and select the most urgent and appropriate books for the Coleg's staff and students. All the Coleg's academic staff were contacted asking for nominations to a shortlist of books considered for inclusion. Books had to satisfy at least one of the following three criteria: that they were 1) core material for Welsh-medium modules, 2) on students' reading lists, or 3) key research material. A fourth criterion which was not explicitly mentioned, but which did come into play when difficult decisions had to be made, was that books that were used in more than one subject area and were relevant for more than one course or university would be given precedence.

A total of 189 titles were nominated, by 41 different academics, spanning a range of over 20 subjects, mostly in the Arts, Humanities and Social Sciences. The lack of nominations in the hard sciences reflects not only the comparative lack of original writing in Welsh on these subjects, but also the fast pace of scientific discovery since the late twentieth century and the fact that academic works in other disciplines tend to remain relevant for longer periods of time. Of the titles nominated, 101 of them were core reading material for modules currently being taught, with the others citing a number of different reasons, including 24 that were named as important for academic research.

Eighteen of the publications were nominated by several lecturers; the most popular book being *Y Traddodiad Rhyddiaith yn yr Oesau Canol* (Bowen ed. 1974), which is a volume of essays on the Welsh medieval prose tradition by 11 eminent contributors.

Philosophy was another subject area for which many nominations were received, as it was a domain that had seen vigorous academic discourse in the Welsh language in the 1960s and 70s. In the intervening years it had largely disappeared as an academic subject through the medium of either Welsh or English from Welsh universities, but it is now being re-established by the Coleg. At least fifteen titles published to date belong to this academic discipline, although some of them are also claimed at least in part by other disciplines. Nine of these titles are the work of the Welsh philosopher J. R. Jones (Dictionary of Welsh Biography, 2009), a body of work that was out of print in its entirety before the advent of the DECHE project.

Books from two important series were included in the project, the *Meddwl Modern* (Modern Mind) series, originally published by Gwasg Gee, and the *Be' Ddwedodd…* (What … Said) series, originally published by the Colegiwm Cymraeg. Both series present synopses of the work of important European thinkers such as Marx, Lenin, Weber, Darwin and Durkheim, and have wide relevance for the fields of History, Philosophy, Social Science, Politics and Science. Their original publishing houses are no longer active, showing the importance of 'rescuing' works of value through digitization and e-publishing.

One very rare title published by the project is *Y Rhyfel Mawr: Apêl at y Bobl* (The Great War: An Appeal to the People) by David Lloyd George, the Welshman, Liberal politician and Prime Minister of the British Government (1916-22). It is a pamphlet containing a speech given by Lloyd George at Bangor in 1915, encouraging his fellow countrymen to support the war. Only one copy of the pamphlet was known to exist, kept at Bangor University's archives, which meant that not even the National Library of Wales had a copy. The digitization of this important document ensured its preservation for the future. In a departure from the usual procedure when creating e-books for this project, images of the original pamphlet's pages were included in the e-book and a copy of the digital images created by Bangor University archive for the project transferred to NLW for additional security.

All but one of the contacted publishers and copyright holders, consented to have their selected books included in the project. The publisher who withheld permission is the main academic publishing house in Wales, holding 46 out of the 189 titles nominated for digitization. The publisher cited their own desire to republish in either electronic or paper versions as the reason for their refusal, and were therefore provided with a list of their nominated 46 titles. It had never been the aim of the project to digitize titles that could be profitably reprinted by the original publishers, and it was encouraging that a publisher thought that it could republish as a commercial venture. However, to date none of these works has reappeared in print or electronic formats.

6. Dissemination

In theory, electronic books should be easier to publish and distribute than their paper counterparts, having no need for physical storage, bookshops or distribution channels. In reality however, some obstacles remain to their publication and dissemination. For example, Welsh language books suffered exclusion from Amazon's Kindle portal for a time, as they said it was an "unrecognised" language (WalesOnLine, 2013). This decision was later reversed following protests and a petition in Wales, but remains an obstacle for other languages. Some Welsh publishers have started selling e-books directly from their own websites, which cuts out the middleman, and can help improve the profitability of what are, essentially small but valuable businesses in a less-resourced language community.

However, in the case of the e-books produced as part of the DECHE project, the Coleg became its own publisher, using its online digital library and e-learning platform, Llyfrgell Adnoddau (Resource Library), to publish the e-books. This did not prove straightforward, as the e-learning platform initially chosen was designed for video audio and PDF dissemination rather than the distribution of e-books. As a result, the EPUB and MOBI downloads, which are better suited for use on mobile devices, were not easy to find, being hidden below more prominent PDF versions. There were also technical problems with the e-learning platform, preventing the distribution of e-books through iTunes U, as was originally intended. It is hoped that these issues will be resolved with the Coleg's proposed move to a new e-learning platform from a different supplier.

In addition to placing the e-books in its own website, the titles are now also being made available through the main library catalogues of Welsh universities as links to the pages in Coleg's digital library. This is important as the e-books are thus mainstreamed and more widely disseminated.

Efforts have also been made to use social media to publicize the new e-books as they appear; messages appear regularly on Twitter, Facebook and other popular media. More traditional leaflets and posters have not been neglected either, and presentations have been made to both students and staff at relevant conferences, seminars and other events.

7. Statistics of Use

Google Analytics was used to count the number of unique visits to the e-books in the Coleg's digital library. Statistics obtained in this manner show the relative popularity of different books, the effect of publicity campaigns, and other general trends.

The most popular e-book by January 2016, according to Google Analytics, was *Crefft y Stori Fer* (Lewis ed., 1949), published on-line in 2014. This was a series of interviews between important Welsh short story writers about their own work, initially broadcast on radio. This had received 200 unique visits so far. Next in popularity were several of the philosopher J. R. Jones' works, with his *Prydeindod* (1966), a treatise on the identity crisis faced by some Welshmen in face of the encroaching British state, placed on-line in 2013, having scored 189 hits.

Reports from the Coleg's Publications Officer indicate that download figures for the e-books series compare favourably with other open Coleg resources posted on their website.

Most Welsh-medium modules only run in two yearly cycles, and it is too soon therefore to see any corresponding pattern between publication and use in specific courses. However, analysis of the unique hits over the five quarters between October 2014 and the end of 2015 reflect the increased activity during the academic year, with a strong peak in the spring semester of 2014, and a dip during the summer holidays, climbing again with the new academic year. The first peak was also helped by the publication of three popular books during this quarter, and a marketing campaign that led to one book jumping from 17 hits during the previous quarter to 133 hits after it had received media attention.

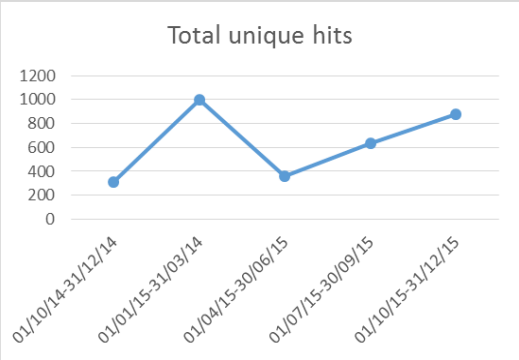

Figure 2: Total unique hits on e-books webpages

8. Student Questionnaire

The project also wanted to quantify students' use of e-book technology, their attitudes towards it, and their awareness of the project in general. As a result, in December 2015, an on-line questionnaire was prepared by the Coleg Cymraeg Cenedlaethol, with input by the project team, using the SurveyMonkey survey tool. E-mail notification, tweets and other social media used to encourage students to respond to it. The questionnaire's aim was to measure student awareness of the e-books, understand what devices and resources students use, and gage students' response to the academic e-books.

Unfortunately, only 27 responses were received from students studying a range of subjects through the medium of Welsh including Law, Welsh, Social Science, and Science on undergraduate and postgraduate level. As the request sent out was a general one, it is not possible to quantify the number of recipients, but the low number of responses was disappointing. However, the Coleg deemed the survey sufficient for their purposes as a broad indication of student attitudes.

Below we discuss a selection of the most relevant questions asked, translating the questions from the original Welsh.

Figure 3 displays the answers to the question: 'What device are you using at this moment?' A list of options were given.

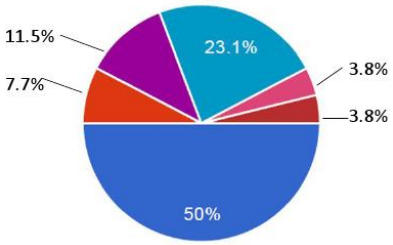

Figure 3: Electronic devices used by students

Windows Computer	13	50%
Apple Mac Computer	2	7.7%
Linux Computer	0	0%
Chrome OS Computer	0	0%
iPad	3	11.5%
iPhone	6	23.1%
Android mobile	1	3.8%
Android tablet	0	0%
Other	1	3/8%

Unfortunately, this question was not well-formulated. It was unclear whether the question referred to the device being used to complete the questionnaire or the device responders were using as e-book readers. Also, only 25 out of the 27 responders answered this question, perhaps because of its ambiguity. It is tempting to ascribe the low reporting of mobile phones to the fact that responders were using their Windows computers to fill in the questionnaire, rather than their lack of mobile phone ownership.

Because of the nature of the question it is difficult to compare the results with the question in the Welsh Books Council questionnaire of 2011, where the general public were asked what device they were then using as e-readers. At the time only 30% of respondents reported that they used an e-reading device (including smartphones). Despite these issues, comparing the results of the two questionnaires still indicates a 20% increase in the years since 2011. It is not clear however whether this is the result of demographic difference, with students being a younger, and therefore more text savvy generation, compared to the general population, or the result of increased use of smartphones and other e-reading devices in the intervening years, or perhaps a mixture of both. The use of paper textbooks continues to decrease due to financial constraints, with electronic devices replacing traditional books in both primary and secondary schools in Wales.

The next question asked 'Where do you usually find electronic resources (e.g. e-books and e-periodicals) for your course?'. A list of options were given. Figure 4 shows the results:

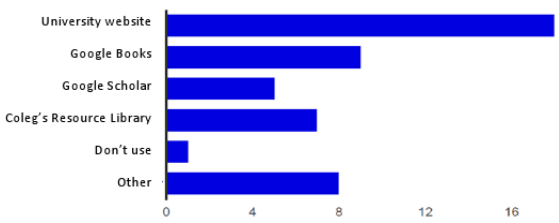

Figure 4: Students' choice of website for resources

university website	18	69.2%
Google Books	9	34.6%
Google Scholar	5	19.2%
Coleg's Resource Library	7	26.9%
I don't use electronic resources	1	3.8%
Other	8	30.8%

Only 26.9% used the Coleg's Resource Library, indicating lack of awareness. Coleg has recognized that it needs to market its resources, as a new establishment it's focus in the first years was to establish the brand.
Students were then asked 'Did you know about the e-books in the Coleg's Resource Library before starting to fill in this questionnaire?' 50% answered that they hadn't heard of the project beforehand. This supported the project team's concern that both undergraduate and postgraduate students were not aware of the e-books' existence, or the website where they're hosted, and the other resources found there. See figure 5 below.

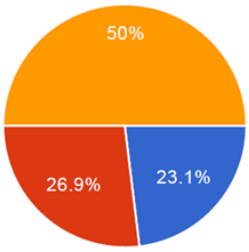

Figure 5: Student awareness of the project's e-books

Yes, and I have used the e-books before now	6	23.1%
Yes, but I haven't used the e-books before	7	26.9
No	13	50%

The questionnaire then asked the student to open a link to one of the e-books and give their first impressions. The answers were written in sentences; one student didn't answer, two answered saying that they hadn't seen the e-books (or maybe hadn't understood the question) and another two that they didn't like reading on screen, preferring hard copies. All the other answers were positive, with remarks such as 'high quality', easy to use', 'good range of subjects', 'very useful', 'easy to download', 'hope to see more published', 'some books I can't get hold of in the library', 'brilliant', 'space-saving'.
Question 7 asked: 'How easy was it to use and read the e-books?' See Figure 6.

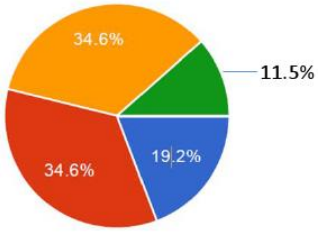

Figure 6: E-books functionality

Very easy	5	19.2%
Easy	9	34.6%
Neutral	9	34.6%
Difficult	3	11.5%

Taking into account that five of the 27 students who answered the questionnaire hadn't answered the previous question, asking them to open an e-book on their device/computer, the answers to question 7 are positive. Students would also be unfamiliar with e-books in Welsh, as the lack of Welsh e-books in general, and for academic purposes in particular, may mean that the conventions of e-book use would be new to them.
Another question listed the useful features which make e-readers ideal tools for academic reading. Students were asked: 'Do you consider the features listed below useful?' The results are found in Figure 7.

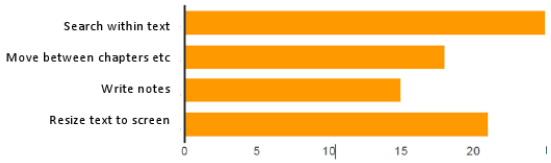

Figure 7: Useful features of EPUB

Being able to search within the text	25	96.2%
Being able to move easily with hyperlinks between chapters, end notes etc	18	69.2%
Being able to write notes within the e-reader and save them	15	57.7%
Being able to resize text to fit the screen	21	80.8%

Searching within the text, and being able to resize text to fit the screen, features that are present in the EPUB and MOBI formats, were both considered very important to the students.
When asked: 'How likely are you to use one or more of the academic e-books in the future?' the students answered:

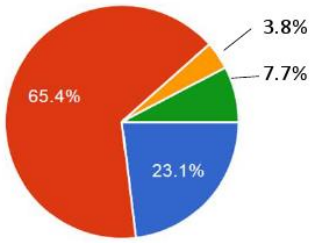

Figure 8: Further use of e-book

Very likely	6	23.1%
Likely	17	65.4%
Neutral	1	3.8%
Unlikely	2	7.7%

This may be criticized as being a leading question, but the questionnaire itself was partly intended as a marketing tool to raise awareness of the e-books, and the answers

therefore show its success in this respect.

Despite the small size of the questionnaire sample, the results confirmed that there was a need to increase the marketing of both the e-books themselves and the Coleg's digital library in general. More could be done to promote the use of digital resources, the advantages of e-books and the advanced features they offer, such as cross-referencing and note-taking. Improved explanation of the different formats would also help, e.g. using the wording 'for iPads' instead of the unfamiliar term 'EPUB', and 'for Kindle' instead of the unfamiliar term 'MOBI'.

9. Conclusions

This project has highlighted a successful method of republishing academic books in a less-resourced language. Weaknesses in dissemination and uptake have also been identified, but these are currently being addressed, and it is hoped that the Coleg's forthcoming move to an improved e-learning platform will address some of the general issues regarding access to the resources in the digital library. Establishing the resources as part of the general library provision to all university students in Wales is also likely to increase their visibility and consequently their use. The positive effect of publicity campaigns as evidenced in the website statistics was noted, and this is one area that the project team would like to see emphasized further.

The e-books produced by this project are also sustainable for the long term, as unlike traditional paper-based books they will not go out of print, and are easily accessible and downloadable from the web. The creation of the DECHE corpus of Welsh academic writing is an additional output which has long-term value, without having added much extra work or cost to the project.

It is hoped that the e-books will also stimulate original scholarly discourse through the medium of Welsh. Already an academic conference on the work of the philosopher J.R. Jones has been organised (Celebrating J.R. Jones Conference, 2016), an event which would be far less likely to be held without the republication of his entire works through the DECHE project.

In the context of efforts to revitalize a minority language, issues of perceived status and participation in the digital media are also relevant. If the rest of academia was moving to electronic resources, leaving Welsh-medium courses with old-fashioned paper books, the impression could be given that the Welsh language itself is old-fashioned and unable to take a full part in the digital world.

To date, efforts in this project have concentrated on reaching its intended academic target audience. Although there is more work to be done in publicizing the e-books already published, their reception and uptake has been encouraging, and e-publication for Welsh academic works, both new and reprinted, has been firmly established. The technology itself is mature, and although further improvements could be made to Welsh OCR to speed up the digitization process, a practical, proven methodology exists for reprinting Welsh academic books in electronic form.

Given that the books produced by this project are available for anyone to download, it is also possible that there may be a wider readership outside academia. Whilst it is not the Coleg's responsibility to market the books to a wider audience, the existence of this substantial body of classic Welsh books, freely available in electronic format, could be of interest to the book reading public in general. It is also a signal that the Welsh language intends to remain contemporary and relevant in the digital age, and that less-resourced languages can use new technology to their own advantage.

10. Acknowledgements

The DECHE Digitization, E-publishing and Electronic Corpus project reported on in this paper was made possible with the aid of a grant from the Coleg Cymraeg Cenedlaethol. The authors would like to thank the Coleg, the National Library of Wales, the original publishers and copyright holders for their help and support in achieving the aims of this project.

11. Bibliographical References

Andrews, T. and Prys, G. (2016). Terminology Standardization in Education and the Construction of Resources: The Welsh Experience, *Education Sciences, Basel.* http://www.mdpi.com/2227-7102/6/1/2/htm. [Accessed 09/02/16].

Bowen, G. ed. (1974). *Traddodiad Rhyddiaith yr Oesoedd Canol.* Gwasg Gomer. Llandusul.

Celebrating J.R. Jones Conference (2016). http://www.colegcymraeg.ac.uk/en/thecoleg/projects/conferences/celebratingjrjonesconference/ [Accessed 01/04/16]

Coleg Cymraeg Cenedlaethol (2016). *What is the Coleg Cymraeg Cenedlaethol?* http://www.colegcymraeg.ac.uk/en/aboutus/whatisthecoleg/. [Accessed 09/02/16]

Dictionary of Welsh Biography (2009). http://yba.llgc.org.uk/en/s2-JONE-ROB-1911.html. [Accessed 11/02/15].

Digital Preservation Strategy, 2012-2015. (2012). National Library of Wales, Aberystwyth. https://www.llgc.org.uk/fileadmin/fileadmin/docs_gwefan/amdanom_ni/dogfennaeth_gorfforaethol/dog_gorff_strat_cad_dig_12_15S.pdf [Accessed 09/02/16].

Finkelstein, D. and McCleery A. (2005). *An Introduction to Book History.* Routledge, New York and London.

Goyal, K. (2006-), Calibre. https://calibre-ebook.com/ [Accessed 01/04/16]

Language Technologies Unit (2012). *E-publishing: Technical guidelines for Publishers in Wales.* http://www.wbti.org.uk/12520.html?diablo.lang=eng. [Accessed 09/02/16]

Llyfrgell Adnoddau (n.d.). Llyfrgell Adnoddau y Coleg Cymraeg Cenedlaethol (The Coleg Cymraeg Cenedlaethol's Resource Libarary – only available in Welsh). https://llyfrgell.porth.ac.uk/ [Accessed 09/02/16]

OmniPage Ultimate (n.d.) http://www.nuance.co.uk/for-business/by-product/omnipage/ultimate/index.htm [Accessed 01/04/16]

Prosiect Digido (n.d.)

https://llyfrgell.porth.ac.uk/library?tag=prosiect%20digi do Coleg Cymraeg Cenedlaethol [Accessed 16/02/16]

Prys, D. Prys; G. Jones D.B. & Chan, D. (2011). *E-publishing in Welsh: A Report for the Welsh Book Council.* Bangor University.

Prys, D.; Roberts M. and Jones, D.B. (2014). DECHE and the Welsh National Corpus Portal. *Proceeding of the First Celtic Language Technology Workshop*, COLING 2014. Dublin. https://www.aclweb.org/anthology/W/W14/W14-46.pdf

Sigil (n.d.) http://sigil.en.softonic.com/ [Accessed 01/04/16]

StatsWales (2011) https://statswales.wales.gov.uk/Catalogue/Welsh-Language/WelshSpeakers-by-LocalAuthority-Gender-DetailedAgeGroups-2011Census [Accessed 01/04/16]

Tesseract (n.d.): https://github.com/tesseract-ocr/tesseract [Accessed 01/04/16]

WalesOnline (11.04.2014). Cardiff. http://www.walesonline.co.uk/news/wales-news/amazon-sparks-language-row-not-2582850. [Accessed 10/02/16].

Welsh Book Council. http://www.cllc.org.uk/. [Accessed 08/02/16].

Welsh Government (2012). *A living language: a language for living*. p. 49. Cardiff. http://gov.wales/docs/dcells/publications/122902wls201217en.pdf [Accessed 09/02/16]

Supporting Language Teaching Online

Cat Kutay
Computer Science and Engineering, UNSW
Sydney, 2052
E-mail: cat.kutay@unsw.edu.au

Abstract

This paper presents the work done in a project to support the teaching of New South Wales (NSW) Indigenous languages through online and mobile systems. The process has incorporated languages with a variety of resources available and involved community workshops to engage speakers and linguists in developing and sharing these resources with learners. This research looks at Human Computer Interaction (HCI) for developing interfaces for Indigenous language learning, by considering the knowledge sharing practises used in the communities, and we compare this work in Australia with similar findings on language reclamation with the Penan indigenous people Malaysia. The HCI studies have been conducted in workshops with linguists and community members interested in studying and teaching their language. The web services developed and used for various languages uses processes of tacit knowledge sharing in an online environment.

Keywords: under-resourced languages, indigenous knowledge sharing

1. Introduction

This work was conducted over many years with languages spoken in Sydney, but often with roots in other regions of New South Wales, Australia (NSW). These are languages spoken by a limited number of elders, and with language resources that may be a dictionary, some archival tapes held at Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS), documents written by missionaries or recent tapes. Also the use that the community wishes to make of these resources may vary. Some wish to teach in schools, others in community. Most have a limited group of people interested or able to engage in the work.

To support the reclamation of these endangered language we have undergone a process of developing web and mobile tools for sharing and teaching language begins with consultations with the community members who have requested web support for their languages, and continues through workshops to train people how to use the system.

The software developed is stored as a simple unit on github, where the different options can be selected to suit community needs. Also a mobile app to utilise the data created for the web site has been used for one language. We are now working on the interface to the next generation app, which will include language comparison between dialects or related languages, and focuses on weekly wordlists or topics.

In this paper we compare our work to similar studies for language reclamation for the Malaysian Penan, a very different language under similar threat of extinction, and compare to their use of knowledge sharing resources. We have found some similarity in the development process for community consultation and designing learning tools so out of interest in the generalisation of this study we have referred to their work where relevant.

The research and development of the web service for language teaching does not depend on the language structure so much as language usage. Hence we believe this work has application to oral languages across the globe, and provides an opportunity for Indigenous people to share their experience in learning and teaching under-resourced languages.

2. Background

We are working with both coastal and inland NSW languages where the structure and features of the language varies, although they are all of the Pama–Nyungan family. The similarity in the projects presented here is the learning focus, where community speakers or those trying to reclaim the spoken language are teaching in school or community groups and require support for language learning.

We are dealing with cultures that are under threat from encroaching western culture, where the elders wish to retain their languages, as a way to describing and understanding the country were they live, and the culture they still maintain.

The process of language reclamation has been two fold in Australia. Firstly to gain recognition of the importance of the maintenance of these language in school curriculum and in the community, and secondly to collect and understand the language resources that exist for each language and build on these to assist learners.

The focus of the teaching is on story telling and learning in context. The teaching uses the Accelerate Second Language Acquisition method (ASLA) developed by Stephen Neyooxet Greymorning, an Arapaho teacher from Montana. The Muurrbay Aboriginal Language and Culture Cooperative promoted this method for language teaching as it provides a context for the learning.

Hence the research focuses on the protocols and methods used in language sharing and teaching in the Australian Indigenous context. The specific features of the software include some language analysis to develop the simple parsing provided on the site. However this parsing focuses more on providing links between text, audio and image material and uses generic analysis of the language itself in terms of recognising common elisions and word forms.

3. Tacit knowledge sharing

The study of oral knowledge sharing begins with an understanding of the process of tacit knowledge sharing, and study of Indigenous cultures in the Pacific has strengthened our understanding of this process (Zaman et al., 2011). The storytelling process is used in many Aboriginal Australian communities as a way to carry on knowledge, so it is instructive to understand what works and what protocols are needed to carry out this teaching. In the telling or retelling of a story there are various rules that have relevance to providing stories in a permanent online repository:

Authority to speak: A significant feature of traditional storytelling is that only those with authority to speak are permitted to present a story. Authority comes from 'being there' in person or through a close relation, being part of the group involved in the story or having some kinship connection to the story (Povinelli 1993).

Community narrative: When a story is told at a community gathering such as a corroboree, many people will contribute the part they know, what they have experience in. First a theme of the story is established, then the many performers add their knowledge.

Deferral to others: When Aboriginal storytellers are speaking, they tend to include or invite other speakers into the story, either as a way of varying the story to keep the listener's attention, as a way of emphasising main points by getting corroboration, or to allow alternative view points to be expressed as a way to help the learner understand.

Knowledge is given not requested: While the teller of the story may start at any point in the narrative, it is their decision where to start. To elicit information a learner must give their present knowledge first as a statement of understanding, rather than a question, so the teller knows where to start and how to direct their story

In this work we use these criteria to evaluate the interactive tools that were developed. The analysis comes from data collecting, workshops, meetings and discussions. There was little opportunity for formalised study of the students or staff working with the system, however we were able to collect feedback from a variety of sources. Basing this on the traditional protocols provides grounding to the evaluation.

4. Grammar of knowledge sharing

In languages it is grammar that provides the cohesion of knowledge. To provide more than just a system for information sharing we developed a knowledge cohesion system that is respectful of the culture being shared.

The first aspect of the Aboriginal language that was instrumental in initiating the revitalisation process in NSW was the naming of place. This arises from the strong cultural tie to land and the fact that languages are used to name the land and create ties between people and land.

This is expressed by Aboriginal knowledge sharing practises in that stories are remembered and re-expressed as located in place. Also Langton (1997) notes that through the cyclic nature of the kinship system, a person's mother's mother and father's father will be the same moiety, and hence will often relate in the same manner to the same country, which reinforces this link to place to the grandchild. Starting with this relationship mode we look at how to support cultural knowledge.

For thousands of years**,** Indigenous people have been sharing knowledge through oral means on how to live in and maintain both themselves and their physical and social environment. While some of this knowledge have been recorded in language, and some is available on the Internet, the online framework for this knowledge is highly unstructured. Research is being done on providing ontologies and frameworks that will provide online learning spaces for this knowledge, especially while retaining the oral format (Kutay and Ho, 2009).

The conception of an oral storytelling grammar is to support the sharing of Aboriginal knowledge online while respecting the cultural representation of knowledge. It is recognised that Aboriginal people have avoided colonisation in many aspects of their culture while living within the mainstream (e.g. Schwab, 1995), and wish to retain alternative means of living and knowing.

Online repositories of stories, supported by the cultural grammar, are becoming a learning tool for those within the culture, as well as those outside the culture to increase their understanding. It also enables Aboriginal trainers to access resources from the community to provide a broader range of cultural training.

The first aspect of the grammar is the protocols listed above which determine how stories are shared by community. The second aspect is the context: schools, community or University. The third aspect is the content, what resources are available in the language and what resources do we have that will help us to develop more material (such as living speakers). From this we have considered how the speakers and students can interact with the language on the web.

4.1 Syntax

Aboriginal Australian story telling is a communal form of oral history designed to fit the community inheritance structure. While social status is granted to people based on their skills and experience, this authority is shared with others of equal skill in other areas, those with the same kinship and hence the same social and environmental responsibilities.

Aboriginal people use a group story telling process (performed as a corroboree) to select the stories that are valuable and hence worth repeating at ceremonies. Tis process also determines what is retained over time, and what is retained as knowledge of the environment.

This process is comparable to the social constructivist learning process (Berger and Luckmann,1996). However the particular theme under which any story is presented may vary over time as priorities and events change.

Stories are placed in a story-path according to themes. These may be relating to morality e.g. how to uphold the

law or the suffer penalties, and be presented with examples both from the Dreamtime and under the new non-Aboriginal law. Alternatively, stories may relate to preservation of the land, and the processes used by ancestors which may relate to a path across the land that people can travel. The story will then describe in sequence the features, seasonal food etc. that can be found and the different aspects of the environment such as the star locations at that time.

Any story path may be related to an area in space, a path in time or a story on a theme (see Figure **2**). These provide the framework within which the story is presented. Also this provides a context into which future telling of the story can be repeated and reinforced.

