Proceedings of the Language Technologies for All (LT4All) , pages 407–436
Paris, UNESCO Headquarters, 5-6 December 2019.
NC

LT4All 2019 Oceanian Languages Poster Session

P.1.1: Building capacity for community-led documentation in Erakor, Vanuatu
Authors: Ana Krajinovic, Rosey Billington, Lionel Emil, Gray Kaltaƥau and Nick Thieberger
Country: Germany

Abstract: We discuss a collaboration between community members and visiting researchers in Erakor, Vanuatu aiming to build the capacity of community-based researchers to undertake language and cultural documentation projects. We focus on the outcomes and benefits of the community-led project in Erakor after initial training, which include: a) long-term documentation of linguistic and cultural practices calibrated towards community's needs (using the PARADISEC repository for ongoing access), and b) collections of large quantities of data of good phonetic quality, which, besides being readily available for research, have a great potential for training and testing language technologies, such as automatic speech recognition.

Native language: Komam utilusus teflan naɱer nig natkon go naɱer weswes ni nlaun nakon nen ruto saof natkon, teflan rufaitau go rutafnau weswes ni natkon raki teflan ruktɱer wesweski nawesien ni namtrirsokwen go nakraksokien ni nafsan go suƥ. Komam ule toop pak nua nawesien ni teflan nataɱol ni natkon ɱas itɱen weswes eɱrom ni Erakor, go ntakun ni tete nfaitauen go nafregnrogwen toklos, a) nawesien ni nakraksokien ni nafsan go suƥ raki naɱitwen nig nfanu, b) nakraksokien ni data kelaap go tenen misleo knen iwi, nen rukta tae ler nametmatuan knen ɱas mau me nen ruktae pregi rupi teni nafregnrogwen ni nanrogwen ni nafeswen.

P.1.2: PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures)
Authors: Amanda Harris and Nick Thieberger
Country: Australia

Abstract: Language archives play an important role in keeping records of the world's languages safe. Accessible audio recordings held in archives can be used by speakers of small and endangered languages, and their communities, and provide a base for further research and documentation. There is an urgent need for historical analog tape recordings to be located and digitised, as they will soon be unplayable. PARADISEC holds records in 1228 languages. We run training for language documentation and are developing technologies to localise access to language records. A concerted effort is needed to support language archives and sustain language diversity.

Resume: Wanpela kain ples olsem akaiv i save lukautim ol rekod or pepa bilong ol kain kain tok ples bai stap gut long bihain taim. Planti ol liklik tok ples ol klostu dai nau. Tok ples bilong ol manmeri na komuniti mas usim akaiv long helpim wok bilong painim aut moa na raitim ol pepa bilong ol tok ples. PARADISEC i gat 1228 tok ples. Mipela painim ol rikoding long taim tumbuna we ol tok ples i stap long keset tep bilong dijitaisim nau o sapos nogat bai ol bagarap. Taim nau long sapotim ol tok ples akaiv long lukautim planti kain kain tok ples.

P.1.3: Bloom Books
Authors: Paul Nelson
Country: United States

Abstract: Bloom is a tool/system that facilitates the creation of literacy materials in any language. Beginning with shell books, a user can simply translate the stories into many languages to build libraries of books. Bloom books can be published as PDFs that are printed, ePubs, Bloom Reader apps (to be read on Android devices) and web sites.

Bloom Enterprise provides extra functionality to create books for the deaf, comprehension questions to ascertain if the readers understand, branding for organizations who sponsor literacy programs, and a dashboard to understand what books are being read, for how long, and how comprehension is trending.

P.1.4: Creating a Synthetic Te Reo Māori Voice
Authors: Isabella Shields, Catherine Watson, Peter Keegan, Rebekah Berriman and Jesin James
Country: New Zealand

Abstract: We have made a synthetic male te reo Māori voice, which runs on MaryTTS and synthesises speech from any Māori text. The voice was created from recordings of 1030 sentences, chosen to ensure full diphone coverage, from the story Ngā Mahi a Ngā Tūpuna. The recordings were made in a soundproof booth and condensed to two hours of continuous speech. Phonetic labelling was first determined automatically using a model obtained from Montreal Forced Aligner. The labelled data, along with a 10,000-word Māori lexicon with phonetic transcription and stress mark up, is passed into MaryTTS to create a synthetic voice.

Native language: Kua hangaia e mātou he reo rorohiko, he reo tāne. Ka noho tēnei reo rorohiko ki roto i a te pūmanawa MaryTTS. Ka whakaurua e tangata he tuhinga, ā, ka hurihia he reo, arā he kōnae tangi. Kua hopukia tēnei ki roto i tētahi wharau karo tangi. I ahu mai ngā kōrero i te pukapuka Ngā Mahi a Ngā Tupuna, ā, kua wehewehe ngā kōrero kia rerenga kōrero poto (1030). Ka āta whiriwhiria ētahi anō kia mau te katoa o ngā tangi oro o te reo Māori. Kia hangai atu ngā oro tuhi ki ngā oro tangi, ka whakamahia te Montreal Forced Aligner, nā te tangata anō i atā tirotiro te tika o ngā hononga. Ka tukua ēnei raraunga ki a MaryTTS, ka tukua hoki he papakupu whakaahua (10,000 + ngā kupu).

