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

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.

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).