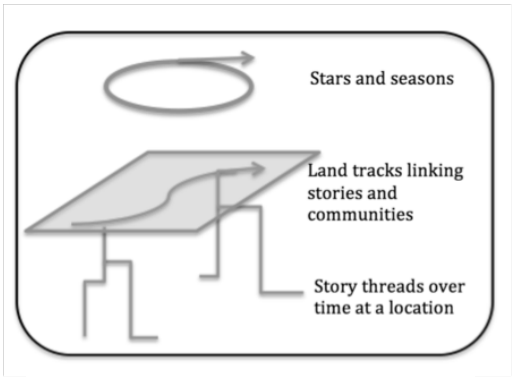

Figure 2 Depiction of story components (Kutay & Ho, 2009)

However it is important to note that this is only a framework. At any performance, only some stories will be told, and only some parts. These will be chosen for their significance to the community at that time, in that season and given recent social and political events.

It is this flexibility in the knowledge sharing that has both ensure the Aboriginal knowledge system has survived the huge upheaval of invasion, but also the highly variable climate of Australia.

In the online environment this approach to knowledge sharing focuses on the ability to incorporate new language words and stories from the community at all times, and we need to be able to support this community conribution.

5. Interface design model

Various studies have been run on the way Aboriginal people view the online environment, how they could use this for knowledge sharing, and the format in which the sharing might be done (Kutay 2011 & 2012). To provide interfaces for community use, we relate our work to existing knowledge sharing practises to reduce the cognitive load of the community members engaging with the system. Similar work has been done for language collection and sharing with the Penan (Zaman 2015), where the process of language classification for sorting words in the interface provided inside into the generational differences in language comprehension. We will discuss this later as it is an interesting aspect of cognitive load for users of mobile interfaces. While the Penan language is not of similar structure or provenance to Australian languages, the way that people live and organise knowledge is similar, as well as the challenged in enabling engagement with information technology.

The model described here provides a conceptualisation or representation of searching, in this case for language information, from the perspective of Indigenous learning within the corroboree setting, where the re-enactment of the real environment assists the user in the construction of their knowledge. The model shown in Figure 1 is a process by which we can analyse the web systems we develop and ensure we cover the complete aspects of the system including the information gathering and learning process.

Pirolli and Card (1997) conceptualized searching for and making sense of information by using concepts borrowed from evolution, biology, and anthropology together with classical information processing theory called the information foraging food-theory (IFT). They describe searching strategies in terms of making correct decisions on where to search next, influenced by the presence or absence of "scent."

Starting with an ecological framework (Bishop 2007) which provides the levels of analysis of the user's environment, (shown as the left column in Figure 1), we mapped this to the key components of the knowledge grammar: the content, context and cohesion of knowledge within the site as pattern attributes (Kutay & Ho 2009).

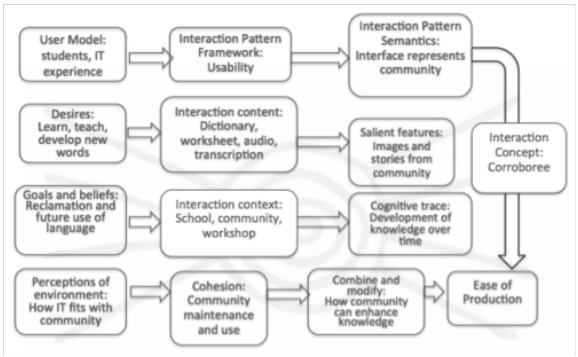

Figure 1: Shows the mapping from the ecological model for participation, to knowledge sharing to HCI aspects,

These were then mapped to the functionality and interface analysis techniques, which cover learning semantics, salient features and cognitive trace plus the final step of combining and modifying information on the site to form knowledge (Rogers, 2004). In particular the ease of production of this knowledge as a sharable resource was the focus of the analysis framework.

The framework provides a way to ensure that the learning system was consistent with the design criteria, the community processes and the learning needs. It was used to assist and evaluate the development of the software platform through workshops and discussion

with linguists, teachers and community members.

5.1 Visual representation of oral knowledge

The aim of the visual design of knowledge learning web sites is to assist learning and enable engagement with the oral components of the language through four aspects, adapted from Rogers & Scaife (1998).

As interfaces become more ubiquitous and pervasive new design paradigms are emerging (Rogers, 2004) as well as the possibility of creating affordances within these interfaces in a manner that may not relate to pre-existing real-world objects or interactions. This assists in reducing the cognitive load, an issue with online learning sites.

1. **Representations within real world**, or juxtapositions that can represent processes in real life narratives. This provides the narrative for the user to follow.
2. **Visual representations of temporal and spatial constraint** such as dialect, speaker, learning environment (worksheet, dictionary, etc), that provide constraints and affordances to assist the learning enquiry, and select the artefacts relevant for further enquiry, and the authority of different annotations available on these artefacts. This forms the resources for the activity.
3. **Artefacts found for further enquiry**, such as audio example, and the authority of different annotations available on these artefacts. This is based on the thematic structure of the activity.
4. **Graphical elements which provide affordances or constrain the inferences** that can be made about the relevance of the search artefacts and relevance to their focus audience, including the level of language used in the document. This provides a context for the user's search activity.

When working online with knowledge artefacts, there will be no 'elders' or over-arching knowledge holders online to tie the information together into knowledge to be understood. In effect an online system provides isolated media packets from which the user has to draw sense. The interface design framework we provide here has been developed around this need to design tools for the processes of information selection (thematic content), the interface format (context), and information linkage (cohesion) to create a knowledge repository.

Any support for the users' external cognition arises from the interaction between internal and external representations when performing tasks that reduce the user's cognitive effort through the use of external representations. The aim is to do this without reducing the information provided.

We used these properties and design dimensions to determine which kinds and combinations of graphical, audio and linkage representations would be effective for supporting different activities. The matrix of affordances provided the semantics of the interface pattern language (see Figure 1).

We will now look at similar models developed for interface design for Indigenous people. Then we provide an example of the use of the model in the development of language sites. These sites are developed to support both teachers who are searching for related material to present to students, and the students doing their own searches to collate knowledge.

6. Previous work on Indigenous interface design

A study by George et al. (2011) of urban Aboriginal people, used Hofstede's (1991) cultural model to analyse websites and provide a method of classifying salient features. They stated that cultural schema must be supported within a context before the culture can be conveyed. In our case the schema is the linkage of knowledge through story, the ability for community to contribute to develop the knowledge, and the levels of access to knowledge. This emphasis is on the multiple layers of knowledge representation within the culture (Pumpa & Wyeld 2006) is also reflected in work with the Penan in Malaysia (Zaman & Winschiers-Theophilus 2015)

Workshops run with the Penan found that the older community members have different schemas for language classification to the younger members, which will make the development of a suitable interface complex, or requiring adaptation. Similarly workshops run with Aboriginal language speakers has shown that there are many different design needs for the representation of language online.

Another project developing a website for sharing the alternate Arandic sign language used in Central Australia, in various contexts by people who also use spoken language (Green et al, 2011). This work required extensive community consultation on how the words are delineated and constructed, as well as how the signing should be authentically represented in an online environment.

The complex process of designing language repositories is repeated with every new project, as the communities deal with a variety of different environmental and social factors that provide a unique system of knowledge and understanding.

The next stage of the work is then to apply the methods developed and the patterns extracted to new situations and so establish the features of each specific site or module developed for culture sharing.

7. Implementation for Language Learning

The design concepts that we developed from the studies described above were divided into the three main systems in the language learning sites we developed. These language sites (Dalang, 2015) are based on a python system that provides easy access to the data to mobile apps and the ability to parse and analyse data uploaded to the site.

The language sites have three main context components. Firstly we provide a searchable dictionary, which is

usually developed by a linguist and transferred to a database. Then archival material in the form of audio and (possibly) transcripts are uploaded and linked to the dictionary words where possible through time-aligned text. These audio files are often in the form of songs, which emphasise the imagery of the subject matter through sound (Magowan, 2001). Therefore much of the wealth of the material lies in the complete tape rather than any segmentation into word or phrase examples. Thirdly teachers can edit and save wordlists and use these to develop worksheets (see Figure 3), which are exercises based on a topic or wordlist.

These worksheets are developed around wordlists, based on the ASLA. The basic process involves a wordlist that is used each week and the learning exercises are based around this list and designed in face-to-face teaching to include actions (eg the girl touched the tree is linked to a picture of this action).

Figure 3 Language worksheet showing sliding popup support tools

Finally language cohesion is supplied by software that links all parts of the web site, including archival material, time-aligned text transcripts, further community contributed recordings, and text material.

The site has four functionalities to support the knowledge design proposed above;

Firstly the data structure is highly interactive, allowing community to add resources, which are then updated as part of the site information without requiring extensive tagging.

Secondly there are functions that enable example sentences to be linked to example recordings of their constituent words or phrase. Hence when linguist's transcripts include recordings that contain a searched phrase**,** these will be automatically linked the words in the example. In the dictionary learners can search for words, and will be provided with example sentences linked to the word.

Thirdly there are various JavaScript functions that provide support for users. Data hidden on the page can be accessed by JavaScript to verify the status and access of the user (e.g. whether teacher or learner) and then support can be tailored to their needs. These links are created interactively as the pages are loaded, since the resources on the site can change over time.

For instance if a sound file exists for a group of words (on the first search of a page) or a single word (on the second search) it is shown as a link for listening. Also resources are provided to searching for a word (student), or creating a word list (teacher). While working on a lesson, and learners have access to the wordlist to help them complete lessons.

Fourthly the site provides an application programming interface (api) for mobile users to access words and wordlists for weekly exercises, based on worksheets. This enables teachers to set up wordlists for users to practise on their mobile. A new mobile app being developed also relies much more on swiping to move around between different views of words (e.g. full text, word list, examples, memory game, etc.) to enable the user to change context while retaining the wordlist as content.

Also another benefit of using python system is this language is designed for text manipulation and parsing. We have linked in the python natural language toolkit (NLTK) to parse example sentence provided by linguists to provide examples relevant to a word. Also we can parse answers entered to exercises by users to verify if they are related to the expected answer. This provides functionality on the site for teachers to set interactive questions with auto-answers, while retaining some options for flexibility in the answers. However it is acknowledged that Aboriginal language structure is highly fluid, using markers to distinguish parts of speech and so sentence comparison can be difficult.

The language system is provided open source as a web service, which can be used with different languages through changing the setup variables, and loading the data for a different language. However the language parser requires more work to assist with generalising between languages. At present we focus on the grammatical features that are common to provide some consistency in support, such as the type of elisions for combining words and the use of markers for parts of speech.

Given the limitations of parsing, we therefore refer first, where possible, to linguistic example sentences that are similar to the example given by a teacher or student in preference to parsing. When teachers enter a search word or phrase, we look for existing examples with those words, and then parse for further examples. We also search for pages within the site that may refer to that word.

7.1 Example of a site as use of the system

The site has been set up for some NSW languages and a separate interface used for each one to provide for material specific to that group, for instance in Sydney the issue around language and culture is that Aboriginal identity is often disputed. The population was decimated early after the English arrived through disease and hunting parties. Also the language has few archival materials and a limited wordlist. The site supports learning language by providing recent recordings, and place an emphasis on linking wiki-style pages on local history and genealogy.

Another language is the Bundjalung language of the northern coast of New South Wales and extending into the state of Queensland. This language comprises five dialects or sub-languages. By combining these on the one site we enable learners of a language that may lack a useful word to select from a neighbouring language.

The web service code is on github (Language, 2015) and is continually updated for the functionality of all sites supporting the differently resourced languages. In this way the functionality and resources can be shared across sites, while retaining the different cultural needs of the various language groups.

8. Verifying the interface design

Using the model developed above we consider now one workshop where the language site was discussed. For this workshop with teachers, we had the printed dictionary, which had been converted from a colour coded word document into a toolbox database to be used on the site. The dictionary includes example sentences after many words, which are also on the site as examples. For this language we had many audio archival tapes, as well as a language speaker. There was some transcription of tapes, but that was not in a searchable order.

The language centre involved in this project has a series of images on various topics that we included in the workshop and teachers could use these to make a collage on the table, depicting scenes and activities. As mentioned we are using the ASLAN process of learning and considering how to support this online.

The proposal was to provide weekly wordlists on mobile and the website on worksheets. The goal is to assist teachers to produce these and students to use and share them in their learning (Ease of production).

We looked at the semantics of the website, how we are to create meaning in the language when many students and teachers have limited vocabulary and grammar. By understanding how users interact with the site we can assist them in this meaning making.

The interaction patterns with teachers working from existing offline resources showed that they would search the dictionary for word, and seek their own dialect first. A second option was a list of neighbouring dialects but the order chosen for these differed for the different dialects. They would then ask others for a pronunciation, or seek an audio version. Then they would check a usage example, also in the dictionary. Finally they would give a changed example that related to their world.

One of the techniques in language teaching is to use archival examples and change one or two words to make a "new" sentence that may be more topical. So the man walked to the mountain may become the man walked to the shops, sometimes using a language word that describes new phenomena such as a shop.

The interaction content they used was the dictionary and each other. They also were keen to listen to the old tapes to hear 'how the language is supposed to be pronounced' but they did not see this as a resource to take apart or to make relevant to any particular topic. Hence the audio tapes do not provide a salient feature for specific language learning. We hope that with transcription this may change.

We also had the images that are shared between many language groups at the centre and teachers produced ones they had made which we have included on the site. There was an emphasis on the need to have images to point to and make the language more active and visual.

The cognitive trace through the material was the theme or word they had chosen and were following up. The conversation may start with 'what is that word' and an attempt to say a half-remembered word, or it may be 'what was a person talking about the other day' or `how do we say this'. The last format came more when planning a lesson, not so much out of community interest. Again the emphasis was to get the version for their specific dialect if possible.

Then the content was combined to provide example sentences with images and sound to provide exercises the students could do while they spoke and listened to each other. When reproducing online this was seen as the need to share graphics, link to audio to help practice and for community to upload new audio when needed.

Once the teachers had a list of words to describe a theme, they collected examples for that theme, through various techniques, including taking examples from the dictionary and slightly altering these to be more topical or to direct the example more towards the specific theme.

Figure 4 Linguistic Toolbox dictionary on the site, showing five dialects

8.1 Cognitive Load

The cognitive load for a learner was considered under the four aspects above. The learners and teachers (who are also learning) are focused on reconstructing the language as close as possible to how it was originally spoken. Hence to reduce this load we looked at:

1. Representations within real world, or juxtapositions that can represent processes in real life narratives.
   The worksheet system is set up with editing tools, shared images and parsing support to assist the teacher creating the sheet and assist the learner who is following the learning material to link new words

with sentences and audio. The links are done within the sheet where possible or in sliding popup windows that follow the user down the page.
We are also providing for male and female voices to be selected as preferred option. This arose from our observations of language learning that when students try to align the sound of their pronunciation with that of the speaker, this is harder to do across gender.

2. Visual representations of temporal and spatial constraint that provide constraints and affordances to assist the learning enquiry, and select the resultant artefacts.
The learner is provided with colour coded dialect options, and a map to show where the dialects are located, as they will use nearby words if no local one available (see Figure 4).
3. Artefacts found for further enquiry, such as audio example, and the authority of different annotations available on these artefacts. This is based on the thematic structure of the activity.
Audio examples are chosen by the dialect option used by the teacher or learner. Also the gender of the speaker can be chosen, which is not used yet as we only have male speakers for most examples.
4. Graphical elements which provide affordances or constrain the kinds of inferences that can be made about the relevance of the search artefacts and relevance to their focus audience,
The dialects are shown in order of locality to the user. Also audio examples are sought as a word group and provided as a link to that group, before individual words, as the sound will change in context.

8.2 Evaluating protocols

The second part of the evaluation is to verify adherence to the design principles developed in the study of the protocols for tacit knowledge sharing. We list here the features developed for this:

Authority of speaker: We need permission to use any archival tapes from the eldest living relative/descendant, which required a long trip around northern NSW to collect. Then we provide these tapes and any community recordings as complete files on the site. However, we are also working with a team to develop a suitable transcription tool to allow segmentation to extract words and phrases for the dictionary interface, so we will need to retain some information on the speaker when segments of their recordings are used across the site.

Community narrative: The site can collect a continually updated series of recordings from community members. This will provide a student with a variety of audio forms, including different dialects and genders to support the different forms of speaking.
We use a map and colour coding to show the source of words as being different dialects as this is important for each community. Also the order of related words is chosen based on the communities expressed perception of what is the more related dialect.

Deferral to others: Where there are many version of a word, including audio versions, we allow various forms to be shown or linked as audio, and the community of language learners and teachers will verify or moderate these. This process is still being worked out so for the present new material is not made public except through the language centre recordings.

Knowledge is given not requested: We are encouraging teachers to develop their own thematic lessons and utilise the material as they wish. The language they know and understand is the material they will teach with best. While teachers can share sheets, we expect they will use their own where possible.

9. Conclusion

This paper presents a study on the cultural sharing of Aboriginal language online, developed in line with an analysis of the community knowledge sharing process. The focus is on improving computer support for oral learning as a way to provide teaching resources that can adapt to the learners' needs and the teacher's focus.
We have attempted to implement the traditional culture of Australian Aboriginal people into the teaching process wherever possible, not just through material, but also method. The work is also designed to be adaptable by teachers to suit the variety of Aboriginal cultures and histories in Australia.
However the important aspect of learning is the student's construction of their knowledge within the social context of the language community and the web can only really provide the information from which learners can do this. Hence we have emphasised the importance of the involvement of human teachers to create the teaching resources, and how the study is on how we can support their work.

10. References

Berger, P. L. and Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge, Garden City, NY: Anchor Books.

Bishop, J (2007) Increasing participation in online communities. In A framework for human–computer interaction. Computers in Human Behavior. 23. pp. 1881-1893.

Dalang, 2015. http://www.dalang.com.au. Retrieved 10.8.15

Green, Jenny; Woods, Gail and Foley, Ben. "Looking at language: Appropriate design for sign language resources in remote Australian Indigenous communities," in Thieberger, Nick; Barwick, Linda; Billington, Rosey and Vaughan, Jill (eds.,).Sustainable data from digital research. Humanities Perspective on digital research. Custom Book Centre, University of Melbourne, 2011

Hofstede G, 1991. Cultures and Organisations. McGraw-Hill, London.

Kutay, C and Ho, P (2009) Australian Aboriginal Grammar used in Knowledge sharing, in Proceedings

of IADIS International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2009), Rome, Italy, November

Kutay, C (2011). HCI study for Culturally useful Knowledge Sharing, Proceedings of the 1st International Symposium on Knowledge Management & E-Learning (KMEL) Hong Kong, 2011.

Kutay, C (2012). Trust Online for Information Sharing, Proceedings of KMIS, Barcelona, October 4-7

Language, 2015. Software for providing language dictionary or wordlist, wiki pages and archival material with tools to support teachers developing learning resources, available on github https://github.com/ckutay/Language. Retrieved 10.3.15

Langton, M. (1997). Grandmothers' Law, Company Business and Succession in Changing Aboriginal Land Tenure Systems, in Galarrwuy Yunipingu, Our Land is Our Life, (pp. 84-117). St Lucia, Queensland: University of Queensland Press.

Magowan, F (2001). Crying to Remember, in Bain Attwood and Fiona Magowan (eds) Telling Stories, Crows Nest, Allen and Unwin.

Pirolli, P, and Card, S. (1997). The evolutionary ecology of information foraging. Technical Report, UIR-R97-0l. Palo Alto, CA: Xerox PARCo (1997).

Povinelli, E (1993). Labour's Lot: The power, history and culture of Aboriginal action, University of Chicago Press.

Pumpa, M & Wyeld, TG (2006). Database and Narratological Representation of Australian Aboriginal Knowledge as Information Visualisation using a Game Engine, Tenth International Conference on Information Visualization (IV'06), IEEE Computer Society, London, United Kingdom, 5-7 July 2006.

George, R, Nesbitt, K, Donovan, M, Maynard, J. (2012) Evaluating Indigenous Design Features Using Cultural Dimensions, in Proceedings of the Thirteenth Australasian User Interface Conference (AUIC2012), Melbourne, Australia

Rogers, Y. and Scaife, M (1998). How can interactive multimedia facilitate learning? In J. Lee, (Ed.) Intelligence and multi modality in multimedia interfaces: Research and applications (pp. 68-89). Menlo Park, CA: AAAl Press.

Rogers, Yvonne (2004). New theoretical approaches for human-computer interaction. Annual Review of Information Science and Technology, 38(1), pp. 87–143.

Schwab, R.R (1995). The calculus of reciprocity: principles and implications of Aboriginal sharing. Centre for Aboriginal Economic Policy Research Discussion Paper No 100.

Zaman, T, Winschiers-Theophilus H (2015) Penan's Oroo' Short Message Signs (PO-SMS): Co-design of a Digital Jungle Sign Language Application. Presented at INTERACT 2015.

Zaman, T., Yeo, A. W., & Kulathuramaiyer, N. (2011). Harnessing Community's Creative Expression and Indigenous Wisdom to Create Value: Tacit-Implicit-Explicit (TIE) Knowledge Creation Model. IKTC2011: Embracing Indigenous Knowledge Systems in a New Technology Design Paradigm

Assessing Digital Vitality: Analytical and Activist Approaches

Maik Gibson
SIL International & Redcliffe College
College Green, Gloucester, GL1 2LX, UK
Email: maik_gibson@sil.org

Abstract

The digital vitality of a language is of concern to many, including linguists and members of the communities that use the language. Kornai (2013) has established a framework for measuring digital vitality which he has used to map the numbers of languages at four different levels - Digitally thriving, Vital, Heritage and Still. This overview is very useful in understanding the overall challenge that minority languages face in digital use. However, when working with a particular community which speaks a minority language, we argue that it is useful to add two more levels of analysis, which would be difficult to justify in an empirical study using mass comparison. As activists we need to be able to identify Emergent use which may not be visible to web crawling, for example in private messaging. Furthermore, identifying where conditions exist for possible digital ascent (e.g. literacy practices, intergenerational transmission of the language in the home) justifies a level we name Latent, which is by its very nature not something to be observed empirically. However the potential for digital development of a language at this level, is, we argue, different from one which is truly Still, unlikely to ascend. The inclusion of these categories might assist digital language activists and communities in effective planning for digital ascent, while not being useful categories for empirical mass comparison. Therefore we propose that action research with a linguistic community will benefit from a five-level framework, and demonstrate how it may be used.

Keywords: digital vitality, multilingualism, language development, language activism, texting, sociolinguistics, language vitality

1. Kornai's Scale of Digital Presence

Frameworks, taxonomies, and scales serve multiple purposes. They can provide an easy way to express a useful overview of large sets of data; to perceive which categories are the most significant; to suggest possible relationships between categories; and to assist in identifying appropriate actions where practical engagement is desired. When considering digital language diversity, Kornai's (2013) framework performs the first function, the useful overview, admirably. When looking at practical engagement with a particular language, we argue that a richer framework introduced in (Gibson, 2015), of which we present an adapted version below, may be useful.

First we provide a brief summary of Kornai's framework, as it is the structure we build upon to introduce further categories.

Kornai's Scale of Digital Presence

Thriving	**T**
Vital	**V**
Heritage	**H**
Still	**S**

Our goal in expanding Kornai's framework is to assist with increasing (and not just analysing) digital language vitality at the lower end. However, for the sake of completeness, our summary starts at the top end of the scale. Here we find **Thriving**, referring to languages with a high level of digital use by both native and foreign speakers, and which also benefit from OS-level support on both Microsoft and Apple platforms (2013:5). Examples at this level include English, French and Chinese.

The next step down, **Vital**, is also for languages "used for communication by native speakers" (2013:5); they are necessarily vital in spoken domains, being at the very least at level 5 in the EGIDS classification (Lewis & Simons, 2010; 2016:104-141). They are used digitally by a significant number of their language community, and will generally also have some level of technological support (e.g. a spellchecker), along with a Wikipedia, and a Bible (Kornai, 2013:6), but will tend not to be used by second language speakers, or have full OS support. Examples cited include Slovak, Assamese and Hausa.

The primary distinction between languages classed as digitally Vital, and those in the **Heritage** category, is the lack of the spoken (as opposed to digital) use, and thereby the identity function. These languages are not used digitally for "two-way contexts such as social networks, business/commerce, live literature" (2013:3), but the digital use instead documents the language – there is a focus on the language itself, the form, rather than the message being conveyed, the function. Examples include Classical Chinese, Sanskrit and Latin – all languages with relatively rich digital resources, but without a community that speak the language natively. To these

classical languages, with large bodies of literature, we may add languages which are recently deceased, but for which digital remains are present, such as Dalmatian. The languages in this category are inherently unable to ascend digitally, as they do not have any native speakers. Kornai (2013:2) says of the digitisation of these language resources that "such efforts, laudable as they are, actually contribute very little to the digital vitality of endangered languages." We therefore argue, that despite their digital presence, Heritage languages should be absent from an activism-oriented scale, because they cannot ascend to a higher level, which is not inherently the case for Still languages; this avoids any misinterpretation of the Heritage stage being intermediate between Still and levels further up the scale.

Furthermore, we note that while it is possible to document dormant or dead languages on the internet, hence raising them from Still to Heritage, this is not a path which will significantly empower a community through language, as it will not be a major medium of communication. Exceptions would include being able to text greetings in a language which has an identificational function for a community, but which is not the primary vehicle for intracommunal communication. So it would be possible to enable inhabitants of Munichis in Peru to use some phrases from the now obsolescent Munichi language while texting, but the purpose would be as a group marker, or as a support to learning the language, rather than a furtherance of the domains of the vernacular language of the home, which in this case is now Spanish. However, the primary concern of the framework presented is to enable speakers of living languages to express themselves in digital writing. Digital literacy, like its non-digital counterpart, is not generally sufficient to reintroduce a spoken language that is no longer used by the community as its vernacular; the Heritage level is not a step on the way to Vital, but something which may be of value to a community which still draws some of its identity from a formerly vibrant language.

Kornai's final category, **Still,** is where there is no observed digital use of the language. He finds that the vast majority (over 95%) of the world's languages, while potentially having spoken vitality, are in this category. This is an empirical observation, based on extensive web crawling combined with automated language recognition software; no trace of the vast majority of the world's spoken languages was found on the internet, so Kornai's judgement that these languages are Still is a fair one, backed up by an abyss of silence. We can conclude therefore that as a scale used for empirical observation of digital language vitality, Kornai's framework functions admirably.

The digital revolution is a relatively recent phenomenon, and continues today, with internet connectivity and the use of smartphones still mushrooming in the developing world. Is it therefore a little early to draw the conclusion stated by Kornai (2013:1), that the "vast majority of the language population, over 8,000 languages, are digitally still, that is, no longer capable of digital ascent"? Here Kornai's analysis shifts from the empirical to the possible, and for those working on the digital development of minority languages, a cut-off date of 2013 for digital development to have begun may seem premature. We will therefore propose that what is Still territory in Kornai's framework be divided into the three categories of Emergent, Latent and Still. This division is especially useful when looking at one language, with rich data sources, rather than taking Kornai's bird's eye view of global trends. Before doing so, however, we will take a brief excursus on the nature of digital writing.

2. The Nature of Digital Writing

Until the arrival of the internet and the mobile phone, Abercrombie's (1963:14) comment that "writing is a device developed for recording prose, not conversation" held not just for the development of literacy, but also for its practice; pre-digital literacy was in the domain of the permanent. And while not all writing was of necessity trying to conform to a standardised norm, this was the usual pattern. Pre-digital exceptions such as graffiti flout conventions both in their placement as well as spelling – a dual rebellion against the established order (see, for example, Pennycook (2010:67) on Singlish graffiti in Singapore). In addition to writing's prose function and its sensitivity to standard norms, we note the relative permanence of non-digital writing. This stands in opposition to conversation, which is in the moment and then disappears, open to selective recall and dispute as to intent.

Text messaging on mobile phones, which is conversational in nature, showed clear signs of deliberate divergence from norms early on in its development, initially attributed in part to the difficulties of the interface, but the subsequent greater ease of texting has not stopped textspeak in English (see Crystal, 2008) having deliberately different norms, especially for informal communication. Brown and Gilman's (1960) distinction between pronouns of solidarity and power can be usefully extended to phenomena such as the period/full stop being deemed a sign of insincerity in texting (Gunraj et al, 2016) or even of anger (Crair, 2013) – both of which we may interpret as an expression of the lack of solidarity. Note that this perception of the full stop as insincere applies only to text messages, not to handwritten notes, so is domain specific, not just a

reflection of informal communication. To text messaging we may add computer-mediated communication, whether messaging services or Web 2.0 platforms such as Facebook and Twitter, which encourage interaction between users. Snapchat accentuates the conversational nature of messaging by speedy deletion of previous interactions. This deliberate divergence from formal norms can play out in more dramatic ways in multilingual societies. We note Paolillo's (2005:63) finding that it is interactivity which encourages the use of the vernacular on the internet – he notes cases from both South Asia and the United Arab Emirates where vernaculars are used far more for chatting than in discussion forums or email, where English prevails.