P.2.1: Poio - Open Source Technology for Language Diversity
Authors: Peter Bouda
Country: Portugal

Abstract: The Poio project publishes open source tools to support under-resourced languages on computers and mobile devices. Our main product is the Poio Text Prediction, a corpus-based text input support system that simplifies the way people enter text in any language. We believe that technology should assist the renewal of local languages and cultures by allowing people to actively teach, learn, extend, and spread their language in their community. Our aim is to give people the ability to use their mother tongue in everyday electronic communication, no matter where they are or what language they speak.

Native language: Das Projekt Poio entwickelt Open-Source-Lösungen um Sprachen auf Computern und mobilen Geräten zu unterstützen für die es nur wenige Daten gibt. Unser Hauptprodukt ist Poio Text Prediction, ein korpus-basiertes Eingabeunterstützungssystem, das jedem die Eingabe jeder beliebigen Sprache erleichtert. Wir glauben, dass Technologie die Erneuerung lokaler Sprachen und Kulturen unterstützen sollte, indem sie Menschen erlaubt ihre Sprache in ihrer Gemeinschaft zu lehren, zu lernen, zu erweitern und zu verbreiten. Unser Ziel ist es allen Menschen die Möglichkeit zu geben ihre Sprache in alltäglicher, elektronischer Kommunikation zu verwenden egal wo sie sich befinden oder welche Sprache sie sprechen.

P.2.2: Towards a Global Lexicographic Infrastructure
Authors: Simon Krek, Thierry Declerck, John Philip McCrae and Tanja Wissik
Country: Germany

Abstract: The poster describes the European Lexicography Infrastructure (ELEXIS), showing how it includes, integrates and cross-links also non-European dictionaries, and how anyone can contribute, either with data, scientific exchange or with an institutional cooperation, which can be implemented in the form of Observer status.
ELEXIS provides for tools and services to develop new dictionary databases or process and enrich existing ones. It also provides for grants for short scientific visits for helping the visitor to get acquainted with the ELEXIS technologies. With its linking strategies, ELEXIS ensures that lexical data of each language is getting a high visibility and accessibility.

Native language: L'affiche décrit l'infrastructure lexicographique européenne (ELEXIS), en montrant comment elle inclut, intègre et interconnecte les dictionnaires (également non européens), et comment chacun peut contribuer, soit par des données, des échanges scientifiques ou une coopération institutionnelle, qui peut être mise en œuvre sous la forme du statut d'observateur.
ELEXIS fournit des outils et des services pour développer de nouvelles bases de données de dictionnaires ou pour traiter et enrichir des bases de données existantes. Il prévoit également des bourses pour de courtes visites scientifiques afin d'aider le visiteur à se familiariser avec les technologies ELEXIS. Grâce à ses stratégies de mise en relation, ELEXIS assure une grande visibilité et accessibilité aux données lexicales de chaque langue.

P.2.3: Tooling up a less-resourced language with NLP : the example of Corsican and the "Banque de Données Langue Corse" (BDLC, Corsican Language Database)
Authors: Laurent Kevers, Stella Retali-Medori, Florian Guéniot and A. Ghjacumina Tognotti
Country: France

Abstract: The current situation regarding the existence of Natural Language Processing (NLP) resources and tools for Corsican reveals their virtual non-existence. Our inventory contains only a few rare digital resources, lexical or corpus databases, requiring adaptation work. Our objective is to use the BDLC project to improve the availability of resources and tools for the Corsican language. We have defined a roadmap setting out the actions to be undertaken: collection of corpora and setting up of a consultation interface (concordancer), language detection tool, electronic dictionary and part-of-speech tagger. The first achievements are already available.

Native language: L’état des lieux concernant les ressources et outils de Traitement Automatique du Langage (TAL) pour le corse révèle leur quasi inexistence. Notre inventaire ne contient que quelques rares ressources digitales, bases de données lexicales ou corpus, nécessitant un travail d’adaptation. Nous nous appuyons sur le projet BDLC pour faire avancer l’outillage de la langue corse. Nous avons défini une feuille de route reprenant les actions à entreprendre : collecte de corpus et mise en place d’une interface de consultation (concordancier), outil de détection de langue, dictionnaire électronique et outil d’annotation en parties du discours. Les premières réalisations sont déjà disponibles.

P.2.4: Language Technology Program for Icelandic
Authors: Anna Nikulásdóttir
Country: Iceland

Abstract: On the 1st of October 2019, work on a five-year Project Plan for Icelandic LT started. The project aims at the development of LT resources and infrastructure software, including speech technologies, machine translation and spell and grammar checking systems. It emphasizes cooperation between academia and industries, with the aim of open and usable software and resources for LT-products being delivered at the end of the program. The self-owned foundation Almannarómur conducts the program on behalf of the Icelandic Government. The research and development work is carried out by teams built across a consortium of nine universities, institutions, and private companies.

Native language: Þann 1. Október 2019 hófst vinna við fimm ára Verkáætlun í máltækni fyrir íslensku. Markmið áætlunarinnar er að þróa málföng og innviði fyrir máltækni, þar sem áhersla verður lögð á taltækni, vélþýðingar og málrýni. Lögð er áhersla á samstarf milli háskólasamfélagsins og fyrirtækja með það að markmiði að áætlunin skili opnum og nothæfum málföngum og hugbúnaði til notkunar í hugbúnaði sem þarfnast máltækni. Sjálfseignarstofnunin Almannarómur stýrir áætluninni fyrir hönd Ríkisstjórnar Íslands. Rannsóknar- og þróunarvinna er unnin af teymum sem mynduð eru í samstarfi níu háskóla, stofnanna og einkafyrirtækja.

P.2.5: A speaking atlas of indigenous languages of France and its Overseas

Authors: Philippe Boula de Mareüil, Gilles Adda, Albert Rilliard and Frédéric Vernier

Country: France

Abstract: The objective is to valorise the linguistic diversity of France through field recordings, a computer-based visualisation of dialectal areas and orthographic transcripts (which represent an object of research in itself). We describe here a website (https://atlas.limsi.fr) presenting interactive maps of Metropolitan France and its Overseas, from which the Aesop fable "The Northwind and the Sun" can be listened to and read in over 300 versions, in regional languages. There is thus both a scientific dimension and a heritage dimension in this work, since a number of regional or minority languages are in a critical situation.

Native language: L'objectif est de montrer et de valoriser la diversité linguistique de la France à travers des enregistrements recueillis sur le terrain, une réalisation informatique (qui permet de visualiser les aires dialectales) et un travail de transcription orthographique. Nous décrivons ici un site web (https://atlas.limsi.fr) présentant des cartes interactives de France hexagonale et des Outre-mer, à partir desquelles la fable d'Ésope «La bise et le soleil » peut être écoutée et lue dans plus de 300 versions, en langues régionales. Il y a ainsi une dimension à la fois scientifique et patrimoniale à ce travail, dans la mesure où un certain nombre de langues régionales ou minoritaires sont en situation critique.

P.2.6: Software and Linguistic Resources for the Tatar language preservation and development: Regional Experience

Authors: Dzhavdet Suleymanov, Aidar Khusainov and Rinat Gilmullin

Country: Russian Federation

Abstract: The poster contains information about the most significant program developments and linguistic resources of the Institute of Applied Semiotics of Tatarstan Academy of Sciences, aimed at supporting the Tatar language in information technologies: the National Corpus of the Tatar language "Tugan tel", the Russian-Tatar machine translation system Tatsoft, Tatar speech synthesis and recognition systems and mobile applications. The main results achieved within of the State program for the preservation, study and development of the state languages of the Republic of Tatarstan and other languages in the Republic of Tatarstan.

Native language: Постерда Татарстан Республикасы Гамәли семиотика Институты тарафыннан башкарылган иң әһәмиятле программа һәм лингвистик ресурслар турында мәгълумат күрсәтелә. Алар арасында – "Туган тел" татар гомумтел корпусы, "Татсофт" русча-татарча тәрҗемәче программа, Татар сөйләмен тавышландыру һәм текстка күчерү системалары, Смартфоннар өчен клавиатура һәм сүзлекләр кушымталары.
Башкарылган эшләрнең төп нәтиҗәләре Татарстан Республикасы дәүләт телләрен һәм Татарстан Республикасындагы башка телләрне саклау, өйрәнү һәм үстерү буенча Татарстан Республикасы дәүләт программасын гамәлгә ашыру кысаларында гамәлгә ашырылды.

P.2.7: Contribution to the Universal Dependencies Treebank of Non-Standard Romanian Texts

Authors: Victoria Bobicev, Catalina Mărănduc, Tudor Bumbu, Ludmila Malahov, Alexandru Colesnicov and Svetlana Cojocaru

Country: Republic of Moldova

Abstract: Cultural heritage preservation is the one non-transferable duty of any given ethnic or social entity, for it is the essence that defines and identifies each one of them among others. In the specific case of the preservation of culturally significant works of writing, this task includes not only digitizing old books to prevent their loss but also optical character recognition, transliteration of old texts and their annotation. We report our latest contribution to the development and enrichment of a universal dependencies (UD) treebank which contains old texts, regional folklore and other non-standard texts from Moldova and Romania.

Native language: Păstrarea patrimoniului cultural este datoria netransmisibilă a oricărei entități etnice sau sociale, deoarece este esența care o definește și identifică. În cazul specific al conservării operelor literare semnificative din punct de vedere cultural, această sarcină include nu numai digitalizarea cărților vechi pentru a preveni pierderea lor, dar și recunoașterea optică a caracterelor, transliterarea textelor vechi și adnotarea lor. Raportăm contribuția noastră recentă la dezvoltarea Treebank-ului de dependențe universale (UD) care conține texte vechi, folclor regional și alte texte non-standard din Moldova și România.

P.2.8: Inquiring about digital use and usability of minority languages: the approach of the Digital Language Diversity Project

Authors: Claudia Soria and Cor van der Meer

Country: Italy

Abstract: We present the results of the Digital Language Diversity Project survey about the digital behaviour, desires, and expectations of minority language speakers. The survey is designed around three conceptual blocks: the digital capacity of the language, its digital opportunities, and speakers' attitudes towards digital use of the language. We believe that the DLDP initiative has the potential to be extended to other languages and deserves to be considered by the community at large as a useful tool for digital language planning. See also http://wp.dldp.eu/reports-on-digital-language-diversity-in-europe/.

Native language: Presentiamo i risultati del sondaggio del Digital Language Diversity Project sul comportamento digitale, i desideri e le aspettative dei parlanti di alcune lingue minoritarie. Il sondaggio è strutturato attorno a tre blocchi concettuali: la capacità digitale del linguaggio, le sue opportunità digitali e gli atteggiamenti dei parlanti nei confronti dell'uso digitale del linguaggio. Riteniamo che l'iniziativa DLDP abbia il potenziale per essere estesa ad altre lingue e meriti di essere considerata dalla comunità in generale come uno strumento utile per la pianificazione linguistica digitale. Vedi anche http://wp.dldp.eu/reports-on-digital-language-diversity-in-europe/.