Multilingual language ecologies tend to have relatively strict norms of what language is appropriate for which type of communication and in which context, i.e. sociolinguistic domains (Fasold, 1984:183). Non-digital writing, being permanent and non-conversational, often triggers the use of more prestigious and state-supported languages (Ferguson, 1959; Lüpke & Storch, 2013:48-76). Writing has not therefore normally been a preferred domain for the vernacular, and we should not be surprised to find speakers of minority languages often more accustomed to writing in other, more widely-spoken, languages; there is little doubt that this will be the pattern for speakers of most minority languages. However, text messaging does not follow formal norms for English, and is a sociolinguistic domain where non-standard and non-normative practices are encouraged. For example, writing about Senegal, Lüpke & Storch (2013:60) note that "Text messages are probably the only context where it is realistic to hope for a use of Baïnounk languages in writing". Likewise, McLaughlin (2009:5) notes, for urban languages, that "Ephemeral media like text messaging on cellular phones or e-mail or internet chat rooms appear to be the prime locations for written forms of urban languages because they are genres that are less formal then other types of writing, and they aim to imitate spoken language."

Along with computer-mediated communication, then, texting provides a written locus for conversation. Texting is therefore a friendlier environment for language varieties which emphasise a relationship of solidarity rather than power – the vernacular, which in multilingual societies is the language with the least overt prestige. This is not to say that other languages are not used in texting and messaging, but that these are the friendliest written environments for the vernacular. So, as Lexander (2011) points out, this is where we are now most likely to see the beginnings of vernacular literacy practices upon which others can be built, in these new sociolinguistic domains, created by technological advances.

3. Emergent and Latent Categories

We see that the most conversational of uses of writing, which as we have noted are the most likely to be where minority languages are found, are also those which are the least likely to be discovered by web crawling, as they are the most private forms of digital communication. When there is widespread vernacular digital literacy, we would expect it to spill over into public uses such as Twitter, or public updates on Facebook. But if this literacy is not widespread, we might well not be able to find it by web crawling. We would therefore categorise use of a language only in private messaging as a case of **Emergent** digital literacy – which will be noted by participant observers, if not from outside the language community. An example of this category is demonstrated by Lüpke and Storch's (2013:59) photo of a text message in Baïnounk Gujaher of Senegal.

The reason that we feel that it is important to add this level to the scale is that it is helpful in proposing optimal interventions for language communities that wish to see their language used digitally. As we saw above, Kornai (2013:2) mentions that while making written texts and dictionaries available online is "laudable," this has little impact on digital vitality. But texting or messaging, because of their conversational nature, are the sociolinguistic domains which are most suited to written vernacular communication, and therefore the best places to start. The creation of new norms for writing can occur within this process – for example the use of numbers to represent certain sounds in the Roman-based Arabizi (used for writing mainly non-standard Arabic in messaging, see Darwish, 2013). Such norms are then available for the language in other domains such as education, and these norms will reflect to some extent a community perspective of what distinctions to write and how; the system will evidently be learnable by new users, as it has already been refined in use. But it is nevertheless unlikely that the norms will be totally stable or universally agreed. We therefore propose that without texting or messaging, other forms of writing are unlikely to take root, and the language will be incapable of digital ascent – if a language is not written in vernacular domains, which are its most natural homes, how will it be used in more formal ones?

Having argued that there is room for an Emergent stage of digital language development, where digital use is found primarily in private media, we propose a distinction between contexts where there is very little chance of any digital ascent (fully Still), and one where the conditions for such digital ascent are in place, but where there is no evidence for current use - **Latent**. As Professor Kornai has pointed out to me in conversation,

evidence for this category in a scale is by its very nature flimsy. Instead, it can only be evoked by reference to other criteria. Before looking at what other conditions need to be met before proposing a language to be at the Latent level, the question of the rationale of doing so must be addressed.

Here our title comes especially into play – while doing a global analysis of the state of digital language vitality, the Latent category will perform no useful function. But when working up close with one or a few language communities, the use of the Latent category may help to identify cases where digital use activities might possibly succeed, and to distinguish them from cases where any digital ascent is extremely unlikely. This distinction will, it is hoped, be of use to those working primarily in the activist rather than purely analytical frame. Different goals benefit from different tools.

The factors that we propose as being significant for identifying a language's status as digitally Latent rather than Still are:

1) stable intergenerational transmission of the language
2) an available model of writing the language
3) the availability of appropriate technology and infrastructure (internet, mobile phone coverage)
4) fonts in which to write the language in the desired script.
5) communal desire to see the language used digitally.

In order to for a language to be classified as at the Latent stage, we propose that all five conditions be met. We observe that these are necessary rather than sufficient conditions for digital ascent. If any one of these is missing, it is unlikely that any community-led digital use of the language will take place. If one is missing, it makes sense to target any activity on that factor rather than on a direct attempt to introduce digital literacy.

1) The first factor, intergenerational transmission, is necessary, as without a vibrant community using the language in day-to-day communication, any digital use will be focused on the language itself rather than the messages that are to be sent and relationships maintained; language is fundamentally a means to an end rather than an end in itself. If a language is not used for active communication, then any use will remain a minority interest, or be found only in marginal cases.

2) Where a language is used in written form in education or other activities such as religious worship (e.g. use in a Bible or hymn book), a model exists for its use. It is rare, though not undocumented, that a community will start writing its own language without some sort of model. The model may sometimes be a language of a neighbouring community, which can raise the question "If them, then why not us?", or be sufficiently understood (whether through intrinsic similarity or familiarity) to serve as a model. For example, in Kenya, languages which have been used in school and church, such as Swahili, Kikuyu and Luo, are commonly found in Facebook posts and comments.

3) The lack of sufficient infrastructure for digital communication is a factor which will apply less over time, as electricity and mobile coverage are rolled out in more and more of the world, but is still relevant in some places. The introduction of a mobile signal to a minority community's area presents an opportunity for language activists to work with the community to establish models of use in the digital domain. And of course many communities will have a diaspora which will have access to the necessary infrastructure in their adopted homes, and might be some of the initiators of vernacular digital use (Gibson, 2012).

4) Where the official script for a language has special characters which are not used for more widely-used languages, speakers often adapt what is already available, focusing more on communication than accurately replicating the standard form. Where the script is not available or easily accessible, people will often adapt what is available, as in the case of Arabizi, which developed not because of a total lack of Arabic script support, but because many devices were not equipped with it (e.g. on early mobile phones). But easy-to-use fonts with the full range of characters will promote digital use, and this is now easier to implement on smartphones and tablets where keyboards are on the screen.

5) The final point is perhaps the most important – if a community shows no interest in developing its language digitally, it is unlikely that any of the above conditions will make any significant difference. Of course today's lack of interest may change tomorrow, and a changing environment (of neighbouring languages being used, or the introduction of a mobile signal) may result in a change of attitude in some sections of the community.

4. An Extended Scale of Digital Presence for Use in Digital Language Development

Here we present the extended scale, including the two new points we have introduced, **Emergent** and **Latent**. We have also excluded the **Heritage** category, as it is not a step of development for living languages; genuine digital ascent occurs only with a language transmitted orally in its community. We do not challenge the fact that this category is useful for understanding digital presence on the web of different languages in Kornai's work, but for the activist hoping to work on increasing digital language development, the Heritage category serves no purpose for a vital language.

Thriving	**T**
Vital	**V**
Emergent	**E**
Latent	**L**
Still	**S**

5. Activities for Moving up to the Next Stage

The framework is proposed in the hope that it will help communities to identify the types of activities that might be the most fruitful for the further digital development of their language. Our primary concern is not with the minority of languages that lie at the top of the table – Thriving and Vital. So we finish by outlining general recommendations for the path towards increased digital vitality for the three levels at the bottom of the scale.

Starting from the bottom up, at the Still level, we would recommend examining the five conditions proposed for a language to be at the Latent level. The first condition, intergenerational transmission, will determine whether any ascent will be towards Emergent or Heritage use; if there is not a community speaking the language, digital uses will either be to document the language, or for linguistically marginal cases such as learning the group's greetings or language of traditional ceremonies. The latter types of activities can be very important in social terms, and are the focus of much of what is called language revitalisation. The subsequent four – a ready model, appropriate infrastructure, useable fonts, and communal desire should then be examined, and action taken to address the lack of necessary conditions. The existing infrastructure is generally something less susceptible to community engagement, as it normally depends on outside agents. But working on the providing a model, or developing fonts are things that activists can contribute towards more easily. Communal interest is of course essential.

At the Latent level, modelling the use of digital literacy alongside appropriate communal activities (given the interest which is part of the definition of this level) may provide the impetus for private digital writing to start. Given that writing in the vernacular has a strong affective dimension, it is hoped that that digital writing in the vernacular will be its own reward, and once started will continue – an issue which would be worthy of fuller empirical investigation.

Once a language has reached the Emergent level, it may be appropriate for the extension of uses that would bring it up to the Vital level. This would mean proceeding from primarily private writing, by way of public posts and discussions on Facebook or Twitter, to more public forms of digital engagement, such as blogs, news sites and community-specific knowledge sharing, whether related to the community's heritage or its current environment. At this stage, tools such as dictionary apps (e.g. for Tuvan, Harrison, 2010:194) can assist both with wider usage of the language and its prestige. To this written digital presence, it would be logical to add audio and video material in the language as part of a broader communal engagement with the internet and digital communication.

References

Abercrombie, D. (1963). *Problems and Principles in Language Study*. 2nd edition. London: Longman.

Brown, R. & Gilman, A. (1960). The pronouns of power and solidarity. *American Anthropologist* 4 (6):24-29.

Crair, B. (2013). The period is pissed. *New Republic*. http://www.newrepublic.com/article/115726/period-our-simplest-punctuation-mark-has-become-sign-anger.

Crystal, D. (2008). *Txtng: the gr8 db8*. Oxford: Oxford University Press

Darwish, K. (2013). *Arabizi Detection and Conversion to Arabic*. http://arxiv.org/pdf/1306.6755.pdf

Fasold, R. (1984). *The Sociolinguistics of Society*. Oxford: Blackwell.

Ferguson, C. (1959). Diglossia. *Word*. 15:325-40.

Gibson, M.L. (2012). Extinct Languages and Languages Close to Extinction: How to Preserve that Heritage? In Vannini, L. & Le Croisier, H (eds.). *NET.LANG: Towards the Multilingual Cyberspace*. 75-88. Paris: C&F Editions.

Gibson, M.L. (2015). A Framework for Measuring the Presence of Minority Languages in Cyberspace. In *Linguistic and Cultural Diversity in Cyberspace. Proceedings of the 3rd International Conference* (Yakutsk, Russian Federation, 30 June – 3 July, 2014). – Moscow: Interregional Library Cooperation Centre.

61-70. http://www.ifapcom.ru/files/2015/khanty/yak_mling_2015.pdf

Gunraj, D. N., Drumm-Hewitt, A.M., Dashow, E.M., Siddhi, S., Upadhyay, N., & Klin, C.M. (2016). Texting insincerely: The role of the period in text messaging. *Computers in Human Behavior* Volume 55, Part B, February 2016, Pages 1067–1075 http://www.sciencedirect.com/science/article/pii/S0747563215302181.

Harrison, K.D. (2010). *The Last Speakers: The Quest to Save the World's Most Endangered Languages*. Washington DC: National Geographic.

Kornai, A. (2013). Digital Language Death. *PLoS ONE* 8(10): e77056. doi:10.1371/journal.pone.0077056 http://www.plosone.org/article/info:doi/10.1371/journal.pone.0077056.

Lewis, M.P. & Simons, G. (2010). Assessing endangerment: Expanding Fishman's GIDS. *Revue Roumaine de Linguistique* 55 (2):103-120. http://www.lingv.ro/RRL%202%202010%20art01Lewis.pdf.

Lewis, M.P. & Simons, G. (2016) *Sustaining Language Use: Perspectives on community-based language development*. Leanpub. https://leanpub.com/sustaininglanguageuse

Lexander, K.V. (2011). Texting and African language literacy. *New Media & Society* 13/3. 427-443.

Lüpke, F. & Storch, A. (2013). *Repertoires and Choices in African Languages*. Berlin: de Gruyter.

McLaughlin, F. (ed.) (2009). *The Languages of Urban Africa*. London: Continuum.

Paolillo, J. (2005). Language Diversity on the Internet: Examining Linguistic Bias. In Paolillo, John, Daniel Pimienta, Daniel Prado et al. *Measuring Linguistic Diversity on the Internet*. Paris: UNESCO. http://www.uis.unesco.org/Library/Documents/measuring-linguistic-diversity-internet-communication-ict-2005-en.pdf

Pennycook, A. (2010). *Language as a Local Practice*. New York: Routledge.

Digital Language Diversity: Seeking the Value Proposition

Martin Benjamin
LSIR, École Polytechnique Fédérale de Lausanne
IC, BC 114, Station 14, 1015 Lausanne, Switzerland
E-mail: martin.benjamin@epfl.ch

Abstract

This paper is a response to the CCURL workshop call for discussion about issues pertaining to the creation of an Alliance for Digital Language Diversity. As a global project, Kamusi has been building collaborative relationships with numerous organizations, becoming more familiar than most with global activities and the global funding situation for less-resourced languages. This paper reviews the experiences of many involved with creating or using digital resources for diverse languages, with an analysis of who finds such resources important, who does not, what brings such resources into existence, and what the barriers are to the wider development of inclusive language technology. It is seen that practitioners face obstacles to maximizing the effects of their own work and gaining from the advances of others due to a funding environment that does not recognize the value of linguistic resources for diverse languages, as either a social or economic good. Proposed solutions include the normalization of the expectation that digital services will be available in major local languages, international legal requirements for language provision on par with European regulations, involvement of speaker communities in the guided production of open linguistic resources, and the formation of a research consortium that can together build a common linguistic data infrastructure.

Keywords: multilingual, philanthropy, funding, participation, open data, language infrastructure

1. Introduction

While the availability of digital resources for a myriad of languages might strike language technology professionals as an obvious good, the topic does not even enter the consciousness of most people in the world. Knowledge of and attitudes toward multilingualism in the digital sphere are important, because the existence of technological resources depends on both the demand for them, and enthusiasm for investment in their production. This paper sets as its problematic the diversity of attitudes toward digital language diversity (DLD), and the challenges that these many perspectives pose for those involved in languages and technology to succeed in a less-than-welcoming funding and policy environment.

The paper first investigates the various parties involved in the intersection between technology and languages, asking for whom DLD is a value. Following the assumption that increasing the quantity and quality of digital resources for diverse languages is an important goal, the paper then asks how to demonstrate the value of DLD for both consumers and policy makers. Finally, the paper discusses an initiative to bring together many groups that are working on various aspects of DLD, to harmonize many currently-atomized projects toward the development of a shared linguistic data infrastructure that will be widely recognized as a valuable goal.

The paper is semi-ethnographic, based on hundreds of conversations with people at many levels of involvement with language and technology from dozens of countries. It is intended to provoke discussion about policy, not to present research results regarding any particular language resource or technology, with the aim of contributing toward action that will open digital resources to billions of people who speak diverse languages that currently sit at or outside the margins of technology. Without a concrete plan that can be pursued over the course of ten or twenty years, action will be haphazard and ineffective. Without understanding the motivations and barriers for the people involved, a concrete plan cannot be developed. This paper seeks to address the social considerations of DLD, in order to foster a hospitable environment for it to thrive.

2. Attitudes Toward DLD

When thinking of DLD, it is first important to recognize that people have a great range of involvement with and perspectives on the subject, ranging from passion to indifference to ignorance to hostility. Many types of actors can be identified, with some broad themes emerging that are nevertheless not universal within categories. In this section, the ways people connect to DLD are differentiated, and the considerations of some of the players are noted.

2.1. Speakers of diverse languages. Language diversity in the digital realm is a fuzzy concept. Without arguing which languages sit where on a scale of tools, data, content, and speakers, it can be seen that languages such as English, French, and German have a great many resources, languages such as Polish and Chinese aspire toward the same, languages like Swahili and Vietnamese enjoy some digital presence but do not necessarily see a future with the same ubiquitous lingosystem as exists for the first category, and a great many languages neither have nor expect a notable involvement in the technological realm.

For people toward the top of the scale, all or many of their regular digital interactions can be conducted in a language

they know, even if services such as speech recognition or autocomplete are unreliable. Speakers assume that language technologies are developing, and services will improve over time. Few give deeper thought to language issues, with many holding the implicit assumption that everyone else should have the wherewithal to maneuver through technology in an available language.

Conversely, people who speak languages lower on the scale do not generally expect that they will ever be represented digitally. People with secondary education often do not find this to be an issue, with technology naturally belonging in the same sphere as their school books, in a language they can read well enough for practical purposes. Those who are not literate in a well-endowed language might be aware of the privileges others enjoy, but have just as much expectation of partaking as they would of being chauffeured in a Mercedes they glimpse on the street. Thus, one either does or does not access technology through a language in which it is already well developed, without demanding or conceiving of services in a local language. Most of the perhaps 95% of people in many African countries who are thereby excluded, for example, do not have exposure to the idea that this could change, while the other 5% do not feel the need.

People will use digital tools that make sense to them. For example, Sri Lankans text in the Sinhala language by transliterating to Latin text, because they do not have useable input devices for their script on their mobile phones, but this temporary expedient would certainly be retired if Android had a well-integrated Sinhala keyboard. As WhatsApp spreads across India, a new method of communications is opening up – conversations as recorded voice messages, detached from real time, that require no greater literacy than the ability to turn on a phone and press the record and stop buttons. This technology is language-neutral on the surface, and greatly enhancing to the linguistic diversity of its users, but will never result in immersion in a lingosystem, only passive improvements in the ability to communicate without entering the technological mainstream.

2.2 Researchers. People who work on NLP and HLT are inherently sympathetic to underrepresented languages. However, research remains stacked in favor of English and a select few other languages, as can be seen in conference programs such as those of the Association for Computational Linguistics.[1] By necessity, researchers tend to develop expertise at the intersection of a certain language or set of languages, and a certain technology or set of technologies. The pity is that subject expertise is difficult to transfer to diverse languages – though much research is generalizable and could be shared in principle, opportunities to do so are rare among established research groups, and impossible for languages that have neither funds nor the research teams to pursue them. Instead, researchers do exemplary work developing digital resources in their language communities, such as solid teamwork among the languages of India, that do not resonate to the benefit of local languages elsewhere. Moreover, teams spend inordinate energy reinventing overlapping tools, such as similar software to build Wordnets in different languages. As discussed below, unifying researchers could both strengthen their individual projects beyond their focus of language or topic, and produce an action agenda that promotes DLD as a whole.

2.3. Governments and policy makers. Among people with important positions in governments and international agencies, four major attitudes prevail. First are those who believe that the path of progress lies with the languages at the top of the scale. This is especially the case in the US, where language policy is geared toward assimilation to English, and support for research on languages low on the scale largely falls to a smattering of funds from the National Science Foundation, the National Endowment for the Humanities, and military budget for languages of strategic interest. However, leaders in other countries also express indifference or hostility to their local languages; for example, official Rome scorns Italy's smaller languages, most Colombian authorities turn their backs on indigenous languages, and some people in high positions in India advocate a focus on English. Second are those who recognize the value of diversity, to the extent that they can promote it for the languages within their ambit. Irish, Welsh, Estonian, Icelandic, even the Sami language spoken by 30,000 people in Finland all enjoy the support of their national governments. The European Commission coordinates major activities toward DLD at the top of the scale; however, EC interest beyond their core 24 official member languages is largely restricted to communication with major trading partners. Spain invests heavily in its regional languages, with Catalan one of the better-resourced languages in Europe. Russia takes an active interest in its minority languages, though print resources prevail over digital. Similarly, South Africa devotes considerable resources to its largest eleven languages, but plays almost no role sharing its expertise elsewhere on the continent. Conversely, the African Academy of Languages has the African Union mandate to promote digital resources for the continent, but no mechanisms to do so. UNESCO, with an established program on multilingualism in cyberspace, has the most holistic global view, but no funds to actuate projects. Third are those who see value in major local languages, such as Swahili or Vietnamese, but do not extend their concern to smaller vernacular languages within their countries. Fourth are those who are interested in their nations' mother tongues, but do not see them as candidates for digital inclusion. An interesting example here is Uganda, where the president was actually the active lead author of print dictionaries for Nyakore and Kiga (Museveni et al 2009 and 2012), but where even the national Luganda language remains at the digital periphery. For Africa,

[1] https://www.aclweb.org/website/node/434

support for digital resources for cross-border languages is a stated policy objective of the African Union, but not one that is buttressed with the resources for implementation.

Precious few international cooperation activities include DLD within their scope; for example, Canada's IDRC supported ITC4D in Africa and Asia for a number of years but has now shifted focus, the British Council is making some investments in supporting mother tongue education at the primary level, and the Swiss SDC has voiced concern for the issue. By and large, however, language remains peripheral to the discourse of international development.

2.4. Donors and foundations. People involved in DLD should recognize that most donors do not find language equity to be a value. Funders have their own agendas, such as curing a disease or saving a forest. Language is rarely on their screens, and may be seen as a hindrance. Getting through doors guarded by program officers who do not see language as part of an organization's mission is almost impossible. Endangered languages do get bits of funding for sentimental reasons, but overall, DLD is seen as unimportant esoterica.

Small private donors largely have no knowledge about language issues. Neither do many DLD projects tread the difficult and poorly trodden path of retail fundraising. Americans in particular are not known for their concern about language, except perhaps for heritage communities that maintain a sentimental attachment to their ancestors or homelands, such as Yiddish. Private donors tend to respond to international concerns when there is a crisis, such as an earthquake or hurricane. People will occasionally respond to heart-tugging appeals about specific endangered languages, but (a) there are too many endangered languages and too few individual donors for that to be an effective strategy toward widespread preservation and documentation activities, and (b) thousands of minority languages that are not on the cusp of vanishing are systematically ignored.

For big donors, language has yet to make a mark as an area of concern. Language barely makes a dent in the grants of the Ford Foundation, for example, with $145,000 spent in 2014 and 2015 on research and development for the emerging Sheng language of Kenya, a $190,000 grant for the Hawaiian language, and $150,000 for a multilingual voter registration platform for Nigeria – not half a million dollars, from an $800,000,000 portfolio [2]. The Gates Foundation has even less interest in language; other than support for English, they have since 2013 granted $100,000 to develop local-language health materials in Burkina Faso, $175,000 for professional development for American teachers of foreign languages, and $100,000 to support language learning for the Makah Nation near their Seattle headquarters, with another $386,000 spent on non-English in prior decades, and no way for prospective grantees to get in the door and make the case for supporting digital language diversity as a path toward the foundations goals of overcoming inequity [3]. For the Hewlett Foundation, language funding equates to English[4]. In an analysis of the grants database of the Foundation Directory[5], Jaumont and Klempay (2015) find that 88% of the roughly 4 billion granted by American philanthropies in Africa over a decade from 2003 went to Anglophone countries, almost entirely for programs conducted in English. Lack of concern for local languages can be further observed in eleemosynary institutions in Europe and elsewhere. Understandably, big donors want projects that can make an immediate, visible impact, whereas language projects have intangible results that might not be evident for decades (if there is ever a way to measure the effect that increased knowledge has on a society, beyond saying that X number of people have used Y resource that contains Z elements). Less benevolently, few philanthropies are amenable to the case that DLD is worth even a moment of their consideration, and neither practitioners nor potential beneficiaries are in a position to demand otherwise.

2.5. Business. The common factor that determines whether a business is interested in DLD is the profit motive, but that can take many forms. Businesses that sell language services often appreciate the value of diversity, though most prefer to focus on languages that promise a bigger return on investment. Other businesses need language resources to communicate with workers, suppliers, or customers. For the first, DLD might have immediate profit motive, such as a translation contract, or might have the long term objective of an expanded usership. However, creating resources for a language for in-house use or external communications, beyond localizing certain material into select languages, is beyond the scope of most businesses. Furthermore, translation agencies have a vested interest in keeping data such as translation memories private. Therefore, companies are often eager consumers of HLT, but not active agents of its production.

A few companies have taken a much longer view toward DLD, with no immediate payoff, but potentially long term value to stockholders. The translation services of Google and Microsoft probably bleed money, requiring vast processing power that is not recovered through sales or advertising revenue. However, as global companies, both understood that most of their potential market does not speak English, so it was logical to start offering services in other languages. Speculatively, as the translation services became increasingly popular, they began to generate their own momentum, and their improvement is now tied as much to the corporations' sense of mission as to any financial aims. Certainly, Google and Bing Translate are espoused as general-purpose public services with unspecified social benefits down the line. At the same time,

[2] https://www.fordfoundation.org/work/our-grants/grants-database/grants-all

[3] http://www.gatesfoundation.org/How-We-Work/Quick-Links/Grants-Database, search term = language

[4] http://hewlett.org/grants/search

[5] https://fconline.foundationcenter.org/

Google's attention to localization, as exemplified by the excellent versions of its software in Swahili produced by its Nairobi office, expands their reach to millions of customers who are gaining increasing access to technology, and who do not have technical literacy in English. While Google could be critiqued for a large range of shortcomings in their language offerings and their approach to sharing data, they and a few other forward-facing companies are helping lead the development of linguistic resources for nearly 100 languages, offering proof positive that a good bankroll and a cutting-edge technological back end can advance development for any chosen language.

3. Constraints on DLD Development

The lack of a profit incentive for languages down the scale means that most DLD efforts are promulgated by people with a greater sense of mission than a budget to implement it. SIL, for example, coordinates the work of numerous dedicated field researchers, from freely available FLEx software for gathering lexical data, to the Webonary system for hosting results in a standard, searchable format. Yet, though each project is bilingual with a major contact language, there is no common core of senses that is shared among projects, and thus no way to link the work that is done on one language with the work that is done on any other, nor to deployment within technologies that build upon linked data. This is an example of how collaboration within the Human Languages Project (HLP) discussed below, particularly mapping emerging sense-specific concept sets that can be used across projects, could save a lot of repetition and confusion.

Academic projects, when funded, also produce results that produce problems. First, the projects are limited to the term for which they receive funding, which means that they might not get all the way through to stated aims, or might reach those objectives – development of a prototype, acquisition of a particular amount of data – and then have to stop. Second, electronic resources need a perpetual host, or they disappear, and digital results all too frequently vanish when funding runs out, or the researcher moves to another university and their original server account is deleted. Additionally, many academic projects are not conceived to integrate with wider efforts, for example as data that can be used for downstream applications, or run into insurmountable barriers regarding copyrights or the expense and time needed to share results beyond the articles that describe them.

The recent growth of technological hotbeds in places such as Nairobi and Accra has not resulted in major new resources for the languages of their countries. Bright young techies have little financial incentive to pursue projects for local languages. As with IT professionals everywhere, they take jobs that have a good chance of financial reward. Usually, that means working on business or e-governance projects that do not include language concerns. As an example, the Kenya Revenue Authority's online tax filing service was an expensive investment that employed skilled programmers, but is not available in any Kenyan language[6]. Meanwhile, adventurous entrepreneurs face enough risk launching startups, without venturing into unproven language markets. While one could argue that localized shells and local content could be profitable, for example with an Android action game, that is not an argument that has attracted many risk-takers in Johannesburg or Bangalore.

Wikimedia's forays into diverse languages demonstrate that creating content and data in a language requires more than an open platform. Though they list Wikipedias in nearly 300 languages[7], far fewer have enough articles or information to attract readers or count as original linguistic content. A large percentage of multilingual Wikipedia content is generated by robots, usually stub articles with formulaic translations, such as this random entry describing some asteroid, typical of the Yoruba Wikipedia: "3585 Goshirakawa jẹ́ plánẹ̀tì kékeré ní ibi ìgbàjá ástẹ́rọ̀ìdì"[8]. Wiktionary has similarly established shells for 172 languages[9], but close inspection shows that much of the content for many languages is useless at best. The utility to the speaker communities is therefore minimal, and uptake for most languages negligible. Nevertheless, the existence of workspaces for the languages sends a dangerous signal that the languages are already taken care of, and that the community will take control of its own resources with no need of further external effort or concern.

Despite the existence of diverse Wikimedia shells, as well as free blogging platforms that support any UNICODE script (though without localized interfaces), most individual speakers do not see themselves in a position to do anything about their own languages. Non-specialists cannot take responsibility for difficult infrastructure; few people install their own water pipes, or write their own word processors, and none can take on all the work necessary to create their own lingosystem. Standard users do not control the technology, cannot localize a piece of software, and cannot issue data into the void. While they could in principle add Wiki or blogging content, few know this is even a possibility, and there is no well-trod path that starts and keeps people involved in content or technology creation.