P.2.9: Language Technologies for Istro-Romanian

Authors: Patricia Serbac
Country: Romania

Abstract: Istro-Romanian is a minority language spoken in the peninsula of Istria, Croatia. It is a severely endangered language, with about 100 speakers left. The last speakers originate from woo Istrian villages, but they have spread now in the rest of the peninsula, in Croatia, in the neighbouring countries and in the world. Due to emmigration, there are Istro-Romanians also in the US and in Australia. The poster presents the own audio and video corpus of Istro-Romanian, the webpage of the community, mail communication etc. The new language technologies could help keeping these speakers together and making them one community.

Native language: Istro-româna e o limbă minoritară vorbită în peninsula Istria, din Croația. E o limbă în pericol grav, având în jur de 100 de vorbitori. Ultimii vorbitori provin din două sate din Istria, dar ei s-au răspândit în restul peninsulei, în Croația, în țările învecinate și în lume. Din cauza emigrării, există istro-români și în SUA și în Australia. Posterul prezintă propriul corpus audio and video al istro-românei, pagina de internet a comunității, comunicarea pe mail etc. Noile tehnologii despre limbă îi pot ajuta pe acești vorbitori să se adune într-o singură comunitate.

P.2.10: Innovative CALL Solutions and The Sustainability of "Nano" Languages in the West-Nordic Arctic Region

Authors: Birna Arnbjörnsdóttir and Auður Hauksdóttir
Country: Iceland

Abstract: This poster describes five innovative CALL solutions that support multilingualism and „nano“ languages in the West-Nordic Region. The non-language specific platforms are developed at the Vigdís Finnbogadóttir Institute at the University of Iceland and include open curated language courses, www.icelandiconline.com, and www.faroese.fa; and tools that enhance oral fluency and communicative skills in Danish as a second and third language in the West-Nordic region. The tools include: www.talboblen.hi.is, that focuses on oral language skills, www.talerum.is , an interactive game based program that encourages interaction, and www.frasar.net a resource that teaches the pragmatics of phrases.

Native language: Veggspjaldið lýsir fimm nýstárlegum tæknilausnum sem styðja fjöltyngi og heimamál á Vestnorræna málsvæðinu. Kerfin eru þróuð við Stofnun Vigdísar Finnbogadóttur við Háskóla Íslands og eru óháð tungumálum. Lýst verður opnum stýrðum tungumálanámskeiðum í íslensku og færeysku, www.icelandiconline.com og www.faroese.fa; auk tæknilausna sem styðja við fjöltyngi og auka munnlega færni í dönsku sem öðru og þriðja máli. Þau kerfi eru www.talboblen.hi.is sem eflir talmál og framburð, www.talerum.is sem er gagnvirkur tölvuleikur sem hvetur til ritunar og talmálsnotkunar og www.frasar.net sem eflir viðeigandi notkun frasa á íslensku og dönsku.

P.2.11: Indigenous/Minority Language Keyboard and Spell Checking Support, for Desktop and Mobile Operating Systems

Authors: Brendan Molloy
Country: Sweden

Abstract: Can you spell out your own name with your current keyboard? Members of several indigenous language communities cannot, ignored by the technological revolution in how we communicate.

Divvun, in collaboration with The Techno Creatives, provides keyboards for mobile and desktop environments, with spell checking integration for Sámi languages and other languages with very complex morphology. We even include Mansi support for this conference!

Our keyboard integration for iOS is so robust you can test our minority language support through QR codes provided on this poster.

Experience a level of minority language support Google and Apple have failed to provide.

Native language: Kan du skriva namnet ditt med tastaturet du har framfor deg? Folk frå mange urfolkssamfunn kan ikkje det, ignorerte som dei er av den tekniske revolusjonen i måten vi kommuniserer på.

Divvun-gruppa, i lag med The Techno Creatives, tilbyr tastatur for mobiltelefonar og datamaskiner, med innebygd stavekontroll for samiske og andre språk med svært kompleks morfologi. Vi har til og med lagt til støtte for mansisk for denne konferansen!

Tastaturingegreringa vår for iOS er so robust at du kan testa støtta for minoritetsspråk med QR-kodane som finst på denne plakaten.

Opplev eit nivå på minoritetsspråkstøtta som Google og Apple ikkje har lyktes med å levera.

2nd Native language: ᚴᛆᚿ ᛑᚢ ᛋᚴᚱᛁᚡᛆ ᚿᛆᛘᚿᛂᛏ ᛑᛁᛏ ᛘᛂᛑ ᛏᛆᛋᛏᛆᛏᚢᚱᛂᛏ ᛑᚢ ᚼᛆᚱ ᚠᚱᛆᛘᚠᚮᚱ ᛑᛂᚵ? ᚠᚮᛚᚴ ᚠᚱᚨ ᛘᛆᚿᚵᛂ ᚢᚱᚠᚮᛚᚴᛋᛋᛆᛘᚠᚢᚿ ᚴᛆᚿ ᛁᚴᛃᛂ ᛑᛂᛏ, ᛁᚵᚿᚮᚱᛂᚱᛏᛂ ᛋᚮᛘ ᛑᛂᛁ ᛂᚱ ᛆᚡ ᛑᛂᚿ ᛏᛂᚴᚿᛁᛋᚴᛂ ᚱᛂᚡᚮᛚᚢᛋᛃᚮᚿᛂᚿ ᛁ ᛘᚨᛏᛂᚿ ᚡᛁ ᚴᚮᛘᚢᚿᛁᛋᛂᚱᛂᚱ ᛔᚨ.

Divvun-ᚵᚱᚢᛔᛆ, ᛁ ᛚᛆᚵ ᛘᛂᛑ The Techno Creatives, ᛏᛁᛚᛒᛦᚱ ᛏᛆᛋᛏᛆᛏᚢᚱ ᚠᚮᚱ ᛘᚮᛒᛁᛚᛏᛂᛚᛂᚠᚮᚿᛆᚱ ᚮᚵ ᛑᛆᛏᛆᛘᛆᛋᚴᛁᛃᚿᛂᚱ, ᛘᛂᛑ ᛁᚿᛂᛒᛦᚵᛑ ᛋᛏᛆᚡᛂᚴᚮᚿᛏᚱᚮᛚ ᚠᚮᚱ ᛋᛆᛘᛁᛋᚴᛂ ᚮᚵ ᛆᚿᛑᚱᛂ ᛋᛔᚱᚨᚴ ᛘᛂᛑ ᛋᚡᛅᚱᛏ ᚴᚮᛘᛔᛚᛂᚴᛋ ᛘᚮᚱᚠᚮᛚᚮᚵᛁ. Vᛁ ᚼᛆᚱ ᛏᛁᛚ ᚮᚵ ᛘᛂᛑ ᛚᛆᚵᛏ ᛏᛁᛚ ᛋᛏᚯᛏᛂ ᚠᚮᚱ ᛘᛆᚿᛋᛁᛋᚴ ᚠᚮᚱ ᛑᛂᚿᛂ ᚴᚮᚿᚠᛂᚱᛆᚿᛋᛂᚿ!

ᛏᛆᛋᛏᛆᛏᚢᚱᛁᚿᛏᛂᚵᚱᛂᚱᛁᚿᚵᛆ ᚡᚨᚱ ᚠᚮᚱ iOS ᛂᚱ ᛋᚮ ᚱᚮᛒᚢᛋᛏ ᛆᛏ ᛑᚢ ᚴᛆᚿ ᛏᛂᛋᛏᛆ ᛋᛏᚯᛏᛆ ᚠᚮᚱ ᛘᛁᚿᚮᚱᛁᛏᛂᛏᛋᛋᛔᚱᚨᚴ ᛘᛂᛑ QR-ᚴᛃᚮᛑᛆᚿᛂ ᛋᚮᛘ ᚠᛁᚿᛋᛏ ᛔᚨ ᛑᛂᚿᛂ ᛔᛚᛆᚴᛆᛏᛂᚿ.

ᚮᛔᛚᛂᚡ ᛂᛁᛏ ᚿᛁᚡᚨ ᛘᛁᚿᚮᚱᛁᛏᛂᛏᛋᛋᛔᚱᚨᚴᛋᛏᚯᛏᛆ ᛋᚮᛘ Google ᚮᚵ Apple ᛁᚴᛃᛂ ᚼᛆᚱ ᛚᛦᚴᛏᛂᛋ ᛘᛂᛑ ᚨ ᛚᛂᚡᛂᚱᛆ.

P.2.12: MultiTAL : an online platform to list NLP tools for under-resourced languages

Authors: Damien Nouvel, Driss Sadoun and Mathieu Valette

Country: France

Abstract: Diversity and variety of human languages raises indisputable difficulties for processing textual data. Regarding under-resourced languages, many softwares have been implemented, but many are poorly referenced and documented. The ERTIM (INALCO) lab has published in 2016 a website (http://multital.inalco.fr) that addresses this issue. Our website lists tools available for languages. For each software, the database provides information concerning : NLP tasks, implemented method, OS compatibility, among others. We do not pretend to be exhaustive, but people populating the database are speakers of concerned languages, they downloaded and tested softwares, and provided technical information for their installation and use.

Native language: La diversité et la variété des langues humaines donne d'incontestables difficultés pour le traitement de données textuelles. Concernant les langages peu dotés, de nombreux logiciels ont été implémentés, mais beaucoup restent peu référencés et mal documentés. L'équipe ERTIM a mis en ligne en 2016 un site (http://multital.inalco.fr) qui réalise ce travail. En faisant la liste des outils par langage, cette base de données apporte des informations sur leur utilisation. Nous ne prétendons pas être exhaustifs, mais les personnes remplissant la base étaient locuteurs des langues concernées, elles ont téléchargé et testé les outils, et ont renseignés des informations sur leur installation et leur utilisation.

P.2.13: Automatic Recognition of mixed Ukrainian-Russian Speech

Authors: Valeriy Pylypenko and Tetyana Lyudovyk

Country: Ukraine

Abstract: This work presents an approach to recognition of conversational speech with code-switching which is widespread in Ukraine now. Both inter-sentential and intra-sentential Ukrainian/Russian code-switching is handled. The approach takes into account closely related Russian and Ukrainian phonetic systems. A cross-lingual ASR system is developed. The acoustic model and pronunciation lexicon are based on Ukrainian phone set. Experiments with different types of code-switching speech (Parliamentary, TV broadcast) were conducted and results are presented. The approach is suitable especially in cases of intra-sentential code-switching where language identification is problematic.