4. Normalization

DLD will not come about of its own accord. There is too much of a gap between the interests in language by the people who create digital resources, and the people who

[6] https://itax.kra.go.ke/KRA-Portal/
[7] https://en.wikipedia.org/wiki/List_of_Wikipedias
[8] https://yo.wikipedia.org/wiki/3585_Goshirakawa
[9] https://meta.wikimedia.org/wiki/Wiktionary#List_of_Wiktionaries

speak diverse languages who are not in positions to effectively demand services. However, no great technological leaps are required to create a full panoply of resources for any given language. The heavy lifting in HLT that is undertaken for languages at the top of the scale can be applied to other languages at relatively high speed and low cost. For instance, speech recognition technology does not need to be invented anew, but rather have existing technology trained with data from a new language. DLD is a matter of the time invested in gathering data, building linguistic models, and creating content. Features inherent to a language, such as diverse writing systems and grammars, are relatively surmountable challenges. However, few people are aware of the pathways, fewer are passionate about the desirability of following them, fewer yet are in a position to work toward implementation, and nobody with money will fund any rigorous effort to address the situation.

What is needed is the normalization of the expectation that each language should have a digital existence. So far, there has been no effort to create public awareness about the possibility for linguistic equity, so people who might wish for good resources think that they are about as likely as their traveling to the moon, and therefore worth about as much time investigating. People who do not know that it is possible to make resources for their language will certainly not demand it. For most people, technology is something that one takes as it comes, without thought of going to the manufacturer and asking for new features. Without economic power to exert, and no political groundswell to demand change, linguistic communities do not even dream of a meaningful presence in the digital sphere.

Beyond the persistent efforts of language technology developers to demonstrate that digital resources can be brought into existence whenever the funds and personnel are available, people interested in DLD can pursue two strategies:

5. Advocacy

The first strategy toward digital inclusion is aggressive advocacy. Ordinary citizens cannot demand language services, but their governments can. However, for governments to make such demands, policy makers need to believe that they are both reasonable and achievable. The case can be made in a few areas, which do not all involve digitization. In most instances, regulations can be adopted directly from existing European directives, both because the wording has been well hammered by lawyers, and because no European country could object to trading partners elsewhere in the world imposing exactly the same linguistic conditions as they demand for themselves.

Language advocates should advance model legislation for interested nations, requiring that corporations provide information and services in major local languages, in areas such as food labels, medicines, aviation, product safety instructions, and any product purchased under government contract.

Aviation is the showcase for decision makers, who tend to fly frequently on international carriers, to recognize the desirability of services in national languages. As Air Canada puts it, "Safety is always our number one concern. For this reason, … earphones are not allowed during critical phases of flight as they would prevent you from hearing safety announcements."[10] Airlines cannot argue with their own insistence that it is essential for passengers to understand instructions from the flight crew, and having those instructions in a language that can be understood by the citizens of the country they are flying to fits directly within that logic. Further, there is almost no additional cost for an airline to train a flight attendant with native language skills; with a short grace period, Air France, British Airways, and other carriers could have speakers of national languages on flights serving their entire route system in a few months.

Government contracts are the next step, with a proven pathway to success. The Brazilian government, for example, will not purchase any product that is not available in Portuguese, so major software manufacturers localize their products to that language without question as part of their normal development cycle. The costs of localization are extremely low versus potential sales to government agencies, whether in software where the purchaser might otherwise be tempted to FOSS solutions, for light bulbs where a competitor could easily claim the market by printing a few extra words on their packaging, or for SUVs where a hundred-page user manual could break the sale of a fleet of expensive vehicles. Of course, if light bulbs are packaged for the government in the local language, the same packaging will make it to ordinary store shelves – which is the wider objective.

EU Regulation (EU) No 1169/2011, on the provision of food information to consumers, states that, where labelling is required, it should be "in a language easily understood by the consumers of the member states where a food is marketed". For medicines, Directive 2001/83/EC of the European Parliament, on the Community code relating to medicinal products for human use, states that, "The package leaflet must be clearly legible in an official language or official languages of the Member State where the medicinal product is placed on the market, as specified, for the purposes of this Directive, by that Member State". Similar regulations exist for medical devices and for other products. It should be beyond dispute that the same rules apply for the first languages of billions of other people around the world.

Beyond software localization, these mandates would promote DLD in two ways. First, many of these

[10] http://gofar.aircanada.com/en/go-far-answers/question/earphones-during-ttl/

requirements will be best provisioned with digital intervention, such as the development of translation systems that can produce results acceptable in a legal context. Second, the growing presence of local languages within national markets will lead to increasing expectations that they will become ubiquitous, including within the digital sphere; if tinned tomato labels can be understood by local purchasers, why not tax filing services or voice commands to a mobile device? Realistically, regulations can only enforce the improvement of resources in select non-European languages, but those few dozen will both satisfy many existing deficiencies, and open the door to DLD for languages even farther down the scale.

6. Production

Where policy makers and the public agree that languages have value, and funds and interest can be mobilized, the expansion of DLD depends on the production of resources within each language. This is not straightforward, because there are many more languages than there are existing advocates, researchers, or business cases.

For an example of production possibilities, I point to the design of the Kamusi Project to enroll speaker communities in the production of data for their own languages. Such data can be used for future technologies, with the goal of digital lingosystems far along the scale; online systems can work for languages with a critical mass of networked speakers, a threshold that has not yet been explored. The systems for community participation have been described elsewhere (Benjamin 2015, Benjamin and Radetzky 2014). In short, games and mobile activities elicit consensus-validated data through targeted microtasks that are designed to be fun and compelling. The tasks are built on premises discussed above, that people do not have the individual ability to develop their own language resources, but will contribute if doing so is easy and well explained, and does not require their own technical or financial investment.

Several incentives are posited to give value to community members to participate in the DLD production process. The first is the creation of resources that can make their own lives easier, for example by producing terms that they see will go directly on product labels for the foods they buy. Second is producing something for their children, including data that can be used in L1 education. Third is producing something for the community; this is expected to be a particularly strong motivation among diasporic populations who wish to give back to their homelands. Fourth are intrinsic rewards, such as pride in seeing one's language grow online, and the recognition within social networks that one is taking an active role in advancing language development. Finally, many people find language play to be inherently enjoyable; people pay for games like Scrabble for languages high on the scale, so there is every reason to suppose that people will enjoy passing time with free games as have never before been available for less resourced languages. Unfortunately, despite extensive development on the back end, the systems have not yet been released publically at the time of writing to test these hypotheses, due to a technical constraint that can be represented thusly in UNICODE: ~~money~~.

Lack of money is the most consistent obstacle for DLD. That is, technology does not present barriers, because most languages can piggyback on prior work for other languages. Nor does DLD necessarily require the agreement of policy makers, although that would help lubricate the finances; as academic researchers and business initiatives show, digital resources can appear whenever someone takes the initiative to create them, regardless of official support. This paper therefore closes by inviting interested readers to participate in an emerging consortium to create a Human Languages Project, along the lines of the Human Genome Project or the Human Brain Project, that unites groups from around the world in the development of tools to produce language data, the development of the data itself for a great diversity of languages, and the development of tools to deploy that data in advanced HLT knowledge and NLP applications. Instead of competing for non-existent funding, banding together within HLP can make the case that digitization of the world's languages is an important and realistic goal, that can be achieved by a network of competent partners with a modicum of philanthropic and intergovernmental support.

7. Conclusions

While those active in producing resources for DLD assume the value to be obvious, the case has not yet been made to the powers of the purse. Researchers can discuss success rates for L1 education (Ouane and Glanz 2011), humanists can wax sentimental about the heritage at the cusp of disappearing in endangered languages (Kornai 2013), and activists can bewail the deep and enduring inequities caused by grossly imbalanced language resources (Osborn 1997). However, for EC funding within Horizon 2020, the argument boils down to one consideration: markets. H2020 calls for cross-lingual data development are entirely focused on "data value chains" of "industrial importance"[11]. Sentiment plays no role. Use value to marginalized people is of no relevance. Self-interest, in terms of European trade and security benefits, are the important features for gaining EC support.

The question of how to gain philanthropic support for DLD, particularly from US foundations, is one for which no answers are evident. No foundation currently expresses DLD as a value, beyond limited support for endangered languages, and they do not entertain proposals that seek to convince them otherwise.

[11] Big Data PPP: cross-sectorial and cross-lingual data integration and experimentation, http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/5093-ict-14-2016-2017.html#

What can be attempted is a united approach by like-minded parties, under the auspices of an organizing framework such as HLP. This is in keeping with the objectives laid out in the "Roadmap toward UNESCO's World Atlas of Languages"[12] to the safeguarding of linguistic diversity through the effective application of ICTs. Languages cannot sell themselves, especially languages with few speakers, that have calculably lower economic value. However, pooling resources can lead to much lower costs per language, creating economies of scale that might just tip the balance toward funding support for languages across the board. Such a consortium could create the enabling environment in which DLD thrives – and is thus offered as the value proposition for funding agencies to create a linguistic data infrastructure for languages at all points along the scale.

8. Bibliographic References

Benjamin, M. 2015. Crowdsourcing Microdata for Cost-Effective and Reliable Lexicography. Proceedings of AsiaLex 2015, Hong Kong

Benjamin, M., and Radetzky, P. 2015. Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification. LREC, 2014

Jaumont, F. & Klempay, J. 2015. Measuring the influence of language on grant-making by U.S. foundations in Africa. Reconsidering Development, 4(1), 51-65.

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

Museveni, Y., Muranga, M.J.K., Muhoozi, A., Mushengyezi, A., and Gumoshabe, G. 2009. Runyankore/Rukiga-English Dictionary. Kampala, Uganda: Institute of Languages, Makerere University, 213 pp

Museveni, Y.K., Muranga, M.J.K., Muhoozi, A., and Gumoshabe, G. 2012. Katondoozi y'Orunyankore-Rukiga. Thesaurus of Runyankore-Rukiga. Kampala: Fountain, 504 pp.

Osborn, D. 1997. Ultimate Development Participation: Institutionalizing Indigenous Language Use in Education and Research. 23rd annual Third World Conference, Chicago, Illinois

Ouane, A. and Glanz, C. 2011. Optimising Learning, Education and Publishing in Africa: The Language Factor. UNESCO/ Association for the Development of Education in Africa/ African Development Bank.

[12] http://unesdoc.unesco.org/images/0024/002438/243852e.pdf

Innovative Technologies for Under-Resourced Language Documentation: The BULB Project

Sebastian Stüker[0,a], Gilles Adda[b], Martine Adda-Decker[b,c], Odette Ambouroue[d], Laurent Besacier[e], David Blachon[e], Hélène Bonneau-Maynard[b], Elodie Gauthier[e], Pierre Godard[b], Fatima Hamlaoui[f], Dmitry Idiatov[d], Guy-Noël Kouarata[c], Lori Lamel[b], Emmanuel-Moselly Makasso[f], Markus Müller[a], Annie Rialland[c], Mark Van de Velde[d], François Yvon[b], Sabine Zerbian[g]

(a) Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany
(b) LIMSI, CNRS, Université Paris-Saclay, France
(c) LPP, CNRS-Paris 3/Sorbonne Nouvelle, France
(d) Langage, Langues et Cultures d'Afrique Noire Laboratory (LLACAN), France
(e) Laboratoire d'Informatique de Grenoble (LIG)/GETALP group, France
(f) Zentrum für Allgemeine Sprachwissenschaft (ZAS), Germany
(g) Universität Stuttgart/Institut für Linguistik, Germany

Abstract

The project *Breaking the Unwritten Language Barrier* (BULB), which brings together linguists and computer scientists, aims at supporting linguists in documenting unwritten languages. To achieve this, we develop tools tailored to the needs of documentary linguists by building upon technology and expertise from the area of natural language processing, most prominently automatic speech recognition and machine translation. As a development and test bed for this, we have chosen three less-resourced African languages from the Bantu family: Basaa, Myene and Embosi.

Work within the project is divided into three main steps:

1) **Collection** of a large corpus of speech (100h per language) at a reasonable cost. After initial recording, the data is re-spoken by a reference speaker to enhance the signal quality and orally translated into French.

2) **Automatic transcription** of the Bantu languages at phoneme level and the French translation at word level. The recognized Bantu phonemes and French words will then be automatically aligned to extract Bantu's morphemes.

3) **Tool development**. In close cooperation and discussion with the linguists, the speech and language technologists will design and implement tools that will support the linguists in their work, taking into account the linguists' needs and technology capabilities.

The data collection has begun for the three languages. We have been using standard mobile devices and a dedicated software—LIG-AIKUMA, which offers a range of speech collection modes (recording, re-speaking, translation and elicitation). LIG-AIKUMA's improved features include a smart generation and handling of speaker metadata as well as re-speaking and parallel audio data mapping.

Keywords: Language documentation, automatic phonetic transcription, unwritten languages, automatic alignment

1. Introduction

Despite tremendous progress in technologies involving speech recognition (e.g., SIRI on iPhones), only a very limited portion of the languages spoken in the world is covered by technology or by scientific knowledge. As for technology, only normative productions of very few languages in very few situations are mastered. When speech is less normative (due to age, spontaneity, speaker's origin, pathology,...) performance drops significantly, with multiplicative effects in case of multiple factors (Gerosa and Giuliani, 2008); this also reflects weaknesses in modelling speech and language. The technological divide becomes most prominent when considering the number of languages spoken in the world: we have a minimally adequate quantity of data for less than 1% of the world's 7,000 languages. Also, most of the world's everyday life speech stems from languages which are essentially unwritten[1], while most technological developments are targeting written languages.

There are thousands of endangered languages for which hardly any documentation exists and time is running out before they disappear: some linguists estimate that half of the presently living languages will become extinct in the course of this century (Nettle and Romaine, 2000; Crystal, 2002; Janson, 2003). Even with the upsurge of documentary linguistics (Himmelmann, 2002; Woodbury, 2011), it is not realistic to expect that without the help of automatic processing the documentary linguistics community will be able to document all these languages before they disappear—given the number of languages involved and the amount of human effort required for the "creation, annotation, preservation, and dissemination of transparent records of a language" (Woodbury, 2011).

In this article, we present the French-German ANR-DFG project *Breaking the Unwritten Language Barrier* (BULB), whose goal it is to develop a methodology and correspond-

[0] apart from the first two authors, the names are in alphabetical order

[1] We include in these languages ethnolects as well as sociolects such as many regional varieties of Arabic, Shanghainese, slang ...

ing tools to achieve efficient automatic processing of unwritten languages, with a first application on three mostly unwritten African languages of the Bantu family (Basaa, Myene and Embosi). Among the languages in danger of disappearing, many of those that have not yet been properly documented are non-written languages. The lack of a writing system makes these languages a challenge for both documentary linguists and natural language processing (NLP) technology. In the present project, we will therefore conduct the necessary research to obtain the technology that is currently missing to efficiently document unwritten languages. Work within the project is divided into three main steps:

1. **Collection** of a large corpus of speech (100h per language) at a reasonable cost. For this we use standard mobile devices and a dedicated software called **LIG-AIKUMA**. LIG-AIKUMA proposes a range of different speech collection modes (recording, respeaking, translation and elicitation). LIG-AIKUMA's improved features include a smart generation and handling of speaker metadata as well as respeaking and parallel audio data mapping (see (Blachon et al., 2016) for further details). After initial recording, the data is respoken by a reference speaker to enhance the signal quality, and orally translated into French.

2. **Automatic transcription** of the Bantu languages at phoneme level and the French translation at word level, followed by the **automatic alignment** of the recognized Bantu phonemes and the French words. The collected oral data (Bantu originals and French translations) contain the necessary information to document the studied languages. Phonetic alignments are highly valuable for large scale acoustic-phonetic studies, phonological and prosodic data mining and dialectal variations studies; cross-language alignments such as the one exhibited for Spanish and English in Figure 1, extracted from (Stüker et al., 2009), but between groups of Bantu phonemes and French words, may also prove very useful for morphological studies, vocabulary and pronunciation elaboration.

3. **Tool development**. Tools will be built upon all these data and alignments. In close cooperation and discussion with the linguists, the speech and language technologists will design and implement tools that will support the linguists in their work, taking into account the linguists' needs and technology capabilities.

2. NLP Technology for Language Documentation

2.1. Language Independent Phoneme and Articulatory Feature Recognition

Systems for language independent phoneme recognition often utilize multilingual models (Kohler, 1996). The idea behind this type of approach is to identify phonemes that are common to multiple languages, e.g., by using global phoneme sets, such as the International Phonetic Alphabet (IPA). Models for phonemes that are common to multiple languages share all the training material from those languages. A multilingual model can be applied to any language that was not originally included in the training languages. The phonemes of this additional language that are not covered by the multilingual model need to be mapped appropriately. By pooling phoneme sets and data from multiple languages one achieves two effects. First, the number of phonemes covered by a multilingual model is in general larger than that of a monolingual model. Second, by pooling data from multiple languages a model can become more robust to slight cross-linguistic variations in the pronunciation of phonemes that are nonetheless denoted by the same symbol. Alternatively to phonemes, methods exist to recognize articulatory features across languages, either with monolingual models from many languages or with multilingual models trained on many languages (Stüker et al., 2003). The advantage of articulatory features-based models is that their coverage for a new language is generally higher than that of phonemes, as articulatory features can be recognized more robustly across languages.

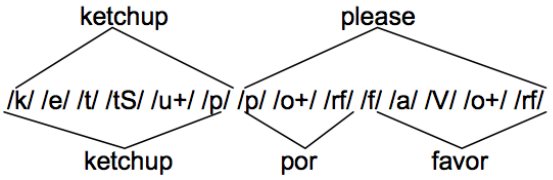

Figure 1: Example of alignment found by GIZA++ between words (in english) and phonemes (in spanish) (Stüker et al., 2009).

Inspired by work in speech synthesis (Muthukumar and Black, 2014) the project decided to pursue a two step approach of first finding the phoneme boundaries in audio recordings of a new languages, and then, in a second step, to classify the identity of the detected phoneme segments. The segmentation step again can be done with multilingual phoneme recognizers from whose output only the detected phoneme boundaries are retained while the detected phoneme identity is discarded (Vetter et al., 2016). The second step can then be achieved based on multilingually detected articulatory features.

2.2. Word Discovery by Word-to-Phoneme Alignment

The feasibility of automatically discovering word units (as well as their pronunciations) in an unknown (and unwritten) language without any supervision was examined by (Besacier et al., 2006). This goal was achieved by unsupervised aggregation of phonetic strings into word forms from a continuous flow of phonemes (or from a speech signal) using a monolingual algorithm based on cross-entropy. This approach lead to almost the same performance as the baseline approach, while being applicable to any unwritten language.

(Stüker and Waibel, 2008) introduced a phone-based speech translation approach that made use of cross-lingual

supervision. This approach works on a scenario in which a human translates the audio recordings of the unwritten language into a written language. Alignment models as used in machine translation (Brown et al., 1993; Och and Ney, 2003) were then learned on the resulting parallel corpus consisting of foreign phone sequences and their corresponding English translation. (Stüker et al., 2009) combined this approach with the monolingual approach above and also did contrastive comparisons. (Stahlberg et al., 2012) and (Stahlberg et al., 2013) then continued to work on this approach by enhancing the alignment model for the task and examined the impact of the choice of written language to which the phoneme sequence is aligned.

Working with a similar goal in mind, and using bilingual information in order to jointly learn the segmentation of a target string of characters (or phonemes) and their alignment to a source sequence of words, (Xu et al., 2008; Nguyen et al., 2010) are building on Bayesian monolingual segmentation models introduced by (Goldwater et al., 2006) and further expanded in (Mochihashi et al., 2009). This trend of research has become increasingly active in the past years, moving from strategies using segmentation as a preprocessing to the alignment steps, to models aiming at jointly learning relevant segmentation and alignment. (Adams et al., 2015) reports performance improvements for the latter approach on a bilingual lexicon induction task, with the additional benefit of achieving high precision even on a very small corpus, which is of particular interest in the context of BULB.

Many questions still need to be addressed. Implicit choices are usually made through the way data are specified and represented. Taking, for example, tones into account, prosodic markers, or even a partial bilingual dictionary, would require different kinds of input data, and the development of models able to take advantage of this additional information.

A second observation is that most attempts to learn segmentation and alignments need to inject some prior knowledge about the desired form of the linguistic units which should be extracted. This is because most machine learning schemes deployed in literature tend to otherwise produce degenerated and trivial (over-segmented or conversely under-segmented) solutions. The additional constraints necessary to control such phenomena are likely to greatly impact the nature of the units that are identified. Supporting the documentation of endangered languages within the framework of BULB should lead us to consequently question as systematically as possible the linguistic validity of those constraints and the results they produce. The Adaptor Grammar framework (Johnson et al., 2007; Johnson, 2008), which enables the specification of high-level linguistic hypotheses appears to be of particular interest in our context.

Another important aspect of the endeavor we are facing lies in the noisy nature of the input produced by the phonemicization of the unwritten language. Processing a phoneme lattice instead of a phonemic transcription, following the work of (Neubig et al., 2010), seems to be a promising strategy here.

More generally, a careful inventory of priors derived from the linguistic knowledge at our disposal should be undertaken. This is especially true regarding cross-lingual priors we can postulate about French on the one hand, and Basaa, Myene and Embosi on the other hand: for lack of taking such priors into account, it is dubious that general purpose unsupervised learning techniques will succeed in delivering any usable linguistic information.

2.3. Preservation of Unwritten Languages by Advanced Technologies

(Bird, 2010) described the model of "Basic Oral Language Documentation", as adapted for use in remote village locations, which are "far from digital archives but close to endangered languages and cultures". Speakers of a small Papuan language were trained and observed during a six weeks period. A technique called re-speaking, initially introduced by (Woodbury, 2003), was used. Re-speaking involves listening to an original recording and repeating what was heard carefully and slowly. This results in a secondary recording that is much easier to transcribe later on (transcription by a linguist or by a machine). The reason is that the initial speech may be too fast, the recording level may be too low, and background noise may degrade the content. For instance, in the context of recording traditional narratives, elderly speakers are often required (and they may have a weak voice, few teeth, etc.) compromising the clarity of the recordings (Hanke and Bird, 2013). In (Bird and Chiang, 2012), the use of statistical machine translation is presented as a way to support the task of documenting the world's endangered languages. An analogy is made between the primary resource of statistical translation models—bilingual aligned text—and the primary artefact collected in documentary linguistics—recordings of the language of interest, together with their translation. The authors suggest exploiting this similarity to improve the quantity and quality of documentation for a language. Details on the mobile application (called AIKUMA) are given in (Hanke and Bird, 2013). AIKUMA is an Android application that supports the recording of audio sources, along with phrase-by-phrase oral translation. In their paper, the concept of re-speaking was extended to produce oral translations of the initial recorded material. Oral translation was performed by listening to a segment of audio in a source language and spontaneously producing a spoken translation in a second language. The process did not involve the touch screen or any buttons, but relied exclusively on audio and proximity sensors for control. After re-speaking, the mobile phone stored the source audio file, along with the translation file and a mapping file. The translation file contained the concatenated recordings of the oral translations. The mapping file specified how each segment of oral translation corresponded to the source audio. Users could listen to the original version, the translation, or interleaved playback of the original with the translation.

Finally, it is also worth mentioning the work of (Kempton and Moore, 2014), who suggest the use of advanced speech technologies to help field linguists in their work. More precisely, they propose a machine-assisted approach for phonemic analysis of under-resourced and under-documented languages. Several procedures are investigated (phonetic similarity, complementary distribution, and min-

imal pairs) and compared.
During the first year of BULB, features were added to the original AIKUMA app to facilitate the collection of the parallel speech data required in the project. The resulting app, called LIG-AIKUMA, is described in Section 3.

3. LIG-AIKUMA recording application

3.1. Motivations and specifications

Within BULB, the use of LIG-AIKUMAis associated with a set of use cases identified by a series of operations to perform. The first one is basic audio recordings. The next two consist of re-speaking and translation. The last one is speech elicitation from displayed texts, images or videos.
As the use of LIG-AIKUMA is driven by these goals, the user interface of AIKUMA had to be changed so as to identify the various modes and focus on them. A number of other changes were also made to further facilitate the use of the application, for instance when it comes to gaining time in saving and loading the meta data of the latest recording, or providing a better feedback on the re-speaking once it is done.

3.2. Recording modes

The core features of the initial AIKUMA for recording, re-speaking and translation have been kept, along with the storage of the speaker meta data. Some parts of the interface have also been reused.
New developments have focused on the setup of 4 modes, dedicated to specific tasks of speech recording. The home view is illustrated in Figure 2 (left). As one can see, the following four modes are identified:

- Free recording of spontaneous speech
- Re-speaking a speech segment (previously recorded with the app or loaded from a wav file): the re-speaking now allows to (optionally) listen to the latest recorded segment so as to check it and re-speak it if needed, before moving on to the next segment. Also, once the re-speaking is done, a summary view displays the new segments and their corresponding original segments, and allows to (optionally) listen to or re-speak any of them before finishing the session. In Figure 2 (middle), original segments are aligned with re-spoken ones. Both can be played while the latter can also be re-recorded if necessary, which is useful for double checking and error correction.
- Translating a recording (previously recorded or loaded): same features as for the re-speaking mode except that the source and target languages must be different.
- Elicitation of speech from a text file (image and video media will follow very soon): the user loads a text file within the app, then reads the sentence, speaks, listens to the recording for checking and goes on to the next sentence, etc. This mode was specifically required for the data collection which took place in Congo-Brazzaville during summer 2015. Figure 2 (right) illustrates the text elicitation mode.

3.3. Current state of development

The interface has been adapted for the large screen of tablets (10 inches), so the app works both on Android powered smartphones and tablets. Apart from the specifications, based on multiple discussions with linguist colleagues, this new version was developed in approximately 3 man/months and generated 5000+ lines of code. All the new code has been put on the LIG forge and is accessible open source[2] for use or development on demand. The application LIG-AIKUMA has been successfully tested on different devices (including Samsung Galaxy SIII, Google Nexus 6, HTC Desire 820 smartphones and a Galaxy Tab 4 tablet).
Figure 3 illustrates the use of the LIG-AIKUMAtablet to collect Embosi verb conjugations (left) or more free conversations including several speakers (right).
Users who just want to use the app without access to the code can download it directly from the forge direct link[3].

4. Documentation of three Bantu Languages

4.1. Bantu languages

In BULB, three typologically diverse northwestern Bantu languages were selected, which stem from different Guthrie zones (areal-genetic groupings, (Guthrie, 1948)): Basaa (A43, Cameroon), Myene (B10, Gabon) and Embosi (C25, Congo-Brazzaville). The Bantu family is one of the largest genera in the world and most of the genetic and typological diversity within this family can be found in the northwestern part of the domain, closest to the Bantu homeland. As northwestern Bantu languages are spoken in the so-called *fragmentation belt*—a zone of extreme linguistic diversity—they differ from their eastern and southern Bantu relatives such as Swahili, Sotho or Zulu in that they are much less studied, protected and resourced.
Our three Bantu languages however have in common that they are relatively well described, as there are also competent native-speaker linguists working on each of them and, at least in the case of Myene, some basic electronic resources are already available (albeit in need of further development to make them suitable for corpus-based linguistic analyses). This was an important criterion in our choice of languages, as the available linguistic analyses will allow us to test the efficiency and improve the outcome of our new tools.