Native language: Ця робота представляє підхід до розпізнавання усного мовлення з переключенням між українською та російською мовами, яке зараз поширене в Україні. Обробляється як міжфразове перемикання, так і всередині фраз. Підхід враховує особливості фонетичних систем тісно пов'язаних російської та української мов. Розроблена багатомовна система автоматичного розпізнавання мовлення. Акустична модель та лексика вимови базуються на українській множині фонем. Представлені результати розпізнавання мовлення з переключенням мови з декількох джерел (Парламентська, ТВ трансляція). Підхід особливо корисний у випадках перемикання мови всередині фраз, де ідентифікація мови є проблематичною.

P.2.14: Apertium: a free/open-source platform for machine translation and basic language technology

Authors: Mikel L. Forcada and Francis Tyers

Country: Spain

Abstract: Apertium is a free/open-source platform for rule-based machine translation and basic language technology. Since 2005, Apertium provides a free/open-source, modular, language-independent machine translation engine, free/open-source linguistic data for a variety of languages and language pairs, with emphasis on less-resourced languages, and free/open-source tools to manage language data, learn rules, and build machine translation engines. The Apertium pipeline contains monolingual modules useful in other human-language technology tasks. The license chosen, the GPL, avoids private appropriation and encourages giving improvements back to the project, creating a community. Apertium is an active research and business platform, and provides a series of stand-alone products.

Native language: Apertium és una plataforma lliure/de codi obert per a la traducció automàtica basada en regles i per a tecnologies bàsiques de la llengua. Des del 2005, Apertium proporciona un motor de traducció automàtica lliure / de codi obert, modular, independent de la llengua, dades lingüístiques lliures / de codi obert per a diversos idiomes i parells d'idiomes, amb èmfasi en llengües amb menys recursos i eines lliures/de codi obert per gestionar dades lingüístiques, aprendre regles i crear motors de traducció automàtica. El 'pipeline' d'Apertium conté mòduls monolingües útils en altres tasques de tecnologia del llenguatge humà. La llicència escollida, la GPL, evita l'apropiació privada i incita a retornar millores al projecte, creant una comunitat. Apertium és una plataforma activa de recerca i negocis i ofereix una sèrie de productes per a usuaris finals.

P.2.15: REDISCOVERING PAST NARRATIONS: THE ORAL HISTORY OF THE ROMANIAN LANGUAGE PRESERVED WITHIN THE NATIONAL PHONOGRAMIC ARCHIVE

Authors: Oana Niculescu, Maria Marin and Daniela Răuțu

Country: Romania

Abstract: Archive. A monumental work, AFLR is the richest, most inclusive and diversified collection of ethno-linguistic recordings in Romania. Only a third of the data have been processed so far and there is a pressing need to digitize the remaining tape recordings. Through the preservation of the archive we can gain access to both individual and collective memories, aiding to a better understanding of our cultural heritage and, at the same time, restoring missing or forgotten pieces of Europe's oral history.

Native language: În această prezentare ne propunem să atragem atenția asupra necesității conservării și digitalizării Arhivei fonogramice a limbii române. În momentul de faţă, AFLR este cea mai bogată şi cuprinzătoare colecţie de texte dialectale din România. Cu toate acestea, doar o treime din material a fost digitalizat, existând riscul ca benzile rămase să se deterioreze, ducând la pierderea înregistrărilor. Protejarea arhivei AFLR contribuie, pe de o parte, la recuperarea narațiunilor individuale și colective, iar, pe de altă parte, la înțelegerea și valorificarea patrimoniului cultural, respectiv redobândirea unor elemente pierdute sau uitate din istoria orală a Europei.

P.2.16: Language technology for indigenous languages: Achievements and challenges
Authors: Sjur Moshagen, Lene Antonsen and Trond Trosterud
Country: Norway

Abstract: Fifteen years of indigenous language technology development by UiT/Sámi Parliament has resulted in spelling and grammar checkers, desktop/mobile keyboards, morphological analysers, MT, speech synthesis, language learning tools and intelligent electronic dictionaries.

This was facilitated by an open source language independent infrastructure, targeted at languages with rich and complex grammar, with integration for host operating systems and apps.

The current primary challenge is integration with closed platforms where we cannot currently support user needs.

Our proposed solution is a “Manifesto for Open Language Technology”, where APIs, localisations and source code are open, while ensuring community intellectual property custodianship, engagement and commitment.

Native language: UiT/Sámedikki 15 jagi eamiálbmot giellateknologiija barggu bohtosat leat sátne- ja grammatihkkadivvunprográmmat, boallobeavdi dihtorii ja mobiltelefovdnii, morfologalaš analysáhtorat, dihtorjorgaleapmi, hállansyntesa, giellaoahppanreaiddut ja intelligeanta digitála sátnegirjjit.

Dát lea huksejuvvon rabas gáldokoda infrastruktuvrras, mii lea heivehuvvon gielaide main lea rikkes ja kompleaksa grammatihkka – infrastruktuvra mii siskkilda geavahanlavttaid ja applikašuvnnaid.

Dál váldohástalus lea integreret prográmmaid giddejuvvon geavahanvuogádagaide, maid siste mii dál eat beasa doarjut geavaheddjiid dárbbsuid.