4.2. Three under resourced Bantu languages

Basaa, which is spoken by approximately 300,000 speakers (SIL, 2005) from the "Centre" and "Littoral" regions of Cameroon, is the best studied of our three languages. The earliest lexical and grammatical description of Basaa goes back to the beginning of the twentieth century (Rosenhuber, 1908) and the first Basaa-French dictionary was developed over half a century ago (Lemb and de Gastines, 1973). Several dissertations have focused on various aspects of Basaa (Bot ba Njock, 1970; Makasso, 2008) and

[2] https://forge.imag.fr/projects/lig-aikuma/
[3] https://forge.imag.fr/frs/download.php/706/MainActivity.apk

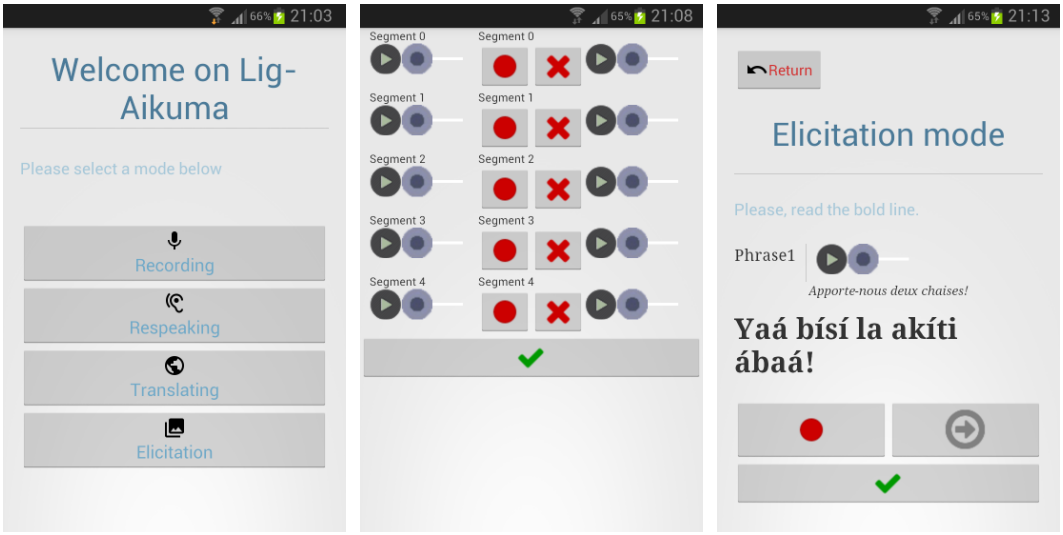

Figure 2: Screenshots from the LIG-AIKUMA application: from left to right, i) home view ; ii) summary view after respeaking is done, speakers may play and edit every segment; iii) elicitation mode

Figure 3: LIG-AIKUMA on Android tablets being used for data collection: elicited verb conjugations spoken by a female native speaker of Embosi (left) and free conversations involving several male native speakers (right).

the language also benefits from recent and ongoing linguistic studies (Dimmendaal, 1988; Hyman, 2003; Hamlaoui and Makasso, 2015).

Myene, a cluster of six mutually intelligible varieties (Adyumba, Enenga, Galwa, Mpongwe, Nkomi and Orungu), is spoken at the coastal areas and around the town of Lambarene in Gabon. The current number of Myene speakers is estimated at 46,000 (Lewis et al., 2013). The language is presently considered as having a "vigorous" status, but the fact that no children were found that could participate in a study on the acquisition of Myene suggests that the language is already endangered. A basic grammatical description of the Orungu variety (Ambouroue, 2007) is available, as well as a few articles on aspects of the phonology, morphology and syntax of Myene ((Van de Velde and Ambouroue, 2011) and references therein).

Our third and last language, **Embosi** (or alternatively Mbochi), originates from the "Cuvette" region of the Republic of Congo and is also spoken in Brazzaville and in the diaspora. The number of Embosi speakers is estimated at 150,000 (Congo National Inst. of Statistics, 2009). A dictionary (Beapami et al., 2000) is available and, just like Basaa and Myene, the language benefits from recent linguistic studies (Amboulou, 1998; Embanga Aborobongui, 2013).

From a linguistic perspective, the three languages display a number of features characteristic of the Bantu family: (i) a complex morphology (both nominal and verbal), (ii) challenging lexical an postlexical phonologies (with processes such as vowel elision and coalescence, which bring additional complexities in the recovery of individual words), and (iii) tones that serve establishing both lexical and grammatical contrasts. Regarding the latter feature, we will be able to build upon the expertise gained in the automatic annotation of the tonal systems of South African languages (Barnard and Zerbian, 2010), although other tonal aspects

of our northwestern Bantu languages will require the development of specific approaches.

4.3. Recording of Bantu Languages

From our experience, we have evaluated the quantity of spoken data to be recorded, re-spoken and translated to 100 hours per language, in order to build reliable models for transcription and alignment, and extract some useful information from them. A part of this data will be transcribed, in order to evaluate the automatic transcription and alignment. At the moment of writing about 50 hours of Embosi have been recorded and partly re-spoken using LIG-AIKUMA while Myene (44 hours[4]) and Basaa (40 hours) have been recorded partly with LIG-AIKUMA and mobile devices, partly with traditional methods. Some of the recordings are obtained with elicitation (dictionary entries, sentences from (Bouquiaux and Thomas, 1976)), and some are conversations (see section 3.). The data collected within this project will be provided after the end of the project to the general scientific community via the ELDA agency.[5]

5. Project perspective and methodology

The development of LIG-AIKUMA continues, with the experience gained from the the first successful fieldwork experiences. We are developing the elicitation modes (with images, videos), as well as the different features needed to save the work done during the field-trips.

BULB's success relies on a strong German-French cooperation between linguists and computer scientists. So far, cooperation has been fostered and strengthened by a series of meetings and courses benefiting the scientific community beyond the present consortium. During the courses, the linguists presented to the computer scientists the major steps to document an unknown language, and the computer scientists introduced their methods to process a "new" language and generate phonetic transcriptions and pseudo-word alignments.

Our three chosen languages, Basaa, Myene and Embosi, have in common a lack of stable orthographic conventions and a lack of texts. Their linguistic resources generally rely on a handful of speakers and few of them are corpus-based. The BULB project will also have the positive outcome of adding to the existing resources (100 hours per language with some transcription and translation) and will thus allow to address new questions with the help of new methodologies (Rialland et al., 2015).

What do endangered languages spoken by few individuals and other unwritten, major languages (e.g., Shanghainese, spoken by 77M people) have in common? They lack written material which drastically limits their access to language processing tools such as speech recognition or translation, not to mention other NLP tools. Our goal is to develop a methodology that can ultimately be applied to any mostly or completely unwritten language, even if it is not endangered.

[4]20 hours were recorded before the project

[5]Evaluations and Language resources Distribution Agency Evaluations and Language resources Distribution Agency `http://www.elda.org`

Acknowledgements

This work was realized in the framework of the ANR-DFG project BULB (ANR-14-CE35-002).

6. References

Adams, O., Neubig, G., Cohn, T., and Bird, S. (2015). Inducing Bilingual Lexicons from Small Quantities of Sentence-Aligned Phonemic Transcriptions. In *12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam.

Amboulou, C. (1998). *Le Mbochi: langue bantoue du Congo Brazzaville (Zone C, groupe C20)*. Ph.D. thesis, INALCO, Paris.

Ambouroue, O. (2007). *Eléments de description de l'orungu, langue bantu du Gabon (B11b)*. Ph.D. thesis, Université Libre de Bruxelles.

Barnard, E. and Zerbian, S. (2010). From Tone to Pitch in Sepedi. In *Proceedings of the Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU10)*.

Beapami, R. P., Chatfield, R., Kouarata, G., and Waldschmidt, A. (2000). *Dictionnaire Mbochi - Franạais*. SIL-Congo, Brazzaville.

Besacier, L., Zhou, B., and Gao, Y. (2006). Towards speech translation of non written languages. In Mazin Gilbert et al., editors, *SLT*, pages 222–225. IEEE.

Bird, S. and Chiang, D. (2012). Machine translation for language preservation. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India*, pages 125–134.

Bird, S. (2010). A scalable method for preserving oral literature from small languages. In *Proceedings of the Role of Digital Libraries in a Time of Global Change, and 12th International Conference on Asia-Pacific Digital Libraries*, ICADL'10, pages 5–14, Berlin, Heidelberg. Springer-Verlag.

Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.-N., Adda-Decker, M., and Rialland, A. (2016). Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. submitted to SLTU 2016.

Bot ba Njock, H.-M. (1970). *Nexus et nominaux en bàsàa*. Ph.D. thesis, Université Paris 3 Sorbonne Nouvelle.

Bouquiaux, L. and Thomas, J. (1976). *Enquête et description des langues à tradition orale.Tome II: Approche linguistique (questionnaires grammaticaux et phrases)*., volume II of *SELAF*. Peeters Publishers.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Crystal, D. (2002). *Language Death*. Cambridge University Press. Cambridge Books Online.

Dimmendaal, G. (1988). *Aspects du basaa*. Peeters/SELAF. [translated by Luc Bouquiaux].

Embanga Aborobongui, G. M. (2013). *Processus segmentaux et tonals en Mbondzi – (variété de la langue embosi*

C25). Ph.D. thesis, Université Paris 3 Sorbonne Nouvelle.

Gerosa, M. and Giuliani, D. (2008). A comparison of read and spontaneous children^{2}s speech recognition. In *The 1st Workshop on Child, Computer and Interaction, WOCCI 2008, Chania, Crete, Greece, October 23, 2008*, page 5.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 673–680, Sydney, Australia, July. Association for Computational Linguistics.

Guthrie, M. (1948). *The classification of the Bantu languages*. Oxford University Press for the International African Institute.

Hamlaoui, F. and Makasso, E.-M. (2015). Focus marking and the unavailability of inversion structures in the Bantu language Bàsàá. *Lingua*, 154:35–64.

Hanke, F. R. and Bird, S. (2013). Large-scale text collection for unwritten languages. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 1134–1138.

Himmelmann, N. P. (2002). Documentary and descriptive linguistics. In *In Osamu Sakiyama and Fubito Endo (eds.), Lectures on Endangered Languages 5, 37-83. Kyoto: Endangered Languages of the Pacific Rim*.

Hyman, L. (2003). Basaá (A43). In Derek Nurse et al., editors, *The Bantu languages*, pages 257–282. Routledge.

Janson, T. (2003). *Speak: A Short History of Languages*. Oxford University Press.

Johnson, M., Griffiths, T. L., and Goldwater, S. (2007). Adaptor grammars: a framework for specifying compositional nonparametric bayesian models. In B. Schölkopf, et al., editors, *Advances in Neural Information Processing Systems 19*, pages 641–648, Cambridge, MA. MIT Press.

Johnson, M. (2008). Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27, Columbus, Ohio, June. Association for Computational Linguistics.

Kempton, T. and Moore, R. K. (2014). Discovering the phoneme inventory of an unwritten language: A machine-assisted approach. *Speech Communication*, 56:152–166, January.

Kohler, J. (1996). Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 4, pages 2195–2198 vol.4, Oct.

Lemb, P. and de Gastines, F., (1973). *Dictionnaire Basaá-Français*. Collge Libermann, Douala.

Paul M Lewis, et al., editors. (2013). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, seventeeth edition.

Makasso, E.-M. (2008). *Intonation et mélismes dans le discours oral spontané en bàsàa*. Ph.D. thesis, Université de Provence (Aix-Marseille 1).

Mochihashi, D., Yamada, T., and Ueda, N. (2009). Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 100–108. Association for Computational Linguistics.

Muthukumar, P. K. and Black, A. W. (2014). Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2613–2617. IEEE.

Nettle, D. and Romaine, S. (2000). *Vanishing Voices*. Oxford University Press Inc., New York, NY, USA.

Neubig, G., Mimura, M., Mori, S., and Kawahara, T. (2010). Learning a language model from continuous speech. In *INTERSPEECH*, pages 1053–1056. Citeseer.

Nguyen, T., Vogel, S., and Smith, N. A. (2010). Nonparametric word segmentation for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 815–823, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Rialland, A., Embanga Aborobongui, G. M., Adda-Decker, M., and Lamel, L. (2015). Dropping of the class-prefix consonant, vowel elision and automatic phonological mining in Embosi. In *Proceedings of the 44th ACAL meeting*, pages 221–230, Somerville. Cascadilla.

Rosenhuber, S. (1908). Die Basa-Sprache. *MSOS*, 11:219–306.

Stahlberg, F., Schlippe, T., Vogel, S., and Schultz, T. (2012). Word segmentation through cross-lingual word-to-phoneme alignment. In *SLT*, pages 85–90. IEEE.

Stahlberg, F., Schlippe, T., Vogel, S., and Schultz, T. (2013). Pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment. In *The 1st International Conference on Statistical Language and Speech Processing*. SLSP 2013.

Stüker, S. and Waibel, A. (2008). Towards human translations guided language discovery for asr systems. In *Proceedings of the First International Workshop on Spoken Languages Technologies for Under-resourced languages (SLTU)*, Hanoi, Vietnam, May.

Stüker, S., Schultz, T., Metze, F., and Waibel, A. (2003). Multilingual articulatory features. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–144. IEEE.

Stüker, S., Besacier, L., and Waibel, A. (2009). Human Translations Guided Language Discovery for ASR Systems. In *10th International Conference on Speech Sci-*

ence and Speech Technology (InterSpeech 2009), pages 1–4, Brighton (UK). Eurasip.

Van de Velde, M. and Ambouroue, O. (2011). The grammar of Orungu proper names. *Journal of African Languages and Linguistics*, 23:113–141.

Vetter, M., Müller, M., Neubig, G., Nakamura, S., Stüker, S., and Waibel, A. (2016). Unsupervised phoneme segmentation on previously unseen languages. In *submitted to SLTU 2016*.

Woodbury, A. C. (2003). Defining documentary linguistics. In Peter K. Austin, editor, *Language Documentation and Description*, volume 1, pages 35–51. London.

Woodbury, A. C. (2011). Language documentation. In Peter K. Austin et al., editors, *The Cambridge Handbook of Endangered Languages*, Cambridge Handbooks in Language and Linguistics, pages 159–186. Cambridge University Press, Cambridge.

Xu, J., Gao, J., Toutanova, K., and Ney, H. (2008). Bayesian semi-supervised Chinese word segmentation for statistical machine translation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1017–1024, Manchester, UK, August. Coling 2008 Organizing Committee.

Corpus collection for under-resourced languages with more than one million speakers

Dirk Goldhahn[1], Maciej Sumalvico[1], Uwe Quasthoff[1,2]
Natural Language Processing Group, University of Leipzig, Germany
Department of African Languages, University of South Africa, South Africa
Email: { dgoldhahn, janicki, quasthoff, }@informatik.uni-leipzig.de

Abstract

For only 40 of about 350 languages with more than one million speakers, the situation concerning text resources is comfortable. For the remaining languages, the number of speakers indicates a need for both corpora and tools. This paper describes a corpus collection initiative for these languages. While random Web crawling has serious limitations, native speakers with knowledge of web pages in their language are of invaluable help. The aim is to give interested scholars, language enthusiasts the possibility to contribute to corpus creation or extension by simply entering a URL into the Web Interface. Using a Web portal URLs of interest are collected with the help of the respective communities. A standardized corpus processing chain for daily newspaper corpora creation is adapted to append newly added web pages to an increasing corpus. As a result we will be able to collect larger corpora for under-resourced languages by a community effort. These corpora will be made publicly available.

Keywords: corpora, under-resourced languages, Web portal, community

1. Introduction

There are about 350 languages with more than one million speakers[1]. For about 40 of them, the situation concerning text resources is comfortable: there are corpora of reasonable size and also tools like POS taggers adapted to these languages. For the remaining languages, the number of speakers indicates a need for both corpora and tools.
The paper describes a corpus collection initiative for these languages. While random Web crawling has serious limitations, native speakers with knowledge of web pages in their language are of invaluable help. The aim is to give interested scholars, language enthusiasts the possibility to contribute to corpus creation or extension by simply entering a URL into the Web Interface. Using a Web portal URLs of interest are collected with the help of the respective communities. A standardized corpus processing chain for daily newspaper corpora creation is adapted to append newly added web pages to an increasing corpus. As a result we will be able to collect larger corpora for under-resourced languages by a community effort. These corpora will be made publicly available directly and as part of the Leipzig Corpora Collection.

2. Crawling strategies and limitations

Random web crawling for smaller languages has several limitations. They are among others related to aspects like the relatively small amount of web pages, the inadequate link structure and the ranking on search engines.

2.1. General crawling problems

The following crawling problem applies to all the following strategies: Due to technical limitations many crawlers cannot follow all links. So-called JavaScript links require the execution of JavaScript code by the crawler which often produces errors. So, if a website heavily uses JavaScript links and there are no other links pointing to special pages (coming from another website, for instance), then all but the main page might be excluded from crawling. At the time of writing (autumn 2015), only the Google crawler is assumed to be able to follow all JavaScript links.
This is a problem especially if the linking density is low, i.e. if the number of static links is not enough to reach most pages. For smaller languages often the web community is in an initial state so complete crawling is difficult.

2.2. Random Web Crawling

The naive algorithm would crawl as much as possible and classify the web pages by language. Modern crawlers like Heritrix[2] [Mohr, 2004] are able to crawl hundreds of millions of pages on commodity hardware, so this should be no problem. The results of such a crawling are even available for direct download: The Common Crawl[3] collected more than 1.8 billion web pages, but the focus is on certain TLDs (Top Level Domains) and on the 40 most prominent languages. Other languages are underrepresented[4].
Language identification in random collections also is more difficult than in restricted collections (see next subsection): The number of pages to be identified in the big collection is comparatively very small, so false positives are a problem. False positives can come from apparently similar languages spoken elsewhere in the

1 http://www.ethnologue.com/

2 https://webarchive.jira.com/wiki/display/Heritrix/Heritrix
3 https://commoncrawl.org/
4 https://docs.google.com/file/d/1_9698uglerxB9nAglvaHkEgU-iZNm1TvVGuCW7245-WGvZq47teNpb_uL5N9/edit

world and from non-text junk documents containing some words of the language, but no actual meaningful text.

2.3. Crawling TLDs

In the simplest case a language is spoken only within one country. Then the natural approach is first to crawl all Web pages of the corresponding top-level domain (TLD) and then to extract the pages in the desired language. The same approach works if a language is spoken in a small number of different countries. Language identification usually also works well [McNamee, 2005] because in most cases the language under consideration is easy to be distinguished from the other languages spoken in the corresponding country.

But for some languages this approach is not productive. Especially in the case of non-official minority languages the community of speakers sometimes prefers another TLD (like .com) instead of the own country's TLD. And in the case of the .com domain, language identification is no more reliable as described above.

2.4. BootCaT approach

In an approach similar to Baroni [2004], frequent terms of a language are combined to form search queries for engines such as Google and to retrieve the resulting URLs as basis for later crawling. As a requirement a small set of frequent terms is needed for each language. Documents such as the Universal Declaration of Human Rights (UDHR), which is available in more than 350 languages, are one possible resource to build such word lists. For an average language, the UDHR contains about 2,000 running words which is still suitable for the task described. An alternative source for word lists are Watchtower documents, which are also available online for about 200 languages[5].

Based on these resources, lists of word tuples of three to five high frequent words are generated. These tuples are then used to query Web search engines and to collect the retrieved URLs. In a next step these Web sites are downloaded and processed further.

Unfortunately, there are certain limitations to the use of this approach for lesser resourced languages. Typically there is a very limited number of resources in the language in question. Since the communities in these languages are small Web sites of interest are typically sparsely linked. Therefore a low page rank score occurs and as a result the Web sites are typically ranked badly on search engines. Experience when using this approach have shown that high ranked results are most likely English Web pages containing few words of the language in question or the short translation of text passages in that language. When checking the Web page for its common language, such URLs are dismissed since English will be detected.

5 https://www.jw.org/

3. Collection method

In this section we propose an alternative method to collect textual resources from the Web for under-resourced languages. The basis of this approach are native speakers with knowledge of Web sites in their respective language. The aim is to give interested scholars, language enthusiasts the possibility to contribute to corpus creation or extension by simply entering a URL into the Web Interface. In order to facilitate the gathering of URLs for a large number of languages a Web portal is currently being developed[6]. Using a basic input mask (see Figure 1) users can easily add URLs together with information on the languages present on this web resource. All additional input fields are optional.

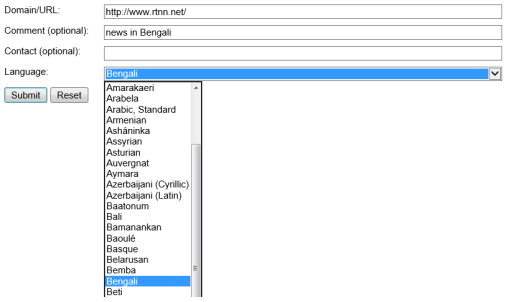

Figure 1: Input mask for URLs of language resources on the Web portal.

The data entered can be viewed online (see Figure 2) and is stored in a local mysql-storage engine where it serves as input for deeper analysis and corpus creation. In a first step parts of a domain are downloaded using Heritrix the crawler of the Internet Archive. Using statistical language identification [Pollmächer, 2012] the languages present in the respective domain are determined. As a data basis for comparison web corpora or documents from sources such as Universal Declaration of Human Rights or Watchtower for several hundred languages are utilized. These sources can include multiple entries for languages using more than one script, enabling the system to create respective corpora.

Bengali (ben)

URL/Domain	Comment
http://www.anandabazar.com/	mostly Bengali
http://www.atnbangla.tv/	TV station
http://www.dainikdestiny.com/	
http://www.jugantor.com	
http://www.rtnn.net/bangla/	news in Bengali
http://www.thedailysangbad.com/	news Website

Figure 2: List view of entries for the Bengali language on the Web portal.

In case the desired language is at least partly present in the documents crawled, in the next step the whole domain is being downloaded. Results of this process are then

6 Available soon at http://small-languages.informatik.uni-leipzig.de

processed utilizing a standardized corpus processing chain for daily newspaper corpora creation[7] which has been adapted to append newly added web pages to an increasing corpus for each language. As a result we will be able to collect larger corpora for under-resourced languages by a community effort. These corpora will be made publicly available.

The Web portal and the adapted processing chain are currently in development. Both will be finished in March 2016.

4. Corpus Processing Chain

We apply a standardized, language-independent pipeline for building corpora from raw data. For details on this processing chain please see Goldhahn (2012). We use own tools for extracting raw text from WARC files (Heritrix output) and HTML pages. Then we apply statistical language identification on document basis. Further processing steps are: sentence segmentation, removal of ill-formed sentences based on handwritten regular expressions, language identification on sentence basis, duplicate sentence removal (a removal of near duplicates such as boilerplates is currently in development), tokenization and word co-occurrence calculation. Finally, the corpora are stored as MySQL databases with a standardized schema. In addition to the basic workflow, additional (possibly language-specific) tools can be applied to some corpora, like POS-tagging, which results in additional database tables.

A couple of technical issues must be taken care of in multilingual processing. As we are using UTF-8 as the sole encoding, proper conversion must be guaranteed at the preprocessing step (HTML/WARC → text). The sentence separator is a rule based tool, which requires a list of sentence-terminating characters. It is important to include such characters for all expected languages and writing systems. Pairs of characters that look similar, but are encoded differently, like Latin semicolon (U+003B) and Greek question mark (U+037E), need special attention. For language segmentation, lists of around 1K most frequent words for each language need to be supplied.

Tests on various input data have shown that our processing chain handles data volumes of up to 200 million sentences. For corpora of 100K - 1M sentences, the running times are typically less than an hour.

5. Conclusion

This paper describes a corpus collection initiative for lesser resourced languages, enabling scholars or language enthusiasts to create and extend corpora for these languages by simply entering a URL. Using this Web portal URLs of interest are collected with the help of the respective communities. As a result we will be able to collect larger corpora for under-resourced languages by a community effort. These corpora will be made publicly available.

7 http://wortschatz.uni-leipzig.de/wort-des-tages/

6. Bibliographical References

Baroni, M., & Bernardini, S. (2004, May). BootCaT: Bootstrapping Corpora and Terms from the Web. In LREC.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In LREC (pp. 759-765).

McNamee, P. (2005). Language identification: a solved problem suitable for undergraduate instruction. Journal of Computing Sciences in Colleges, 20(3), 94-101.

Mohr, G., Stack, M., Rnitovic, I., Avery, D., & Kimpton, M. (2004, July). Introduction to heritrix. In 4th International Web Archiving Workshop.

Pollmächer, J. (2011). Separierung mit FindLinks gecrawlter Texte nach Sprachen. Bachelor Thesis, University of Leipzig.

7. Appendix: Language Lists

The following lists are collected as follows: The languages with more than 1 million of speakers (according to Ethnologue[8]) are divided into two parts: The Leipzig Corpora Collection[9] [Goldhahn et al., 2012] is used to identify under-resourced languages. For simplicity, a language is called under-resourced if there are less than 1 million sentences in the corpora of this collection. The following two tables show both the under-resourced languages and the well-resourced languages with more than 1 million sentences in the collection.

Language	*Code*	*Country*
Dari	prs	Afghanistan
Hazaragi	haz	Afghanistan
Albanian	sqi	Albania
Kabyle	kab	Algeria
Tachawit	shy	Algeria
Kimbundu	kmb	Angola
Umbundu	umb	Angola
Armenian	hye	Armenia
Bengali	ben	Bangladesh
Chittagonian	ctg	Bangladesh
Rangpuri	rkt	Bangladesh
Sylheti	syl	Bangladesh
Vlaams	vls	Belgium
Fon	fon	Benin
Aymara	aym	Bolivia
Quechua	que	Bolivia, Peru
Bosnian	bos	Bosnia and Herzegovina
Hunsrik	hrx	Brazil
Jula	dyu	Burkina Faso

8 http://www.ethnologue.com/

9 http://corpora.informatik.uni-leipzig.de/

Mòoré	mos	Burkina Faso
Rundi	run	Burundi
Central Khmer	khm	Cambodia
Fulah	ful	Cameroon
Kabuverdianu	kea	Cape Verde Islands
Bouyei	pcc	China
Chuanqiandian Cluster Miao	cqd	China
Gan Chinese	gan	China
Hmong	hmn	China
Hmong Daw	mww	China
Khams Tibetan	khg	China
Min Dong Chinese	cdo	China
Min Nan Chinese	nan	China
Northern Qiandong Miao	hea	China
Nuosu	iii	China
Southern Dong	kmc	China
Tibetan	bod	China
Uyghur	uig	China
Zhuang	zha	China
Kituba	mkw	Congo
Baoulé	bci	Côte d'Ivoire
Dan	dnj	Côte d'Ivoire
Alur	alz	Dem. Rep. of Congo
Chokwe	cjk	Dem. Rep. of Congo
Kituba	ktu	Dem. Rep. of Congo
Kongo	kon	Dem. Rep. of Congo
Koongo	kng	Dem. Rep. of Congo
Lingala	lin	Dem. Rep. of Congo
Luba-Kasai	lua	Dem. Rep. of Congo
Luba-Katanga	lub	Dem. Rep. of Congo
Ngbaka	nga	Dem. Rep. of Congo
Songe	sop	Dem. Rep. of Congo
Yombe	yom	Dem. Rep. of Congo
Zande	zne	Dem. Rep. of Congo
Tigré	tig	Eritrea
Afar	aar	Ethiopia
Amharic	amh	Ethiopia
Gamo	gmv	Ethiopia
Hadiyya	hdy	Ethiopia
Oromo	orm	Ethiopia
Sidamo	sid	Ethiopia
Wolaytta	wal	Ethiopia
Tigrigna	tir	Ethiopia, Eritrea
Occitan	oci	France
Abron	abr	Ghana
Akan	aka	Ghana
Éwé	ewe	Ghana

Pontic	pnt	Greece
Quiché	quc	Guatemala
Eastern Maninkakan	emk	Guinea
Fang	fan	Guinea
Kpelle	kpe	Guinea
Pular	fuf	Guinea
Susu	sus	Guinea
Haitian	hat	Haiti
Ahirani	ahr	India
Assamese	asm	India
Awadhi	awa	India
Bagheli	bfy	India
Bhili	bhb	India
Bhojpuri	bho	India
Bodo	brx	India
Bundeli	bns	India
Chhattisgarhi	hne	India
Deccan	dcc	India
Dhundari	dhd	India
Dogri	doi	India
Garhwali	gbm	India
Garo	grt	India
Goan Konkani	gom	India
Godwari	gdx	India
Gondi	gon	India
Haryanvi	bgc	India
Ho	hoc	India
Kanauji	bjj	India
Kangri	xnr	India
Kannada	kan	India
Kashmiri	kas	India
Konkani	knn	India
Kumaoni	kfy	India
Kurux	kru	India
Lambadi	lmn	India
Magahi	mag	India
Mahasu Pahari	bfz	India
Maithili	mai	India
Malayalam	mal	India
Marwari	mwr	India
Meitei	mni	India
Mina	myi	India
Mundari	unr	India
Nimadi	noe	India
Oriya	ori	India
Rajasthani	raj	India
Sadri	sck	India
Santali	sat	India
Shekhawati	swv	India

Surgujia	sgj	India
Surjapuri	sjp	India
Tamil	tam	India
Telugu	tel	India
Tulu	tcy	India
Varhadi-Nagpuri	vah	India
Vasavi	vas	India
Jambi Malay	jax	Indonesia
Bali	ban	Indonesia (Java and Bali)
Betawi	bew	Indonesia (Java and Bali)
Javanese	jav	Indonesia (Java and Bali)
Madura	mad	Indonesia (Java and Bali)
Sunda	sun	Indonesia (Java and Bali)
Banjar	bjn	Indonesia (Kalimantan)
Sasak	sas	Indonesia (Nusa Tenggara)
Bugis	bug	Indonesia (Sulawesi)
Gorontalo	gor	Indonesia (Sulawesi)
Makasar	mak	Indonesia (Sulawesi)
Aceh	ace	Indonesia (Sumatra)
Batak Dairi	btd	Indonesia (Sumatra)
Batak Mandailing	btm	Indonesia (Sumatra)
Batak Simalungun	bts	Indonesia (Sumatra)
Batak Toba	bbc	Indonesia (Sumatra)
Minangkabau	min	Indonesia (Sumatra)
Musi	mui	Indonesia (Sumatra)
Bakhtiari	bqi	Iran
Domari	rmt	Iran
Gilaki	glk	Iran
Iranian Persian	pes	Iran
Kashkay	qxq	Iran
Laki	lki	Iran
Mazanderani	mzn	Iran
Northern Luri	lrc	Iran
Southern Kurdish	sdh	Iran
Central Kurdish	ckb	Iraq
Eastern Yiddish	ydd	Israel
Dholuo	luo	Kenya
Ekegusii	guz	Kenya
Gikuyu	kik	Kenya
Kalenjin	kln	Kenya
Kamba	kam	Kenya
Kimîîru	mer	Kenya
Kipsigis	sgc	Kenya

Lubukusu	bxk	Kenya
Maasai	mas	Kenya
Oluluyia	luy	Kenya
Kurdish	kur	Kurdistan, Iraq, Turkey
Kyrgyz	kir	Kyrgyzstan
Lao	lao	Laos
Macedonian	mkd	Macedonia
Malagasy	mlg	Madagascar
Nyanja	nya	Malawi
Tumbuka	tum	Malawi
Yao	yao	Malawi
Malay	zlm	Malaysia (Peninsular)
Bamanankan	bam	Mali
Maasina Fulfulde	ffm	Mali
Soninke	snk	Mali
Hassaniyya	mey	Mauritania
Halh Mongolian	khk	Mongolia
Central Atlas Tamazight	tzm	Morocco
Tachelhit	shi	Morocco
Tarifit	rif	Morocco
Lomwe	ngl	Mozambique
Makhuwa	vmw	Mozambique
Makhuwa-Meetto	mgh	Mozambique
Sena	seh	Mozambique
Tswa	tsc	Mozambique
Burmese	mya	Myanmar
Pwo Eastern Karen	kjp	Myanmar
Rohingya	rhg	Myanmar
S'gaw Karen	ksw	Myanmar
Shan	shn	Myanmar
Ndonga	ndo	Namibia
Eastern Tamang	taj	Nepal
Nepali	nep	Nepal
Limburgish	lim	Netherlands
Tamashek	tmh	Niger
Zarma	dje	Niger
Anaang	anw	Nigeria
Berom	bom	Nigeria
Central Kanuri	knc	Nigeria
Ebira	igb	Nigeria
Edo	bin	Nigeria
Hausa	hau	Nigeria
Ibibio	ibb	Nigeria
Igbo	ibo	Nigeria
Izon	ijc	Nigeria
Kanuri	kau	Nigeria
Nigerian Fulfulde	fuv	Nigeria
Nigerian Pidgin	pcm	Nigeria

Tiv	tiv	Nigeria
Yoruba	yor	Nigeria
Norwegian	nor	Norway
Baluchi	bal	Pakistan
Brahui	brh	Pakistan
Eastern Balochi	bgp	Pakistan
Lahnda	lah	Pakistan
Northern Hindko	hno	Pakistan
Pahari-Potwari	phr	Pakistan
Seraiki	skr	Pakistan
Sindhi	snd	Pakistan
Southern Balochi	bcc	Pakistan
Western Balochi	bgn	Pakistan
Western Panjabi	pnb	Pakistan
Guarani	grn	Paraguay, Bolivia
Bikol	bik	Philippines
Cebuano	ceb	Philippines
Central Bikol	bcl	Philippines
Filipino	fil	Philippines
Hiligaynon	hil	Philippines
Ilocano	ilo	Philippines
Maguindanao	mdh	Philippines
Pampangan	pam	Philippines
Pangasinan	pag	Philippines
Tagalog	tgl	Philippines
Tausug	tsg	Philippines
Waray-Waray	war	Philippines
Romany	rom	Romania
Bashkort	bak	Russian Federation
Chechen	che	Russian Federation
Chuvash	chv	Russian Federation
Kabardian	kbd	Russian Federation
Tatar	tat	Russian Federation
Rwanda	kin	Rwanda
Mandingo	man	Senegal
Mandinka	mnk	Senegal
Pulaar	fuc	Senegal
Serer-Sine	srr	Senegal
Wolof	wol	Senegal
Mende	men	Sierra Leone
Themne	tem	Sierra Leone
Somali	som	Somalia
Northern Sotho	nso	South Africa
Southern Ndebele	nbl	South Africa
Tsonga	tso	South Africa
Venda	ven	South Africa
Xhosa	xho	South Africa
Zulu	zul	South Africa
Tswana	tsn	South Africa, Botswana
Southern Sotho	sot	South Africa, Lesotho
Swati	ssw	South Africa, Swaziland
Galician	glg	Spain
Sinhala	sin	Sri Lanka
Bedawiyet	bej	Sudan
Dinka	din	Sudan
Tajiki	tgk	Tajikistan
Gogo	gog	Tanzania
Haya	hay	Tanzania
Makonde	kde	Tanzania
Nyakyusa-Ngonde	nyy	Tanzania
Sukuma	suk	Tanzania
Swahili	swa	Tanzania
Northern Khmer	kxm	Thailand
Thai	tha	Thailand
Malay	msa	Thailand, Malaysia
Dimli	diq	Turkey
Zaza	zza	Turkey
Turkmen	tuk	Turkmenistan
Acholi	ach	Uganda
Chiga	cgg	Uganda
Ganda	lug	Uganda
Lango	laj	Uganda
Lugbara	lgg	Uganda
Masaaba	myx	Uganda
Nyankore	nyn	Uganda
Soga	xog	Uganda
Teso	teo	Uganda
Uzbek	uzb	Uzbekistan
Muong	mtq	Viet Nam
Tày	tyz	Viet Nam
Bemba	bem	Zambia
Tonga	toi	Zambia, Zimbabwe
Manyika	mxc	Zimbabwe
Ndau	ndc	Zimbabwe
Ndebele	nde	Zimbabwe
Shona	sna	Zimbabwe

Table 1: Under-resourced languages with more than 1 million speakers and less than 1 million sentences, ordered by country.

Language	*Code*	*Country*
Arabic	ara	various countries
English	eng	various countries
Pushto	pus	Afghanistan, Pakistan
Azerbaijani	aze	Azerbaijan
Belarusan	bel	Belarus
Bulgarian	bul	Bulgaria
Chinese	zho	China
Serbo-Croatian	hbs	Croatia, Serbia, Bosnia and Herzegovina
Czech	ces	Czech Republic
Danish	dan	Danmark
Estonian	est	Estonia
Finnish	fin	Finland
French	fra	France
Georgian	kat	Georgia
Greek	ell	Greece
Hungarian	hun	Hungaria
Gujarati	guj	India
Hindi	hin	India
Marathi	mar	India
Indonesian	ind	Indonesia
Gilaki	glk	Iran
Persian	fas	Iran
Hebrew	heb	Israel
Italian	ita	Italy
Japanese	jpn	Japan
Kazakh	kaz	Kazakhstan
Korean	kor	Korea
Latvian	lav	Latvia
Lithuanian	lit	Lithuania
Mongolian	mon	Mongolia
Dutch	nld	Netherlands
Polish	pol	Poland
Portuguese	por	Portugal
Romanian	ron	Romania
Russian	rus	Russia
Serbian	srp	Serbia
Slovak	slk	Slovakia
Slovene	slv	Slovenia
Afrikaans	afr	South Africa
Catalan-Valencian-Balear	cat	Spain
Spanish	spa	Spain
Tamil	tam	Sri Lanka, India
Swedish	swe	Sweden
Turkish	tur	Turkey
Ukrainian	ukr	Ukraine
Urdu	urd	Urdu
Vietnamese	vie	Vietnam

Table 2: Well-resourced languages with more than 1 million sentences, ordered by country.

Building Intelligent Digital Assistants for Speakers of a Lesser-Resourced Language

Dewi Bryn Jones[1], Sarah Cooper[2]

[1]Language Technologies Unit, [2]School of Linguistics and English Language
Bangor University, Bangor, Wales, UK
E-mail: {d.b.jones, s.cooper}@bangor.ac.uk

Abstract

This paper reports on the work to develop intelligent digital assistants for speakers of a lesser-resourced language, namely Welsh. Such assistants provided by commercial vendors such as Apple (Siri), Amazon (Alexa), Microsoft (Cortana) and Google (Google Now) are allowing users increasingly to speak in natural English with their devices and computers in order to complete tasks, obtain assistance and request information. We demonstrate how these systems' architectures do not provide the means for external developers to build intelligent speech interfaces for additional languages, which, in the case of less resourced languages, are likely to remain unsupported. Consequently we document how such an obstacle has been tackled with open alternatives. The paper highlights how previous work on Welsh language speech recognition were improved, utilized and integrated into an existing open source intelligent digital assistant software project. The paper discusses how this work hopes to stimulate further developments and include Welsh and other lesser-resourced languages in as many developments of intelligent digital assistants as possible.

Keywords: intelligent digital assistants, speech recognition, lesser resourced languages, Welsh

1 Introduction

It is increasingly possible, as a consequence of recent advancements in speech recognition, machine translation and natural language processing and understanding, for users to engage with their devices and computers. They do this via intelligent speech interfaces in order to command and control as well as to receive answers to questions voiced in natural language.

There are four main commercial platforms driving this change, namely Apple Siri, Google Now, Microsoft Cortana and Amazon Alexa. To date these provide their powerful capabilities in English and to a lesser extent some other major languages. There is little evidence so far that they are likely to extend their choice of languages to the long tail of smaller languages, including Welsh, in the near future. Furthermore there are no means for external developers to adapt these systems for any new language.

Indeed languages with smaller numbers of speakers often find themselves lagging in digital innovation including language technologies. As such they are lesser-resourced with regard to the availability and interest in funding. However the Welsh Government through its Welsh Language Technology and Digital Media Fund have since 2012 have followed a strategy to develop 'more tools and resources… to facilitate the use of Welsh, including in the digital environment" (Welsh Government, 2012; 45) as well as develop "new Welsh language software applications and digital services" (Welsh Government, 2013; 12).

With funding from the Welsh Government as well as S4C, (the Welsh-language public service television channel), we have established a project called 'Seilwaith Cyfathrebu Cymraeg' (*Welsh Language Communications Infrastructure*) (Jones and Ghazali, 2016). This project aims to ensure that Welsh language users are not excluded from continued developments in human computer interaction by improving current Welsh speech recognition and applying its capabilities in a prototype Welsh language intelligent digital assistant application. The project is limited to 8 months in duration due to the initial funding programme.

In addition the project will make all resources and software available under very permissive open-source licenses via the Welsh National Language Technologies Portal infrastructure (Prys and Jones, 2015). This provides unrestricted usage to developers involved in commercial, education and volunteer activities within the lesser resourced language community. It also provides unrestricted usage for global companies who wish to extend their range of languages supported in any multilingual intelligent digital assistants.

2 Approach

The 'Seilwaith Cyfathrebu Cymraeg' project initially evaluated the four main commercial intelligent digital assistant platforms that are responsible for popularising a new mode of human computer interaction (Apple Siri, Amazon Alexa, Microsoft Cortana and Google Now) Each platform is complemented by APIs (Application Programming Interfaces) and SDKs (Software Development Kits) that each company is eager for developers to utilise in their commercial products and services. This enables each platform to have its capabilities extended and presence widened into third party apps and products. We initially investigated whether these systems would provide an opportunity to extend the range of supported language (Ghazali et al., 2015)

A generic software architecture and flow of events was realised during the investigation as seen in Figure 1.

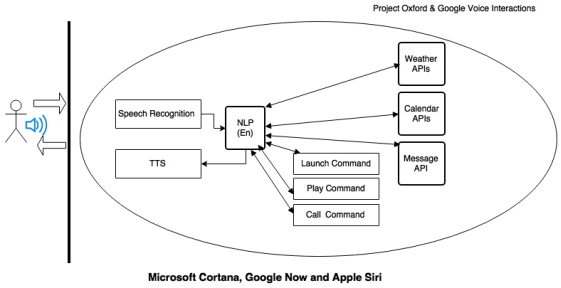

Figure 1 - Generic Architecture of an Intelligent Digital Assistant

A user speaks in natural language to the device or computer:

1. A speech recognition engine captures the audio and converts the user's spoken wish into text. The text is handed over to a natural language processing (NLP) component. For example the user asks: 'DO I HAVE ANY MEETINGS TODAY?'
2. The NLP component understands from the text what the user intends or wants and thus identifies which component, module or API is able to fulfil the request. For example, the NLP component may recognize a question from 'DO I HAVE' and recognizes 'MEETINGS' and 'TODAY' as keywords associated with time and calendar.
3. The NLP component communicates with the identified module or API according to its interface specification. For example it constructs and sends a message to the Calendar API:

   ```
   getEvents(type=meeting,date=today)
   ```

4. In cases where an answer is required, the NLP accepts a response from the obliging module or API and generates a sentence with results included. For example it may receive a result in a format such as JSON:

   ```
   {"success": true, "events": [
   {"date": "2015-09-25",
   "time":"10:30am", "name":"Meeting to
   discuss a new Welsh language
   project","location":"Bangor
   University, Ogwen Building, Room
   234"},{"date": "2015-09-25",
   "time":"12:30pm", "name":"Lunch with
   Delyth","location":"Terrace
   Restaurant, Bangor University"} ]}
   ```

   From which the NLP constructs the sentence:

   Yes you do. At 10:30 this morning you have a meeting to discuss a new Welsh language project in Bangor University, Ogwen Building, Room 234. Then at 12:30 you have lunch with Delyth in the Terrace Restaurant, Bangor University.

5. The natural language sentence result is handed to a text to speech engine for voicing back to the user. For example:

   "*Yes you do. At half past ten this morning you have a meeting to discuss a new Welsh language project in Bangor University, Ogwen Building, Room two hundred and thirty four. Then at half past twelve you have lunch with Delyth in the Terrace Restaurant, Bangor University.*"

We became aware that these commercial architectures have their speech recognition and natural processing components encapsulated into one super-component. As a consequence they only provided access in the language that the speech recognition component supports. This linguistic limitation may be necessary for a functional consistency but does not allow for external developers to integrate support for additional languages with the aid of alternative language technologies that could still leverage the capabilities of the commercial offerings.
A number of other intelligent personal assistant platforms exist. Their suitability as a basis for building a Welsh language digital assistant is feasible only if their architectures are more granular and open and which can fulfil the following criteria:

- A Welsh language speech recognition engine can be integrated
- The NLP for understanding texts from voiced requests can be either adapted or replaced
- Responses can be provided via Welsh language text to speech
- APIs and modules that implement capabilities and fullfill tasks but which are based in English language usage can still be included with novel integration of Welsh to English in requests and English to Welsh machine translation in responses

Figure 2 illustrates a desirable architecture.

The best open alternative candidate was found to be Jasper (Marsh and Saha, 2014). Jasper is a very simple Python application which already integrates a number of speech recognition and text-to-speech engines and provides an easy mechanism for developers to extend its capabilities via simple addition of modules written as Python scripts. Jasper is able to run completely locally on small computers such as Raspberry Pis.

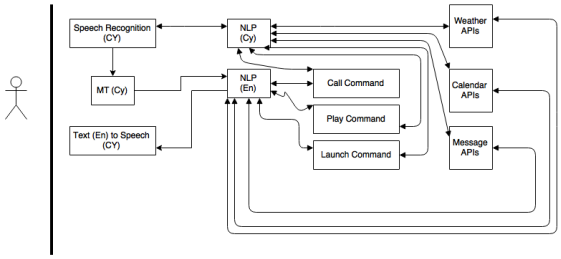

Figure 2 - Desirable Architecture for Building a Welsh Language Intelligent Personal Assistant

Also worthy of consideration, due to the availability of source code and/or sufficient modularity and openness of APIs, were SIRIUS (Hauswald et al., 2015) and Wit.ai (wit.ai, n.d). They contain more complex language technologies and are considerably more complex.

These open alternatives are able to support an incremental development strategy. Initially Jasper can be used for developing a simple speech recognition system in conjunction with an elementary intent evaluator supporting a limited number of domains in Welsh (such as asking the time, news or weather). Subsequently more sophisticated and intelligent capabilities can be developed through further iterations.

Such an approach is compatible to the work to improve the Welsh language speech recognition component and extend its vocabulary and grammar complexity in order to support a growing number of possible requests and questions.

3 Speech Recognition Improvement

Prior to the 'Seilwaith Cyfathrebu Cymraeg' project, a significant amount of work on Welsh language speech recognition had been done in the GALLU (Gwaith Adnabod Lleferydd Uwch, translation: Further Speech Recognition Work) project which built on earlier work on basic speech recognition project during 2008-9 (Cooper et al., 2014). Some earlier work on speech technology work had also been conducted under the WISPR (Welsh Irish Speech Processing Resources) project (Prys et. al., 2004).

The GALLU project succeeded in developing new letter to sound rules, a collection of prompts covering all Welsh language phonemes and subsequently a speech corpus (Cooper et. al., 2015) collected via a crowd sourcing iOS and Android app called Paldaruo (Jones, 2015). Elements of the corpus were successfully used to train HTK acoustic models and their inclusion in an open source speech recognition decoder engine (Julius) that could control the movement of a toy robotic arm (Aonsquared, n.d). The speech recognition component of the 'Seilwaith Cyfathrebu Cymraeg' project would be a continuation of this previous work and means that Julius and HTK would continue to be the basis for developing Welsh language speech recognition.

Improving speech recognition for Welsh would mainly consist of training new acoustic models for all Welsh language phones from the entire Paldaruo speech corpus. The corpus had grown in size since 2014 to contain contributions from 410 speakers. This was in contrast to training during the GALLU project where data from only 20 speakers was used along with a subset of phones. We improved the training scripts from the GALLU project and packaged them with the HTK into a user friendly and portable environment for speech recognition development using Docker. Docker allows for acoustic model training to be easily shared and consistently reproduced amongst other researchers and developers. Scripts were added for downloading the Paldaruo Speech Corpus, prompts text, speaker metadata (e.g. age and accent) as well as a Welsh pronunciation lexicon. We also added scripts that package the acoustic models for use in decoders such as Julius.

At the start of the GALLU project we had not anticipated the automated testing of acoustic models and as such had not collected extra recordings from each speaker for a test corpus. It was left to this project and the Docker based HTK training environment to follow a widely considered 'bad practice' of using training data as test data. A word loop grammar was used so that a test could expect any word after a given word. However this approach was beneficial in evaluating and improving the acoustic models. An initial reproduction and testing of the robotic arm control application from GALLU using the new Docker HTK and scripts environment (Iteration 0, Figure 3,4) showed that the approach was useful in validating contributions from speakers.

Iteration	Speakers	Description
0	20	All contributions used in training GALLU robotic arm control
1	410	All contributions
2	177	All contributions noted as Central and South Wales accents.
3	210	All contributions noted as North Wales accents
4	1	Recording from one contributor validated and verified by ear
5	35	All complete contributions (160 speakers) tested individually to have a Word Accuracy above 70%
6	88	All contributions (410 speakers) tested individually and seen to have word accuracy above 70%
7	88	As 6 but with an improved clustered triphone question for Welsh (tree1.hed)

Figure 3 - Acoustic Model Training Iterations

Iteration	Word Accuracy	Sentence Accuracy
0	90%	42%
1	20%	0%
2	26%	0%
3	24%	0%
4	75%	7%
5	92%	24%
6	91%	19%
7	92%	20%

Figure 4 - HTK Results on accuracy for each iteration

Thus we began by training new acoustic models with all recordings from all speakers (Iteration 1, Figure 3,4). Initially scores were disappointing (Word Accuracy 20%). While attempts at distinguishing between accents did improve word accuracy scores to some extent (Iteration 2, 3 Figure 3,4), we did not deem it sufficient enough to base further iterations on a partitioned speech corpus.

It could be argued that this was to be expected given that the speech corpus was crowd sourced and we had little control on the quality of contributions. Furthermore there could be no human involvement in quality assurance and verification of recordings due to the resources available (given that audio files numbered into their thousands).

However an experiment to train and test acoustic models based on one speaker's contribution, along with amendments to the training and testing scripts, provided a breakthrough which, according to our testing strategy, provided much improved word and sentence accuracy scores (Iteration 4, Figure 3, 4). Subsequently we trained and tested every contribution individually to obtain word accuracy scores for each speaker. This resulted in a wide variety of scores across speakers. All contributions found to have word accuracy scores above 70% were considered better quality.

Of the 410 individuals who had used the Paldaruo app, only 136 had recorded all 43 prompts. We assessed these contributions on an individual basis and 35 of these were found to have a word accuracy score above 70%. When combined together to create speaker independent acoustic models, the word and sentence accuracy improved to over 90% (Iteration 5 Figure 3, 4).

When considering all contributions, regardless of the number of prompts recorded, we found 88 speakers had word accuracy scores higher than 70%. When combined together to create speaker independent acoustic models there was a slight decrease in comparison to the previous iteration in both word and accuracy scores (Iteration 6, Figure 3, 4).

We improved our HTK decision tree clustering script file for Welsh (the language specific tree.hed file) which groups all Welsh phonemes according to their acoustic classes. Our results consequently improved word and sentence accuracy results by 1% (Iteration 7, Figure 3, 4).

4 An Early Prototype Welsh-language Intelligent Digital Assistant

With acoustic models deemed sufficient in quality, the project was able to implement its first iteration of a Welsh language intelligent digital assitant which would support answering questions and fulfill tasks in the domains of news, weather, time, music, proverbs and jokes.

We developed simple grammar and vocabulary files for Julius in order to produce the first release of a Welsh language speech recognition component - julius-cy (Jones and Cooper, 2016). It aims to recognise all possible means of requesting information (such as news, weather and time) in complete and natural sentences. For example:

BETH YDY'R TYWYDD HEDDIW?
What's today's weather?

BETH YW TYWYDD YFORY?
What's tomorrow's weather?

BETH YW'R NEWYDDION?
What's the news?

FAINT O'R GLOCH YDY HI?
What time is it?

CHWARAEA GERDDORIAETH CYMRAEG
Play Welsh music

julius-cy contains scripts and documentation that make the whole process of installation very easy on a Linux based machine such as Raspberry Pi. This could be deemed difficult for a non-specialist given the complexity of the packages and configuration required for the acoustic models, pronunciation lexicons as well as any necessary grammar and vocabulary files. Further scripts make it possible for julius-cy to recognise users' own additions to the grammar and vocabulary files.

The open source project Jasper (Marsh and Saha, 2014) was used as our platform for applying julius-cy in a complete intelligent digital assistant solution. The persona name 'Macsen' was chosen as the wake up word that would instruct the Jasper system to alternate between passive and active listening modes. Further, but minimal, alterations were necessary to permit Jasper to support interaction in languages other than English. Modules written in simple Python were added to integrate various Welsh language websites such as Golwg360 News and S4C Weather. To voice responses to user requests 'Macsen' uses either a Welsh language Festival based text-to-speech voice (Prys et al., 2004) or the more naturally sounding 'Geraint' voice from Ivona's Speech Cloud.

5 Further Work and Conclusions

Much work remains on developing a Welsh language intelligent digital assistant. We are releasing all models, scripts, code and data developed by the project via the Welsh National Language Technologies Portal (Prys and Jones, 2015) as well as GitHub (Jones, 2016a; Jones, 2016b), according to the permissive MIT open source license which allows the widest possible outreach to other developers involved in commercial, education and volunteer activities for Welsh and other languages. We welcome all feedback and pull requests for extending its capabilities and improvements.

We intend for julius-cy to support large vocabulary continuous speech recognition by utilising the large text resources available to us, including the 30 million word Cysill Ar-lein corpus (Prys, Prys & Jones, 2016, forthcoming), to produce language models. This would allow opportunities for Macsen to support more domains and intelligent capabilities.

One aim is to investigate applying machine translation to translate texts recognized by julius-cy, from Welsh to

English in order to consume English medium APIs provided by wit.ai and/or SIRIUS. This allows Welsh speaking users to still use intelligent capabilities that are rooted in the English language.

The work on improving acoustic models in the meantime will continue. Additional funding was secured recently from Bangor University's Undergraduate Internship Scheme to employ a student to recruit and collect recordings from Welsh speaking staff and students in order to expand the amount of quality assured contributions.

Further work on acoustic and language modelling for Welsh are to be the subject of new KESS (Knowledge Economy Skills Scholarships) PhD programmes in partnership with the Welsh Government which will explore developing speech technologies using more recent developments in deep learning and neural network approaches.

Many in the Welsh language community recognise the opportunities and risks posed by the growing number of services and apps that provide intelligent capabilities via speech interfaces. Other similar language communities also recognise the benefit of creating language resources in this domain.

However, lesser resourced languages are required to be innovative in attracting funding. Certain funders desire useful end products and applications with wide reaching public impact. This skews longer term aims for research and development of basic language technologies and resources such as speech recognition for lesser resourced languages. There is a danger of users relying on widely available English language products and services which will impact upon language use, vitality and diversity.

We hope that this work contributes to further development of intelligent digital assistants for lesser resourced languages as well as stimulating developments in the wider industry.

6 Acknowledgements

The 'Seilwaith Cyfathrebu Cymraeg' project reported on in this paper was 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 project.

7 Bibliographic References

AonSquared. [No date]. *Speech recognition using the Raspberry Pi.* Available at: http://aonsquared.co.uk/raspi_voice_control [Accessed: 16 February 2016]

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

Cooper, S., Chan, D., Jones, D.B. (2015). *Corpws Lleferydd Paldaruo.* [http://techiaith.cymru/corpora/Paldaruo]

Ghazali, S. , Jones D. B., Prys D. (2015). *Towards a Welsh Language Intelligent Personal Assistant: A Brief Study of APIs for Spoken Commands, Question and Answer Systems and Text to Speech. Report for the Welsh Government*. Available at : http://techiaith.bangor.ac.uk/towards-a-welsh-language-intelligent-personal-assistant/?lang=en [Accessed: 23 March 2016]

Hauswald, J., Laurenzano, M. A., Zhang, Y., Li, C., Rovinski, A., Khurana, A., Dreslinski, R., Mudge, T., Petrucci, V., Tang, L. & Mars, J. (2015) Sirius: An Open End-to-End Voice and Vision Personal Assistant and Its Implications for Future Warehouse Scale Computers. In *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, New York, NY: ACM.

HTK. [No date]. *Hidden Markov Toolkit.* Available at: http://htk.eng.cam.ac.uk/ [Accessed: 16 February 2016]

Jones D.B. (2015) *Paldaruo – an iOS app for crowdsourcing speech data.* Available at https://github.com/techiaith/Paldaruo [Accessed: 16 February 2016]

Jones D.B. (2016a) *Macsen – A Welsh Language Intelligent Digital Assistant*. Available at https://github.com/techiaith/macsen [Accessed: 16 February 2016]

Jones D.B. (2016b) *An easy to use speech recognition development toolkit using HTK, Julius and Docker* Available at https://github.com/techiaith/seilwaith [Accessed: 23 March 2016]

Jones D.B. and Cooper, S. (2016) *julius-cy – Welsh language speech recognition with Julius,*Available at: https://github.com/techiaith/julius-cy [Accessed: 16 February 2016]

Jones D.B., Ghazali, S. (2016) *Project Seilwaith Cyfathrebu Cymraeg Website*. Available at http://techiaith.bangor.ac.uk/seilwaith-cyfathrebu-cymraeg/ [Accessed: 16 February 2016]

Marsh, C. and Saha, S. (2014) *Jasper Documentation*, Available at: http://jasperproject.github.io [Accessed: 16 February 2016]

MIT. [No date]. *The MIT License*. Available at: http://opensource.org/licenses/mit-license.html [Accessed: 16 February 2016]

Prys, D., and Jones D. B. (2015). National Language Technology Portals for LRLs: A Case Study. Paper presented at *Language Technologies in Support of Less-Resourced Languages, (LRL 2015)* 28 November 2015, Poznan, Poland.

Prys, D., Williams, B., Hicks, B., Jones, D., Ní Chasaide, A., Gobl, C., Carson- Berndsen, J., Cummins, F., Ní Chiosáin, M., McKenna, J., Scaife, R. and Uí Dhonnchadha, E. (2004). WISPR: Speech Processing Resources for Welsh and Irish. In Proceedings of the *Pre-Conference Workshop on First Steps for Language Documentation of Minority Languages*, LREC Conference, Lisbon, Portugal.