Min evttohus lea "Rabas giellateknologiija manifesta", mas API:t, lokaliseren ja gáldokoda leat rabas, muhto seammás giellaservodagat galget hálddašit gáldokoda intellektuealla rivttiid.

P.2.17: Using technology to empower Indigenous knowledge sharing
Authors: Lorna Williams, Tracey Herbert and Daniel Yona
Country: Canada

Abstract: B.C. is an Indigenous language hot spot with 34 languages. In response to threats to the vitality of these languages, communities in B.C. have adopted collaborative approaches to language revitalization and technology. FirstVoices is an example of a collaborative, community-led language revitalization platform where communities manage, curate and control their data. The platform encourages youth to connect with fluent elders, in order to share their knowledge in a digital space. This poster presents realities, challenges, and opportunities of language revitalization and the ripple effect technology, such as keyboards or shared platforms, can have on access to language.

P.2.18: European Language Monitor by EFNIL
Authors: Sabine Kirchmeier
Country: Denmark

Abstract: The poster presents The European Language Monitor (ELM) - a key project of EFNIL, the European Federation of National Institutions for Language. ELM is an online database containing data on language legislation and language planning in Europe. The user can browse and compare language legislation, information on minority languages, and provisions for language use in the educational systems, in business, in the media and for language technology.
ELM contains comments, quotes, links and translations of national legislation wherever possible. The data for ELM are collected every 4 years. The current version, ELM 4, is based on data collected in 2017-2018.

Native language: Denne poster præsenterer The European Language Monitor (ELM) – et nøgleprojekt for EFNIL, den europæiske sammenslutning af officielle sproginstitutioner. ELM er en online database som indeholder data om sproglovgivning og sprogplanlægning i Europa. Brugeren kan søge og sammenligne sproglovgivning, information om minoritetssprog og regulering af sprogbrugen i uddannelsessystemet, i erhvervslivet, i medierne og inden for sprogteknologi. ELM indeholder kommentarer, citater, links og oversættelser af national lovgivning hvor det er muligt. Data til ELM indsamles hvert 4. år. Den nuværende version, ELM 4, er baseret på data som er indsamlet I 2018-2018.

P.2.19: Preserving Endangered European Cultural Heritage and Languages Through Translated Literary Texts
Authors: Amel Fraisse, Ronald Jenn, Shelley Fisher Fishkin and Zheng Zhang
Country: France

Abstract: ResOurceS for Endangered languages Through TranslAted texts (ROSETTA) is a collaborative and interdisciplinary project. Very much as the Rosetta stone helped decipher the demotic and hiero- glyphic scripts thanks to the presence of the Greek translation, the Rosetta project intends to preserve contemporary endangered languages and assist with their survival through translation. Our project puts to use the extant translated versions of a single literary text into a number of endangered languages over a rather long period of time. A first experiment was conducted on existing Basque translations of the well-traveled American novel “Adventures of Huckleberry Finn”.

Native language: روزيتا هو مشروع تعاوني ومتعدد التخصصات. بقدر ما ساعد حجر رشيد في فك رموز النصوص الديموغرافية والهيروغليفية بفضل وجود الترجمة اليونانية، يعتزم مشروع روزيتا على الحفاظ على اللغات المهددة بالإنقراض والمساعدة في بقائها من خلال الترجمة. يهدف مشروعنا إلى استخدام الترجمات الموجودة لنص أدبي واحد في عدد من اللغات المهددة بالإنقراض على مدى فترة زمنية طويلة. أجريت أول تجربة في هذا المشروع على ترجمات الرواية الأمريكية مغامرات هاكلبري فين المتوفرة في اللغة الباسكية المهددة بالانقراض.

P.2.20: Towards ASR that recognises everyone in a country with no spoken standard

Authors: Benedicte Haraldstad Frostad

Country: Norway

Abstract: Norwegian has many dialects, two written and no spoken standard. Speakers are used to linguistic diversity, and dialects are strong identity markers. Changing one's dialect is associated with identity loss. The extra costs, need for specialised expertise in spoken Norwegian and lack of suitable lexica and speech data sets complicate the development of ASR-products for this language community. This poses a democratic problem as public institutions automatise dictation and integrate ASR as a means for interaction. The Language Council is initiating innovative projects to improve ASR for Norwegian and minority languages in Norway and wishes to exchange ideas and experiences.

Native language: Norsk har to offisielle skriftnormer, mange dialekter og ingen offisiell uttalenorm. Språkbrukerne er vant til språklig mangfold, og dialekter er sterke identitetsmarkører. Å endre på dialekten sin forbindes med identitetstap. De ekstra kostnadene, den spesialiserte ekspertisen i norsk talespråk som trengs, og mangelen på passende leksika og taledatasett vanskeliggjør utvikling av talegjenkjenningsprodukter for dette språksamfunnet. Det er et demokratisk problem ettersom offentlige institusjoner innfører automatiske dikteringsverktøy og integrerer talegjenkjenning i sine kommunikasjonskanaler. Språkrådet har tatt initiativet til nyskapende prosjekter skreddersydd for å øke kvaliteten på talegjenkjenning for norsk og minoritetsspråk i Norge og ønsker å utveksle idéer og erfaringer.