Prys, D; Prys, G; & Jones, D.B. (2016, forthcoming) Cysill Ar-lein: A Corpus of Written Contemporary Welsh compiled from an on-line Spelling and Grammar Checker. In *Proceedings of the 10th International*

Conference of Language Resources and Evaluation, LREC 2016. Portorož, Slovenia.

Welsh Government (2012). A living language, a language for living. Available at: http://wales.gov.uk/docs/dcells/publications/122902wls201217en.pdf [Accessed: 25 March 2016].

Welsh Government. (2013). Welsh language Technology and Digital Media Action Plan. Available at: http://wales.gov.uk/docs/dcells/publications/230513-action-plan-en.pdf [Accessed: 25 March 2016].

wit.ai [No date] *wit.ai: Natural Language for Developers.* Available at: https://wit.ai [Accessed: 16 February 2016]

Multiword Expressions for Capturing Stylistic Variation Between Genders in the Lithuanian Parliament

Justina Mandravickaitė[1,2], Michael Oakes[3]
[1]Faculty of Philology, Vilnius University, Vilnius, Lithuania
[2]Baltic Institute of Advanced Technology, Vilnius, Lithuania
[3]Research Institute of Information and Language Processing, University of Wolverhampton, Wolverhampton, United Kingdom

E-mail: justina@bpti.lt, Michael.Oakes@wlv.ac.uk

Abstract

The relation between gender and language has been studied by many authors, but there is no general agreement regarding gender influence on language usage in the professional environment. This could be because in most of the studies data sets are too small or texts of individual authors are too short in order to capture differences of language usage according to gender successfully. This study draws on a larger corpus of transcribed speeches in the Lithuanian Parliament (1990-2013) to explore gender differences in a language with a setting of political debates using stylometric analysis. The experimental set up consists of multiword expressions as features (formulaic language can allow a more detailed interpretation of the results in comparison to character n-grams or even most frequent words) combined with unsupervised machine learning algorithms to avoid the class imbalance problem. MWEs as features in combination with distance measures and hierarchical clustering were successful in capturing and mapping difference in speech according to gender in the Lithuanian Parliament. Our results agree with the experimental outcomes of Hoover (2002) and Hoover (2003), where frequent word sequences and collocations combined with clustering showed more accurate results than just frequent words.

Keywords: multiword expressions, stylometry, parliamentary speeches

1. Introduction

Gender influence on language usage has been studied by many authors, but common agreement has not yet been reached (Lakoff, 1973; Holmes, 2006; Holmes, 2013; Argamon et al., 2003). Understanding gender differences in a professional environment would assist in a more balanced atmosphere (Herring and Paolillo, 2006; Dynel, 2008). Most previous studies relied on relatively small data sets, texts written by the individual authors which were too short to capture the variation in the language usage according to gender (Newman et al., 2008; Herring and Martinson, 2004). Besides, some authors have claimed that gender differences in language depend on the context, e.g., people assume male language in a formal setting and female in an informal environment (Pennebaker, 2011).

In this paper the impact of gender on the language used in a professional setting, i.e., Lithuanian Parliament debates, is explored. We study language with respect to style, i.e., male and female style of the language usage in the Parliament by applying computational stylistics or stylometry. Stylometry is based on two hypotheses: (1) the human stylome hypothesis, i.e., each individual has a unique writing style (Van Halteren et al., 2005); (2) the unique writing style of an individual can be measured (Stamatatos, 2009). From an information retrieval perspective, stylometry allows the derivation of meta-knowledge, i.e., what can be learned from the text about the author (Daelemans, 2013). This can be gender (Luyckx et al., 2006; Argamon et al., 2003; Cheng et al., 2011; Koppel et al., 2002), but also such things as age (Dahllöf, 2012), psychological characteristics (Luyckx and Daelemans, 2008), and political affiliation (Dahllöf, 2012).

As in many other studies of gender and language (Yu, 2014; Herring and Martinson, 2004), biological sex as the criterion for gender was used in this study. Also, we compare differences in the gender related language use at the group level. The Lithuanian language allows an easy distinction between male and female legislators based on their names.

This study seeks not to attribute text samples to female or male MPs (the authorship attribution task), but to explore variation of language use based on gender in political debates of the Lithuanian Parliament (detecting stylistic variation). Since one reason that idiolects differ is that people have different reserves of prefabricated word sequences (Larner, 2014; Johnson and Wright, 2014), in our experiments multiword expressions were used as distinguishing features speeches of female and male MPs. Also, because of the high imbalance in terms of the amount of data (significantly more for male MPs than for female MPs) as well as no gold standard corpus for reference being available, we used unsupervised machine learning methods for detecting stylistic variation between speeches made by female MPs

and male MPs. As most stylometric experiments using formulaic language as a feature were performed for English (e. g., Hoover (2003)), the main question this study seeks to answer is whether variability in language use with respect to style can be successfully captured using fixed word sequences, i.e., multiword expressions as features of Lithuanian which is a highly inflected language.

2. Data set

A corpus of parliamentary speeches from the Lithuanian Parliament [1] was used for capturing stylistic variation between genders. It consists of parliamentary speeches from March 1990 till December 2013. 10,727 speeches were made by female members of Parliament (MPs) and 100,181 by male MPs. The whole corpus contains 23,908,302 words (2,357,596 by female MPs and 21,550,706 by male MPs). Further statistics are shown in Table 1 (Kapočiūtė-Dzikienė and Utka, 2014).

	Number of samples	Number of words	Number of unique words	Average length of a sample in words
Female MPs	10727	2357596	93611	219.78
Male MPs	100181	21550706	268030	215.12
TOTAL	110908	23908302	279494	215.57

Table 1: Statistics of the corpus of transcribed Lithuanian parliamentary speeches.

The number of MPs included is 147, being only those included in the corpus, who produced at least 200 speeches of at least 100 words each. Out of 147 MPs, 129 were male and 18 were female.
All the samples were concatenated into two large documents based on gender. Then these two documents for the sake of faster processing were split into parts of equal size (except for the last parts of each big original document), giving 15 smaller documents of transcribed speeches from female MPs and 15 from male MPs.

3. Method

3.1 Stylistic features

Character n-grams are considered to be the most effective features in stylometric analysis (Kestemont, 2014; Stamatatos, 2009; Šarkutė and Utka, 2015; Kapociute-Dzikiene et al., 2014) because they are language-independent, are able to record style and stylistic differences and do not require external linguistic tools such as a part-of-speech tagger or parser. Using the most frequent words or function words (which in most cases have a high frequency (Hochmann et al., 2010; Sigurd et al., 2004)) as linguistic features is the most popular solution (Burrows, 1992; Hoover, 2007; Eder, 2013b; Rybicki and Eder, 2011; Eder and Rybicki, 2013; Eder, 2013a) for stylometric analysis. Most frequent words (MFW) are considered to be topic-neutral and have been relatively successful (Juola and Baayen, 2005; Holmes et al., 2001; Burrows, 2002).
However, we decided to use multiword expressions as linguistic features for our analysis. The choice was based on the assumption that the speech of politicians in their professional setting is rather formalised, and so uses specific expressions. Also, formulaic language can allow a more detailed interpretation (Antonia et al, 2014; Suzuki et al, 2012) of the results in comparison to character n-grams or even most frequent words. In a broad sense, a multiword expression (MWE) is a sequence of at least two words that are frequently used together (Marcinkevičienė, 2001). MWEs have "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al. 2002).
To obtain a list of MWEs to use in our stylometric experiments we used the Ngram Statistics Package [2] (Pedersen et al, 2011). The corpus of parliamentary speeches in the Lithuanian Parliament was split into word bi-grams and then association measures were calculated for each one. Lexical association measures assess the degree of association between components of possible MWE. For our experiment we chose two widely known association measures – Log-likelihood and Dice. Log-likelihood brings out word sequences with the highest degree of valence which ensures strength of association among the MWE components, while Dice gives higher values for word sequences in the corpus with equal frequencies and ignores sequences that are rare (Hunston, 2002). From the MWE candidates for which we calculated Log-likelihood and Dice values we took only the ones with the highest values and then manually eliminated sequences that were definitely not MWE. Eventually for our stylometric analysis we used a list of 4737 bi-gram MWEs. Examples of some MWE found in the corpus are presented in Table 2.

Dice	**Log-likelihood**
profesinėms sąjungoms (trade unions, dative)	gerbiamieji kolegos (dear colleagues, vocative)

[1] "Automatic Authorship Attribution and Author Profiling for the Lithuanian Language" (ASTRA) (No. LIT-8-69), 2014 – 2015.

[2] http://www.d.umn.edu/~tpederse/nsp.html

šventų atsiminimų (saint memories, genitive)	taip pat (also/as well)
Didžiojoje Britanijoje (Great Britain, locative)	Lietuvos Respublikos (Republic of Lithuania, genitive)
Kristijono onelaičio (Kristijonas Donelaitis, genitive)	bendru sutarimu (by consensus)
chasidų sinagogos(hassidic synagogue, genitive)	įstatymo projektas (bill (law), nominative)
status quo	Seimo nariai (members of the Parliament, nominative)
Drąsiaus Kedžio (Drąsius Kedys, genitive)	iš tikrųjų (indeed/actually)

Table 2: Examples of MWE by lexical association measures (Dice and Log-likelihood).

3.2 Statistical measures and experimental setup

The experiments were performed using the Stylo package for stylometric analysis with R (Eder et al., 2014). For the chosen approach firstly, using the whole corpus, a raw frequency list of features is generated, then normalized using z-scores. The z-scores are calculated by subtracting the mean frequency of a certain feature in one text from its mean frequency in all the texts in the corpus and dividing this difference by the standard deviation (Hoover, 2004a). Using Burrows' Delta measure (Burrows, 2002), the dissimilarity between two texts is the mean of the differences in z-scores over all the features under consideration in those two texts. A distance matrix is generated consisting of all the pairwise dissimilarity scores between the texts. This distance matrix can be visualized using a visualization technique such as a dendrogram produced by hierarchical agglomerative clustering.

Burrows's Delta is possibly the most popular distance measure used for stylometric analysis (Burrows, 2002; Rybicki and Eder, 2011). Delta depends on z-scores, the number of texts and the balance among them in terms of amount, length and number of authors (Stamatatos, 2009). Although this distance measure is effective for English and German texts, it has been less successful for more inflected languages such as Latin and Polish (Rybicki and Eder, 2011). Therefore a variant of Delta was chosen for our experiments. Eder's Delta is a modified standard Burrows's Delta, yet it gives more weight to the frequent features and rescales less frequent features to avoid random infrequent ones (Eder et al., 2014). It was developed for use with highly inflected languages, such as Lithuanian. However, this Delta variant retains sensitivity to the number of samples in the same way as other Delta variations.

The purpose of this paper is to capture stylistic dissimilarities/variations by mapping positions of the text samples in relation to each other according to gender, and therefore (hierarchical) clustering was chosen. Though its sensitivity to changes in the number of features or methods of grouping is well known (Eder, 2013a; Luyckx et al., 2006), in this study it gave rather stable results.

Additionally, the robustness of hierarchical clustering in this study was examined using the bootstrap procedure (Eder, 2013a). This procedure used extensions of Burrows's Delta (Argamon, 2008; Eder et al., 2014) with bootstrap consensus trees (Eder, 2013a) as a way to improve the reliability of cluster analysis dendrograms. Hierarchical clustering analysis lacks standard validation procedures, except for visual examination, and hence we found a combination of hierarchical clustering dendrograms and decision trees a useful tool for the evaluation of results.

4. Results

For our exploration of stylistic variation between female and male MPs from 50 to 4730 most frequent features, in this case MWEs, were chosen. Eder's Delta was combined with hierarchical clustering to visualize the categorization as well as for mapping positions of the samples in relation to each other, i.e. capturing variation in speech according to gender. No culling was applied (Eder et al., 2014; Hoover, 2004b) in our experiments. During the culling procedure words which have most of their occurrences in a single text instead of being distributed throughout the corpus, are eliminated (Stamatatos, 2009).

The results showed that using MWEs as features for stylometric analysis in combination with Delta variants and hierarchical clustering was successful in capturing differences in speech in the Lithuanian Parliament according to gender. The 50 most frequent MWEs were enough to capture the variation between the speeches of female and male MPs.

This recorded variation remained stable up to 1200 most frequent MWEs. This means that the first 1200 MWEs in the list used for analysis were helpful in capturing variation according to gender. The results are shown in Figures 1 and 2, where the data set is clearly divided into clusters corresponding to male and female speakers.

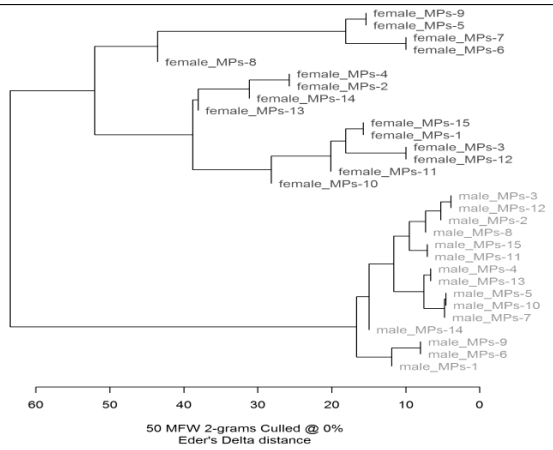

Figure 1: Variation between the speeches of female and male MPs with 50 most frequent MWEs as features.

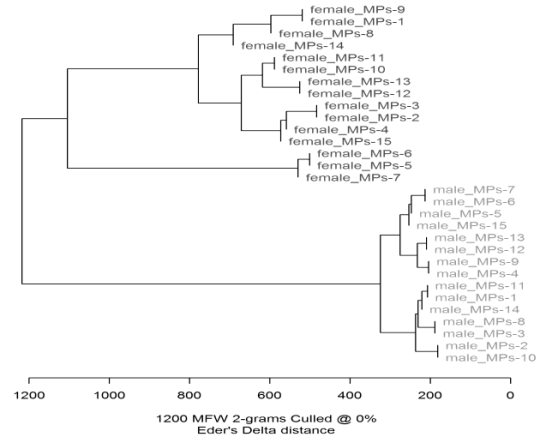

Figure 2: Variation between the speeches of female and male MPs with 1200 most frequent MWEs as features.

The Bootstrap Consensus Tree (BCT) procedure was applied to test the results. It is a combination of hierarchical clustering and decision trees (Eder, 2013a). It works by repeating the clustering with hundreds of different subsets of the original data, and retaining only those linkages between texts which appear in an above threshold proportion of runs. A consensus strength of 0.5 was chosen, i.e., the linkages between two texts retained if they appeared in at least half of the bootstrapping runs. The BCT results for discriminating between male and female legislators in the Lithuanian Parliament are shown in Figure 3.

Among other observations, female MPswere more inclined to use morphological collocations.These are defined as fixed expressions consisting of two or more functional words (inflected or non-inflected) that have a unified common meaning, are non-compositional and also have a syntactic function (Rimkutė, 2009; Rimkutė and Kovalevskaitė, 2010). Of the most frequent MWEs, female MPs used such morphological collocations as *dėl to* (‘therefore‘), *iki šiol* (‘by now‘), *be abejo* (‘undoubtedly‘), etc. more frequently then male MPs. Also, in the transcribed speeches of female MPs there occured more subjunctive constructions (indicating suggestion, certain degree of uncerntainty), for example, *aš siūlyčiau* (‘I would suggest‘), *aš manyčiau* (‘I would think‘), *galėtų būti* (‘[it] could be‘).

Male MPs, among other differences in comparison to female MPs, tended to use more references to other MPs. Moreover, they used more sequences related to power (e.g., *gynybos štabas* (‘defence headquarters‘), *ginkluotosios pajėgos* (‘armed forces‘), *diplomatinis korpusas*, (‘diplomatic corps‘)) economics/finance (e.g., *finansinė atskaitomybė* (‘financial accountability‘), *finansinis tvarumas* (‘financial sustainability‘), *fiskalinis deficitas* (‘fiscal deficit‘)). Also, male MPs used more verbs in the first person plural, for example *ar pritariame* (‘do we agree [?]‘), *galime sutarti* (‘we can agree‘), *būkime biedni* (‘let‘s be poor [but proper]‘ – part of popular Lithuanian saying).

As presented above, using stylometric analysis with MWEs as features, we were able to record certain differences in language usage according to gender. Some of them were topical, others of the nature of lexical or morphosyntactic style. For making more generalisations, further research is needed.

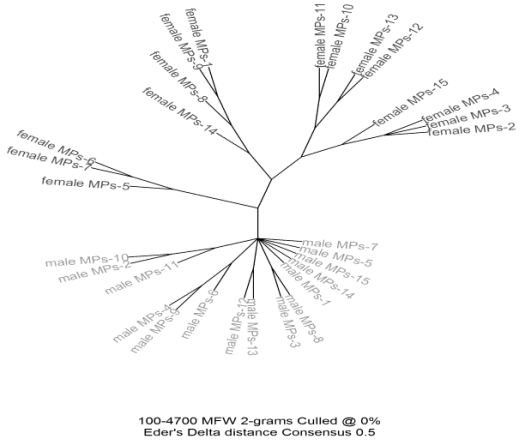

Figure 3: Variation between the speeches of female and male MPs BCT with consensus of 0.5.

5. Conclusions and future work

MWEs as features in combination with distance measures and hierarchical clustering were

successful in capturing and mapping difference in speech according to gender in the Lithuanian Parliament. Our results agree with the experimental outcomes of Hoover (2002) and Hoover (2003), where frequent word sequences and collocations combined with clustering showed more accurate results than just frequent words. However, although Eder (2011) reported increased accuracy using bi- and tri-gram collocations for English, word sequences were useless for other languages, especially Latin. Also, we got useful results with far fewer features than some studies (e.g. Eder (2010), Stamatatos (2006)), suggest for successful analysis. Therefore further, more extensive, experiments are required regarding the usefulness of MWEs as features, as well as the number of features and their range. For example, how many features from the beginning of the feature list are useful, and when we should select features from the middle and when from the end of the list of MWEs ordered by frequency. The effect of culling, the elimination of features with the most occurrences in a single text instead of being distributed throughout the corpus, also needs to be explored. We have shown that MWE can be used as linguistic features to discriminate between male and female speeches in the Lithuanian parliament, Lithuanian being an inflected language, and this approach could contribute to research on different usage of language depending on gender.

6. Bibliographical References

Alexis A., Craig, H., Elliott, J. (2014). Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution. *Literary and Linguistic Computing*, 29(2), pp. 147--163.

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2), pp. 131--147.

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text-the Hague then Amsterdam then Berlin*, 23(3), pp. 321--346.

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3), pp. 267--287.

Burrows, J. F. (1992). Not unles you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7(2), pp. 91--109.

Cheng, N., Chandramouli, R., and Subbalakshmi, K. P. (2011). Author gender identification from text. *Digital Investigation*, 8(1), pp. 78--88.

Daelemans, W. (2013). Explanation in computational stylometry. In *Computational Linguistics and Intelligent Text Processing*, Springer, pp 451--462.

Dahllöf, M. (2012). Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches - a comparative study of classifiability. *Literary and linguistic computing*, 27(2), pp. 139--153.

Dynel, M. (2008). Gendered Discourse in the Professional Workplace. *Journal of Pragmatics*, 9(40), pp. 1620--1625.

Eder, M. & Rybicki, J. (2013). Do birds of a feather really flock together, or how to choose training samples for authorship attribution. *Literary and Linguistic Computing*, 28(2), pp. 229--236.

Eder, M. (2010). Does size matter? Authorship attribution, small samples, big problem. *Proceedings of Digital Humanities*, pp. 132--135.

Eder, M. (2011). Style-markers in authorship attribution: a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, 6(1), pp. 99--114.

Eder, M. (2013a). Computational stylistics and biblical translation: How reliable can a dendrogram be. *The translator and the computer*, pp. 155--170.

Eder, M. (2013b). Mind your corpus: systematic errors in authorship attribution. *Literary and linguistic computing*, 28(4), pp. 603--614.

Eder, M., Rybicki, J., Kestemont, M., and maintainer Eder, M. (2014). Package 'stylo'.

Herring, S. C. & Martinson, A. (2004). Assessing gender authenticity in computer mediated language use evidence from an identity game. *Journal of Language and Social Psychology*, 23(4), pp. 424--446.

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10 (4), pp. 439--459.

Hochmann, J. R., Endress, A. D, and Mehler, J. (2010). Word frequency as a cue for identifying function words in infancy. *Cognition*, 115(3), pp. 444--457.

Holmes, D. I., Gordon, L. J., and Wilson, C. (2001). A widow and her soldier: Stylometry and the American civil war. *Literary and Linguistic Computing*, 16(4), pp. 403--420.

Holmes, J. (2006). Sharing a laugh: Pragmatic aspects of humor and gender in the workplace. *Journal of Pragmatics*, 38(1), pp. 26--50.

Holmes, J. (2013). *Women, men and politeness*. Routledge.

Hoover, D. L. (2002). Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing*, 17(2), pp. 157--180.

Hoover, D. L. (2003). Frequent collocations and authorial style. *Literary and Linguistic Computing*, 18(3), pp. 261--286.

Hoover, D. L. (2004a). Delta prime? *Literary and Linguistic Computing*, 19(4), pp. 477--495.

Hoover, D. L. (2004b). Testing Burrows's Delta. *Literary and linguistic computing*, 19(4), pp. 453--475.

Hoover, D. L. (2007). Corpus stylistics, stylometry, and the styles of Henry James. *Style*, 41(2), pp. 174--203.

Hunston, S. (2002). Methods in corpus linguistics: Beyond the concordance line. *Corpora in Applied Linguistics*, pp. 36--95.

Johnson, A., Wright, D. (2014). Identifying idiolect in forensic authorship attribution: an n-gram textbite approach. *Language and Law/Linguagem e Direito*, 1(1), pp. 37--69.

Juola, P. & Baayen, R H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, 20(Suppl), pp. 59--67.

Kapočiūtė-Dzikienė J., Šarkutė, L. and Utka, A. (2015). The effect of author set size in authorship attribution for Lithuanian. *Nordic Conference of Computational Linguistics NODALIDA*, pp. 87--96.

Kapočiūtė-Dzikienė, J. & Utka, A. (2014). Seimo posėdžių stenogramų tekstynas autorystės nustatymo bei autoriaus profilio sudarymo tyrimams. Linguistics/Kalbotyra, 66, pp. 27--45.

Kapociute-Dzikiene, J., Sarkute, L., and Utka, A. (2014). Automatic author profiling of Lithuanian parliamentary speeches: exploring the influence of features and dataset sizes. *Proceedings of the Sixth International Conference Baltic HLT 2014*, pp. 99--106.

Kestemont, M. (2014). Function words in authorship attribution from black magic to theory? *EACL 2014*, pp. 59--66.

Koppel, M., Argamon, S., and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), pp. 401--412.

Lakoff, R. (1973). Language and woman's place. *Language in society*, 2(01), pp. 45--79.

Larner, S. (2014). A preliminary investigation into the use of fixed formulaic sequences as a marker of authorship. *International Journal of Speech Language and the Law*, 21(1), pp. 1--22.

Luyckx K. & Daelemans, W. (2008). Personae: a corpus for author and personality prediction from text. *LREC 2008*.

Luyckx, K., Daelemans, W., and Vanhoutte, E. (2006). Stylogenetics: Clustering-based stylistic analysis of literary corpora. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*.

Marcinkevičienė, R. (2001). Tradicinė frazeologija ir kiti stabilūs žodžių junginiai. *Lituanistica*, 4(48), pp. 81--98.

Newman, M. L., Groom, C. J., Handelman, L. D., and Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3), pp. 211--236.

Pedersen, T., Banerjee, S. and McInnes, B. T. (2011). The Ngram statistics package (text::nsp): A flexible tool for identifying ngrams, collocations, and word associations. *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (ACL)*, pp. 131--133.

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, 211(2828), pp. 42--45.

Rimkutė, E. (2009). Gramatinė morfologinių samplaikų klasifikacija. *Kalbų studijos* 14, pp. 32-38.

Rimkutė, E. & Kovalevskaitė, J. (2010). Sudėtinės ir suaugtinės lietuvių kalbos morfologinės samplaikos. *Kalbų studijos* 16, pp. 79--88.

Rybicki, J. & Eder, M. (2011). Deeper delta across genres and languages: do we really need the most frequent words? *Literary and linguistic computing*, 26(3), pp. 315--321.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., Flickinger, Dan. (2002). Multiword expressions: A pain in the neck for NLP. *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, pp. 1--15.

Sigurd, B., Eeg-Olofsson, M., and Van Weijer, J. (2004). Word length, sentence length and frequency–zipf revisited. *Studia Linguistica*, 58(1), pp. 37--52.

Stamatatos, E. (2006). Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence Tools*, 15(05), pp. 823--838.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3), pp. 538--556.

Van Halteren, H., Baayen, H., Tweedie, F., Haverkort, M., and Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1), pp. 65--77.

Yu, B. (2014). Language and gender in congressional speech. *Literary and Linguistic Computing*, 29(1), pp. 118--132.

Open Source Code Serving Endangered Languages

Richard Littauer, Hugh Paterson III
Saarland University, SIL International
Saarbrücken, Germany, Oregon, USA
richard.littauer@gmail.com, hugh_paterson@sil.org

Abstract

We present a database of open source code that can be used by low-resource language communities and developers to build digital resources. Our database is also useful to software developers working with those communities and to researchers looking to describe the state of the field when seeking funding for development projects.

Keywords: open source, under-resourced languages, database, endangered languages, code, computational resources

1. Introduction

Almost half of the approximately 7,000 currently spoken languages are expected to become extinct this century; it is estimated that less than 5% of these will be used online or have significant digital presence (Kornai, 2013). Languages which do not have significant digital resources (often called low-resource, under-resourced, or minority languages) risk extinction from loss of prestige, specific domain usage (such as online), and ultimately loss of speakers.

Many language communities and academics working with them try to prevent this by developing tools, websites, and resources for their languages. However, these approaches are fragmented, often incurring large developmental and funding costs for single, non-extensible use cases.

To make this process easier for all stakeholders, we have built the first database (to our knowledge) of all open source code projects related to low-resource languages. It is available here: https://github.com/RichardLitt/endangered-languages[1].

Our database is structured as a simple list (in Markdown format), hosted in a GitHub[2] repository. GitHub is the largest online network of open source code and allows for parallel collaborative development, while providing critical collaborating support via a built in wiki, issue tracking and comments on new suggestions. The features provided by Github are increasingly important to academics and the field of education (Zagalsky et al., 2015). These features were not available via previous code and text sharing solutions such as SourceForge and are often not available via institutional repositories.

Our list is structurally simple, shareable, easily updated, and part of a wider cultural movement on GitHub of using Markdown files as simple databases [3]. Our list monopolizes on the low barrier of entry on GitHub (Storey et al., 2014). Because the list is part of GitHub, and because it is not maintained, funded, or dependent upon a large institution or funding body (but rather run by independent contributors), the list itself is a novel way of crowd-sourcing data and resources from the linguists, computational linguists, and coders who would use the tools themselves.

Our list currently describes over 241 open source projects, which includes specific sections for extensible code for over 26 different languages.

Our solution is open source, not just in the sense of code and data availability (or disclosure) but in the sense of collaborative involvement and inclusive discussion. By using a solution like GitHub we were able to reach out to both developers and researchers. Our list is updatable more rapidly than previous solutions like Lingtransoft [4], or other resources like GOLD [5], ELCAT [6], or OLAC [7]. We mitigate the risk of a single point of failure (a recent issue with LinguistList [8]) by working in a distributed fashion. The data in our list is open and can be copied and reused by anyone, something not necessarily true of previous solutions. By choosing such a solution, we sidestep many of the institution issues associated with archives and repositories currently servicing academics.

Our ultimate goal is a collaboratively built and maintained resource for highlighting useful, extensible code for low-resource languages. We would like to share our current efforts and welcome communication with the wider academic linguistics community.

2. Database structure

The list itself largely consists is a single Markdown file. This is useful for a couple of reasons; first, there's no need for a gateway or endpoint to access the content of the database, as there would be if it were coded in SQL, RDF, or some other relational database. Secondly, people can search the list using their browser and the standard search feature for any website. Finally, the list can be digested immediately instead of depending on searches to get complete coverage of the data.

2.0.1. Categories

Within the list, there are several main sections where we attempt to categorize the resources, based on user input and

[1] https://github.com/RichardLitt/endangered-languages

[2] https://github.com

[3] Sindre Sorhus's list of awesome-lists. https://github.com/sindresorhus/awesome

[4] http://lingtransoft.info

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

[6] http://www.endangeredlanguages.com

[7] http://www.language-archives.org

[8] http://linguistlist.org

upon the best guesses of the maintainers about the functionality of the various tools. These sections are: Generic repositories (which includes massive dictionary and lexicography projects, single language lexicography projects, utilities, presentations of data, and software), i18n-related repositories, audio automation, text-automation, experimentation, flashcards, natural language generation, computing systems, android applications, Chrome extensions, FieldDB, FieldDB web-services and components and plugins, academic research paper-specific repositories, example repositories, and language and code interfaces.
We also have two other lists: One of other open source linguistic organizations, on GitHub, and other OSS (Open Source software) organizations, and another for language-specific projects, which includes subsections for code which is relevant to: Amharic, Arabic, Bengali, Chichewa, Estonian, Georgian, Guarani, Hausa, Hindi, Hognorsk, Inuktitut, Irish, Japanese, Kinyarwanda, Korean, Lingala, Malay, Malagasy, Migmaq, Minderico, Nishnaabe, Oromo, Quechua, Sami, Scottish Gaelic, Secwepemctsín, Somali, Tigrinya, Zulu.
Finally, we also include a short list of closed source resources which can still be utilized for free.

2.0.2. Example entry

Each entry is a single line, containing the name of the resource, a link to the resource, and a short description. If the resource is also a GitHub repository, we include a link to a badge that shows the amount of stars (similar to likes or favorites on other social media sites, and, generally, a good proxy for usage and developer uptake of the resource) for that repository.
Here is one such entry, for Scannell's Chichewa code[9]:

```
* [Chichewa ![GitHub stars]
(https://img.shields.io/github/stars
/kscanne/chichewa.svg)]
(https://github.com/kscanne/chichewa)
 NLP resources for Chichewa.
```

The "GitHub stars" link automatically includes an SVG image file of the amount of stars that particular repo has gotten, which allows readers to easily see how popular a repository is on GitHub. Note that this would normally be one line of code.

3. Personas and Stakeholders

The list that we are maintaining is not aimed at any one group in isolation. Instead, there are several key groups whom we think may derive value from this list. These are as follows: Project Managers, Software Developers, Community Developers, and Linguists.

3.1. Project Managers

Project managers are generally linguists or community members who have been tasked with developing language resources, but generally don't have the background to understand the technical aspects on their own (and thus are different from software developers). Many project managers do not have strong information technology project management backgrounds. However, they are often skilled linguists who have access to grant funding. The finer details of carrying out a project in a manner which benefits more than one language community is simply out of scope for many first time managers, and projects. We hope that project managers should be able to look at our list to be able to determine two things: 1. Has the project task, goals, or relevant deliverables already been accomplished for another language (or indeed, for the same language), and 2. where can they find the code base for that project so that they can integrate it into their own workflow. We hope, as well, that some project managers might be able to have a project used case and that they can use our list to answer the question of how to get funding for their idea.

3.2. Software Developers

Software developers are often looking to find pre-existing solutions, and to find pre-existing modules which can be applied to new use cases they are asked to solve. Every problem which has already been solved by someone else is time that does not have to be spent developing. Every major project today uses open source code in some capacity, partially for this reason. This is in some ways the opposite perspective from the project manager, who are looking to develop something new and may not be looking for extensible solutions to their problem. The developer is looking for use cases to which they can apply or reapply their code. We hope that our list makes this easy.
One developer, at least, has said that this was the case:

> Thanks a lot for pointing me to the list. It is awesome! It has some really good tools and resources which would be very useful in a lot of things that I am doing. I shall definitely add some of my resources and tools to this list....I have also come across a very useful library - Poio-api[10] - on your list which is a parser for most of these XML files that I work with.

(Personal communication, 2015)

3.3. Community Developers Doing Language Development

This may include language development organizations, or individuals who are "community members" looking to deploy their own solutions. They want to know "does it work with my language?" by which they often, but not always, mean written language. An additional complexity is that there are different perspectives on what "does it work" really mean. The degree of technological uptake in the majority culture may affect the expectation about what kind of tasks can be easily accomplished in the low-resourced language. Tasks might be, for example, sending SMS messages in a particular script. But sometimes, users of low-resource languages have a very different expectation and interaction with technology. For instance, members of deaf

[9] https://github.com/kscanne/chichewa

[10] https://github.com/cidles/poio-api

communities require video integration in their digital solutions more than many other kinds of low-resource languages, but deaf communities may still need other tools commonly shared with written languages - like dictionary tools.

3.4. Linguists

We define linguists here as researchers who are not tech savvy, but who are working in academia or directly with language communities from an academic perspective. Linguists, as such, are generally looking for patterns in language data. They want tools which are going to be easy to use to find the patterns they are looking for and to present the data in ways which help others to understand the purpose and meaning of those patterns. As end users, they are more likely to be looking for tools which are useful out of the box, and so may not be able to appreciate all of the items in the list, but still may benefit from a quick search through it.

As well, we provide many links to tools that can be used with ELAN[11], Praat[12], and other audio software which are used on a day-to-day basis by many linguists themselves.

4. Commitments

Unlike most projects, which must depend upon institutions, private or public funding, this project has no single point of failure. The list itself is currently hosted by Littauer's GitHub account[13], but due to the nature of a git[14] repository and of collaborative work on GitHub, any replication of the list can be edited, stand alone, and be used if the original project goes down for any reason. The decentralized quality of git repositories makes this list a much less brittle solution than individually hosted database or institutional repositories.

As well, there is a very low possibility of misuse; if any one person using the program decides to enforce their own viewpoint, perspective, or rules at the cost of any other user group or of the community, it is entirely possible for anyone else to make a copy of the list and to then use that as the main source of truth going forward.

One possible issue with the list is that if a malicious user takes over the main list. Then, it would take some time for any other list to have the same clout in the community. This is, however, true of all existing databases. A good example, although not open data, is the Ethnologue[15], which recently added a paywall to their database about languages (but not to the ISO 639-3 standard which SIL International[16] also stewards). The loss of a previously free accessible resource became a major source of contention for many linguists, although SIL International had legitimate financial reasons for doing so.[17].

One of the future goals of the project is to develop a community where anyone can ask questions about the list and its resources, and other users can help out and give advice easily. While this is possible with GitHub issues, it depends upon a higher amount of usage of the list itself than currently exists. Marketing the list in tech conferences is one possible solution.

5. Conclusion

We have here outlined our reasons for developing a list of open source software for endangered languages. We hope that this resource is used by the community, and that this paper fosters discussion and awareness of open source resources.

6. Bibliographical References

Kornai, A. (2013). Digital language death. *PloS one*, 8(10):e77056.

Storey, M.-A., Singer, L., Cleary, B., Figueira Filho, F., and Zagalsky, A. (2014). The (r) evolution of social media in software engineering. In *Proceedings of the on Future of Software Engineering*, FOSE 2014, pages 100–116, New York, NY, USA. ACM.

Zagalsky, A., Feliciano, J., Storey, M.-A., Zhao, Y., and Wang, W. (2015). The emergence of github as a collaborative platform for education. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1906–1917. ACM.

[11] http://www.mpi.nl/corpus/html/elan/

[12] http://www.fon.hum.uva.nl/praat/

[13] https://github.com/RichardLitt

[14] https://git-scm.com/

[15] http://ethnologue.com/

[16] http://www.sil.org/

[17] http://www.ethnologue.com/ethnoblog/m-paul-lewis/ethnologue-launches-subscription-service

Morphology Learning for Zulu

Uwe Quasthoff[1,2], Dirk Goldhahn[1], Sonja Bosch[2]
[1]NLP Group, Department of Computer Science, University of Leipzig, Germany
[2]Department of African Languages, University of South Africa, South Africa
Email: quasthoff@informatik.uni-leipzig.de, dgoldhahn@informatik.uni-leipzig.de, boschse@unisa.ac.za

Abstract

Morphology is known to follow structural regularities, but there are always exceptions. The number and complexity of the exceptions depend on the language under consideration. Substring classifiers are shown to perform well for different languages with different amounts of training data. For a less resourced Bantu language like Zulu the learning curve is compared to that of a well-resourced language like German, the learning curve of which might be considered as an extrapolation for the less resourced languages in the case of larger training set sizes. Two substring classifiers, TiMBL and the Compact Patricia Tree Classifier are tested and shown to give comparable results.

Keywords: morphology, machine learning, classification, Zulu, less resourced languages

1. Introduction

Machine learning of morphology has several interesting aspects. First, morphology is known to follow structural regularities (i.e. repeatedly occurring substrings often have the same segmentation), but there are always exceptions. The number and complexity of the exceptions depend on the language under consideration. So the first question is to investigate the quality of machine learning methods for this task: Which classifier works well, and how much training data is needed for a certain quality. The second aspect is linguistic: For many languages (especially the less resourced languages) there are no extensive resources for morphological analysis. So the amount of training data is usually very limited and any information about the size of the necessary training data is helpful. We aim to compare suffix learning in two languages, on the one hand German, a highly resourced language, and on the other hand Zulu, a considerably less resourced language (cf. Quasthoff et al, 2014). For German, much larger training data are available and this might help to extrapolate the Zulu learning curve.

The paper is organized as follows: after an introduction to Zulu morphology and the available training data in section 2, the classification tasks using TiMBL and the Compact Patricia Tree Classifier (CPTC) are explained in sections 3 and 4. The results of the classification for Zulu prefixes and suffixes are presented in section 5. Section 6 compares the situation for Zulu and German suffixes for different sizes of training data. Section 7 discusses the main classification errors. The last section gives a perspective on future work and an outlook for the application to other Bantu languages.

2. Zulu Morphology

Zulu [ISO 639-3: zul], like other members of the Bantu language family, has a complex morphological structure, characterised among others by the extensive use of prefixes as well as suffixes in the formation of words. The basis from which Zulu words are constructed is the root, which is the lexical core of a word while affixes usually add a grammatical meaning or function to the word (inflection and derivation). Hence, morphological segmentation is essential for any kind of information retrieval from text corpora. Some of the morphological analysers for Zulu that are reported on, are e.g. machine learning Zulu analysers (Spiegler et al. 2008; Shalanova et al. 2009), and a bootstrapping approach (Joubert et al. 2005). However, none of these morphological analysers is freely available. A finite-state morphological analyser ZulMorph, reported on by Bosch and Pretorius (2011) is accessible as demo, but analysing a single word at a time[1].

Let us look at the complex nature of affixes in Zulu necessitating morphological decomposition. The noun classification system of the Bantu languages in general, categorises nouns into a number of noun classes (18 classes in the case of Zulu), which manifest themselves as regularities in noun prefix morpheme variations. These play a pivotal role in the morphological structure of the language, where noun-class agreement prefixes express agreement with verbs (subject and object prefixes), adjectives, possessives, pronouns and so forth (see Kosch, 2006). In addition to class-dependent prefixes, there are further examples of class-independent prefixes in the case of nouns, viz. copulative, locative and adverbial morphemes; and in the case of verbs, there are negative, tense and aspect morphemes.

Suffixes in Zulu do not correspond to nominal classification and are therefore fewer in number and variation. Examples of class-independent suffixes in the case of nouns are diminutive, augmentative and locative morphemes; and in the case of verbs we have for instance the verb terminative morpheme (positive/negative, tense) and extension morphemes.

In many cases of affixes combining with other affixes or roots, morphophonological changes between underlying and surface levels may occur. In other words, the same

[1] http://gama.unisa.ac.za/demo/demo/zulmorph

morpheme may be realised in different ways depending on the environment in which it occurs. For instance, with the suffixation of the diminutive morpheme *-ana*, various phonological changes may occur, depending on the nature of the final syllable of the noun stem[2], for example: *ingubo* (blanket) > *ingutshana* (small blanket). The same applies to the suffixation of e.g. the passive extension suffix *-w-* where palatalisation occurs in the verb root when it ends in certain syllables, for example: *-loba* (write) > *-lotshwa* (be written).

The available training data used, are Set 1: the isiZulu NCHLT Annotated Text Corpora (Language Resource Management Agency, 2013). The latter are lemmatised, part of speech tagged and morphologically analysed corpora based on documents from the South African government domain crawled from gov.za websites and collected from various language units. The corpora were developed during the Department of Arts and Culture's National Centre for Human Language Technologies (NCHLT) Text project. A description of these corpora is given in Eiselen and Puttkammer (2014); and Set 2: the Ukwabelana (2013) word list with segmentations and labelled morphological analyses described by Spiegler et al. (2010). It should be noted that, to the best of our knowledge these two mentioned data sets are the only freely available segmented and tagged Zulu data sets.

Set 1 is available under the terms of the Creative Commons Attribution 2.5 South Africa License[3], while Set 2 is released with the GNU General Public License Version 3, 29 June 2007[4]. Since we reported on the resource scarceness of Zulu corpora in 2014 (Quasthoff et al., 2014) the situation has not changed much except for the fact that Set 1 has become available.

The two different data sets that were used, contain:
Set1: 19,439 words
Set2: 9,224 words

The intersection has the following size:
| Set1 ∩ Set2 | = 8,652 Elements

Because of the complicated structure in the case of multiple prefixes and suffixes in Zulu, the situation was simplified in the following way: First, we only consider the open word classes noun and verb, and the combined word classes adjective/adverb (abbreviated to N, V and A). For each word, multiple prefixes were concatenated to one single prefix, and multiple suffixes were concatenated to one single suffix. Every word now has the form prefix-root-suffix, where both prefix and suffix might be empty and the root is classified as N, V or A. There is a very small number of words that have a more complex structure which was ignored.

[2] We distinguish between "stem" and "root" as in Faaß et al. (2012).
[3] http://creativecommons.org/licenses/by/2.5/za/legalcode
[4] http://www.cs.bris.ac.uk/Research/MachineLearning/Morphology/Resources/Ukwabelana/LICENSE.txt

Although circumfixes, that is simultaneous affixation of a prefix and a suffix (e.g. a negative prefix which requires a negative suffix), do occur to some extent in Zulu morphology, these were not taken into consideration at this stage.

It should be noted that in the Ukwabelana (2013) corpus, adjectives and so-called relatives are grouped together under the POS adjectives, the reason probably being that these two morphologically distinct categories fulfil the same syntactic function. For instance in morphological terms *-dala* is an adjective stem while *-luhlaza* is a relative stem, however both are marked as <ar>, adjective root in this corpus. There are also cases of inconsistency in the data used, which could influence the results, e.g. the following is an example from the Language Resource Management Agency (2013) data indicating that the same stem is simultaneously classified as adjective and relative: esi-nzima{RelStem} / ne-nzima{AdjStem}.

3. Morphology Learning with TiMBL

The set of attributes was selected in the following language independent way: The only assumptions are that a suffix boundary can be identified if we consider long enough tail substrings of the words and have enough training data showing all possibilities. Similarly for prefixes: We only assume that the end of a prefix can be identified if we consider long enough head substrings of words. so a possible dependence of prefix and suffix is not considered. Moreover, part-of-speech is assumed to be learnable inspecting the head and the tail of a word. So we have three classification tasks, where we can use the same attributes for classification:

POS: Identify part-of-speech (i.e. N, V, or A) given head and tail of a word.
Prefix: Identify the prefix given the head of the word (and the tail as a kind of dummy attributes).
Suffix: Identify the suffix given the tail of the word (and the head as a kind of dummy attributes).
Table 1 shows some data with the following attributes:

Word:	the word to be classified
Attributes:	up: does the word contain any uppercase characters?
	l1 … l7: the n leftmost characters, n=1...7
	r7… r1: the n leftmost characters, n=1...7
Classification Targets:	
	POS (N, V, or A)
	pre: prefix
	suf: suffix

The classification targets can be learned independently. Another useful strategy is to learn POS first and to use is later as an additional attribute for finding the affixes.

It is noteworthy that the Zulu "words" in Table 1 are the underlying morpheme strings after decomposition and deletion of annotations that is, without morpho-phonological changes.

For the classification, two different tools were used: TiMBL used the attributes above and different classifiers and distance measures were tested. The following selection gave the best overall results:
classification algorithm: TRIBL2 (starting with the decision tree algorithm IGTree and later switching to k-NN and k=1)
metric: Dice coefficient
feature-weighting: Gain Ratio
In section 6, it is shown that the same parameters are working for German data.

4. Classification using Compact Patricia Trees

As a second classifier, a pure decision tree suffix classifier was used. The Compact Patricia Tree (CPT) classifier (Morrison, 1968) generates a decision tree having the last character as first attribute, the second-last as second attribute and so on. In contrast to TiMBL, there is no limit in the length of the substrings as in the above approach. For prefix learning, we use the word in reverse ordering, of course. Using the implementation from Witschel (2005), CPT classifiers are trained to return a class when given a string (Eiken, 2006). With this approach the training set is perfectly reproduced. Due to the compact representation and an efficient search mechanism in the tree, CPTs can be used as lexical components for millions of words. The feature of highest importance of CPTs is their ability to generalise, that is, to return a classification for unseen strings. When an unseen word is classified, the class will be chosen based on training words with the longest common affix. Therefore the same class will be assigned for similar strings. In case of several training words of different classes matching with the same affix, the class is chosen based on the class distribution. CPTs can be trained on beginnings or endings of strings making it easy to apply them to e.g. prefix- and suffix-classification.

5. Classification results for Zulu prefixes and suffixes

Tables 2 and 3 show the classification results for the different tasks of learning POS, suffixes and prefixes. In every case, 10% of the data are used as test data, and for training another 10%, 20%, …, or 90% are used.
The results differ slightly for the different sets of training data. The bigger difference is for the different tasks. POS and suffix detection result in reasonable quality around or above 90%, prefix detection seems to be more difficult. This relates to the entropy for the distribution of the values for the classification target.
For a more linguistic interpretation of the results see section 7.

6. Comparison with German

For a better interpretation of the results the same task was repeated for German POS and suffixes. For prefixes, the situation is much simpler compared to Zulu. For German, there are much more training data available and hence the results might give an impression of what would happen with more training data for Zulu.
The German data set contains 70,000 words with the same POS tags N, V or A. Here, suffix is understood as inflexional suffix, derivational suffixes are not considered. The data come from wortschatz.uni-leipzig.de and are not free of errors.
For the classification task, 10% of the data are taken as test data and different amounts of the remaining data are used for training (see table 4 and 5). The percentages start lower to make the absolute numbers comparable.

7. Error analysis

Some typical classification errors for German and Zulu words together with possible reasons are given in Tables 6 to 10. The list is ordered by error frequency. There was no difference for the two classifiers TiMBL and CPTC noticed.

Error[5]	*Examples*	*Explanation*
A/V	dotierten, markierten, notleidenden	participles are often misclassified
A/N	Schweizer, Zahlreiche, Heiligen	capitalization at sentence beginnings or as proper nouns
V/A	verhinderte, geschnitten, gelangten	participles are often misclassified
N/A	Barometer, Acryl, Nirvana	foreign words
N/V	Clou, Biss, Stern	short words
V/N	Münden, Blickt, Gekocht	capitalization at sentence beginnings

Table 6: Typical classification errors for German POS.

Error	*Examples*	*Explanation*
en/n	Reisen, Terminen	general difficulty to recognize nouns ending in -e
n/en	Pannen, Ruinen, Vorlieben	general difficulty to recognize nouns ending in -e
-/s	Orleans, Glas	short or foreign words

Table 7: Typical classification errors for German suffixes.

[5] The correct classification is followed by the error made by the classifier

The following tables contain selected examples for misclassified Zulu words together with a possible reason in each case:

Error	*Examples*	*Explanation*
A/N	kumnyama	general difficulty to distinguish between identical noun stems and relative stems
N/V	kwelikamlotha	general difficulty to distinguish between identical noun stems and verb roots
N/V	israyeli	difficulty to recognize foreign noun stems

Table 8: Typical classification errors for Zulu POS.

Error	*Examples*	*Explanation*
engiyi/engi	engiyizwayo	general difficulty to recognize monosyllabic verb roots
sengiyam/ sengiya	sengiyam-esaba	general difficulty to recognize vowel-initial verb roots
baka/ba	bakageorge	difficulty to recognize foreign noun stems

Table 9: Typical classification errors for Zulu prefixes.

Error	*Examples*	*Explanation*
ela/a	oweqela	difficulty to recognize short verb roots
a/wa	esizwa	general difficulty to recognize verb roots that include a last syllable that resembles a verb extension suffix

Table 10: Typical classification errors for Zulu suffixes.

8. Future work

As more training data becomes available from the Language Resource Management Agency (2013), it is planned to extend the machine learning of affixes to other less resourced Bantu languages of South Africa as well. Bootstrapping the results of the described Zulu case for purposes of closely related languages such as Xhosa, Swati and Southern Ndebele, will also be investigated.

Moreover, the training data together with the classifier will be made available as a tool for the morphological segmentation for all these languages. After removal of prefixes and suffixes this tool can be used to identify new stems used in these languages. At this stage, the pure affix removal does not generate a linguistic stemmer, because some further transformation of a stem identified might be necessary.

9. Bibliographical References

Bosch, S.E. and Pretorius, L. (2011). Towards Zulu corpus clean-up, lexicon development and corpus annotation by means of computational morphological analysis. South African Journal of African Languages 31(1):138-158. ISBN 0257-2117. http://uir.unisa.ac.za/bitstream/handle/10500/5539/bosch_sajal_v31_n1_a11.pdf?sequence=1&isAllowed=y Accessed on 23 March 2016.

Eiken, U.C., Liseth, A.T., Richter, M., Witschel, F., and Biemann, C. (2006). Ord i Dag: Mining Norwegian Daily Newswire. Proceedings of FinTAL, Turku, Finland.

Eiselen, E.R. and Puttkammer, M.J. (2014). Developing text resources for ten South African languages. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland. p. 3698-3703.

Faaß, G., Bosch, S. and Taljard, E. (2012). Towards a Part-of-Speech Ontology: Encoding Morphemic Units of Two South African Bantu languages. Nordic Journal of African Studies 21(3):118–140. ISSN 1459-9465. http://www.njas.helsinki.fi/ Accessed on 23 March 2016.

Joubert, L., Zimu, V., Davel, M. and Barnard, E. (2004). A framework for bootstrapping morphological decomposition. Available: http://www.meraka.org.za/pubs/joubertl04morphanalysis.pdf Accessed on 23March 2016.

Morrison, D. (1968). Patricia- practical algorithm to retrieve information coded in alphanumeric. Journal of ACM, 15(4):514–534.

Quasthoff, U., Bosch, S. and Goldhahn, D. (2014). Morphological Analysis for less-resourced Languages: Maximum Affix Overlap applied to Zulu. Workshop "CCURL 2014 - Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era" 26th May 2014. LREC, 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland. pp. 52-55.

Shalonova, K., Golenia, B. and Flach, P. (2009).Towards learning morphology for under-resourced languages. IEEE Transactions on Audio, Speech and Language Processing, 17(5):956–965.

Spiegler, S., Golenia, B., Shalonova, K., Flach, P. and Tucker, R. (2008). Learning the morphology of Zulu with different degrees of supervision. IEEE Spoken Language Technology Workshop, pp. 9–12.

Spiegler, S., van der Spuy, A. and Flach, P.A, (2010). Ukwabelana - An open-source morphological Zulu corpus. Proceedings of the 23rd International

Conference on Computational Linguistics (COLING). p. 1020-1028.
Witschel, H.F. and Biemann, C.: Rigorous dimensionality reduction through linguistically motivated feature selection for text categorisation. Proceedings of NODALIDA, Joensuu, Finland (2005).

10. Language Resource References

Language Resource Management Agency (2013). isiZulu NCHLT Annotated Text Corpora. http://rma.nwu.ac.za/index.php/resource-catalogue/isizulu-nchlt-annotated-text-corpora.html Accessed on 12 February 2016.
Ukwabelana - An open-source morphological Zulu corpus. (2013). Available: http://www.cs.bris.ac.uk/Research/MachineLearning/Morphology/resources.jsp Accessed on 11 February 2016.
Wortschatz Universität Leipzig. (2016). Accessed on 12 February 2016. http://wortschatz.uni-leipzig.de/ Accessed on 23 March 2016.

word	up	l1	l2	l3	l4	l5	l6	l7	r7	r6	r5	r4	r3	r2	r1	POS	pre	suf
abahlali	0	a	ab	aba	abah	abahl	abahla	abahlal	bahlali	ahlali	hlali	lali	ali	li	i	V	aba	i
abakwenziwe	0	a	ab	aba	abak	abakw	abakwe	abakwen	wenziwe	enziwe	nziwe	ziwe	iwe	we	e	V	abakw	iwe
esesemabangeni	0	e	es	ese	eses	esese	esesem	esesema	bangeni	angeni	ngeni	geni	eni	ni	i	N	esesema	eni
lainvume	0	l	la	lai	lain	lainv	lainvu	lainvum	ainvume	invume	nvume	vume	ume	me	e	N	lain	*
ngaukugcwalisa	0	n	ng	nga	ngau	ngauk	ngauku	ngaukug	cwalisa	walisa	alisa	lisa	isa	sa	a	V	ngauku	isa
ngauMsombuluko	1	n	ng	nga	ngau	ngauM	ngauMs	ngauMso	mbuluko	buluko	uluko	luko	uko	ko	o	N	ngau	*
saukufakela	0	s	sa	sau	sauk	sauku	saukuf	saukufa	ufakela	fakela	akela	kela	ela	la	a	V	sauku	ela
ukubuya	0	u	uk	uku	ukub	ukubu	ukubuy	ukubuya	ukubuya	kubuya	ubuya	buya	uya	ya	a	V	uku	a

Table 1: Sample data with letter n-grams as attributes.

Task	*Entropy*	*Percentage used as training data*								
		10	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*
Number of training data		923	1,846	2,769	3,691	4,613	5,535	6,457	7,379	8,301
RMA-prefix	7.08	67	70	71	72	73	73	74	76	78
RMA-POS	1.00	84	87	89	89	90	90	91	92	92
RMA-suffix	2.10	91	93	94	94	95	96	96	96	96
Spiegler-prefix	5.93	61	63	63	63	64	64	66	67	72
Spiegler-POS	0.85	80	84	84	85	86	87	87	87	87
Spiegler-suffix	2.54	85	88	90	91	91	91	92	92	92

Table 2: Quality of TiMBL results for Zulu in %.

Task	*Entropy*	*Percentage used as training data*								
		10	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*
Number of training data		923	1,846	2,769	3,691	4,613	5,535	6,457	7,379	8,301
RMA-prefix	7.08	69	70	72	73	74	75	75	76	78
RMA-POS	1.00	79	82	85	86	87	88	89	89	89
RMA-suffix	2.10	90	93	94	95	95	95	95	95	95
Spiegler-prefix	5.93	63	61	64	65	67	68	68	69	71
Spiegler-POS	0.85	75	80	81	82	83	84	84	84	85
Spiegler-suffix	2.54	86	89	91	92	92	92	93	92	93

Table 3: Quality of CPTC results for Zulu in %.

Task	*Entro-py*	*Percentage used as training data*													
		3	*6*	*9*	*12*	*15*	*18*	*21*	*30*	*40*	*50*	*60*	*70*	*80*	*90*
Number of training data		2,036	4,148	6,223	8,349	10,473	12,565	14,605	20,889	27,884	34,973	42,004	48,970	55,974	62,976
deu-POS	0.86	87	88	88	89	89	89	89	90	91	91	91	92	92	92
deu-suffix	1.94	89	91	92	93	93	94	94	94	95	95	95	95	96	96

Table 4: Quality of TiMBL results for German in %.

Task	*Entro-py*	*Percentage used as training data*													
		3	*6*	*9*	*12*	*15*	*18*	*21*	*30*	*40*	*50*	*60*	*70*	*80*	*90*
Number of training data		2,036	4,148	6,223	8,349	10,473	12,565	14,605	20,889	27,884	34,973	42,004	48,970	55,974	62,976
deu-POS	0.86	93	93	94	94	95	95	95	95	95	95	95	95	95	95
deu-suffix	1.94	88	91	91	92	93	93	93	94	94	95	95	95	95	95

Table 5: Quality of CPTC results for German in %.

Proceedings of the LREC 2016 Workshop
"CCURL 2016 – Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity"

23 May 2016 – Portorož, Slovenia

Edited by Claudia Soria, Laurette Pretorius, Thierry Declerck, Joseph Mariani, Kevin Scannell, Eveline Wandl-Vogt

http://www.ilc.cnr.it/ccurl2016/

Acknowledgments: the CCURL 2016 Workshop is endorsed by the Erasmus+ DLDP project (grant agreement no.: 2015-1-IT02-KA204-015090).