P.2.21: Komi Latin-Alphabet Letters Not Found in Unicode

Authors: Jack Rueter

Country: Finland

Abstract: The two literary languages Komi-Permyak and Komi-Zyrian have used numerous alphabets and orthographies from the 13th century on. There are approximately six years of extensive publications in without recognizable texts. The typography is considered inconsistently poor, and therefore it can be categorized as a transitional alphabet. The term transitional alphabet, in turn, means that Unicode characters from mixed ranges can be used to satisfy many of the missing letters. The poster will illustrate missing letters in the Latin range with discussion and derive the minimal alphabetical requirements for the documentation of these two Uralic languages of Russia.

Native language: Кыкнан коми кывъяслӧн татчӧдз гижанногныс уна вӧлісны дас коймӧд нэмсянь. Оз тырмыны UNICODE-ын шыпасъяс квайт кымын волӧн гижӧдъяслы, кодъяс вывті коланаӧсь 1930-ӧд вояссянь. Шуӧны типографияыс пӧ омӧль, да та вӧсна колӧ чайтыны, мый вужӧдана тайӧ латиницаӧн гижанногыс. Вужӧдана – тайӧ позяс вӧдитчыны быдлаысь шыпасъястӧ – латиницаысь да кириллицаысь тшӧтш. Петкӧдлам став тырмытӧм шыпасъяс, кодъяс пиысь медся этша, мый колӧ, медым Россияса на кыкнан урал кывъяслы 6 во лыддянтортӧ сетны электроннӧй ногӧн.

P.2.22: Developing technologies for low-resource Uralic languages: Case studies on Saami and Komi varieties

Authors: Niko Partanen, Michael Rießler and Thierry Poibeau

Country: Finland

Abstract: The Uralic languages are spoken in northern Eurasia, and almost all of them are endangered. Language technology can play a major role in documenting and describing these languages better, and in making related workflows faster and more efficient. However, applying modern methods effectively in this context remains a challenge.

We have developed language technology for Komi and Saami, with a focus on a low-resource scenario. Besides providing an overview of this work, we detail what we see as the main challenges. Although we focus on individual languages, our experiences also translate to the wider situation of endangered languages outside Eurasia.

Native language: Uralilaisia kieliä puhutaan laajalla alueella Pohjois-Euraasiassa, ja valtaosa niistä on uhanalaisia. Kieliteknologialla voi olla merkittävä rooli näiden kielten kuvaamisessa ja dokumentaatiossa, erityisesti tehden näihin toimiin liittyvistä käytännöistä tehokkaampia ja nopeampia. Kieliteknologian nykysovellusten hyödyntämisessä tässä kontekstissa on silti runsaita ratkaisemattomia haasteita.

Työryhmämme on kehittänyt kieliteknologiaa saamelaiskielille ja komille, erityisesti tilanteeseen, jossa käytettäviä resursseja on vähän. Kuvaamme aiemmin tehdyn työn sekä keskeisimmät ongelmakohdat. Vaikka keskitymme yksittäisiin kieliin, ovat kokemuksemme sovellettavissa myös muihin uhanalaisiin vähemmistökieliin oman alueemme ulkopuolella.

P.2.23: Understanding culture and society with the language resources and tools offered through the CLARIN Research Infrastructure

Authors: Maria Eskevich and Franciska de Jong

Country: Netherlands

Abstract: Europe's Common Language Resources and Technology Infrastructure (CLARIN) aims at making language resources and tools from all over Europe and beyond accessible for research purposes through a single sign-on platform. CLARIN supports academic researchers, students, journalists and citizen-scientists interested in digital language resources, such as parliamentary records, social media data, newspaper archives and spoken corpora, and also functions as a knowledge sharing ecosystem. CLARIN adheres to the FAIR data principle. Open access to digital language resources that capture social and cultural diversity can help advance the social sciences and humanities at large.

Native language: De Europese onderzoeksinfrastructuur CLARIN (Common Language Resources and Technology Infrastructure) maakt taaldata vanuit de hele wereld en digitale analysetools voor taal via een 'single sign-on' platform toegankelijk voor onderzoeksdoeleinden. CLARIN ondersteunt academische onderzoekers, studenten, journalisten en citizen-scientists die gebruik maken van taalmaterialen (zoals parlamentaire verslagen, social media data, krantenarchieven en gesproken corpora), en functioneert tevens als een ecosysteem voor het delen van kennis. CLARIN is gebaseerd op de principes van FAIR data. Digitale taalmaterialen die vindbaar en toegankelijk zijn en hun inherente sociale en culturele diversiteit zijn van belang voor het domein van de sociale en geesteswetenschappen in brede zin.

P.2.24: A Multimodal Database of Russian Sign Language
Authors: Alexey Karpov, Ildar Kagirov, Dmitry Ryumin and Alexander Axyonov
Country: Russian Federation

Abstract: We present a multimodal database of Russian sign language (RSL) - TheRuSLan. It includes lexemes from RSL within one subject area demonstrated by 14 informants that were recorded with Kinect 2.0 sensor in FullHD video, infrared and depth map modes. RSL has an official status in the Russian Federation, and over 120K deaf people in Russia and some neighbour countries use it as their main language of spoken communication. RSL has no written system, poorly described and has very few electronic resources. We compare our database with other RSL corpora, and formulate some basic principles of gesture lexeme description.

Native language: Мы представляем многомодальную базу данных (тезаурус) русского жестового языка (РЖЯ) - TheRuSLan. База данных включает демонстрации лексем РЖЯ, относящиеся к одной предметной области и показанные 14 информантами. Данные были записаны при помощи устройства Kinect 2.0 в формате FullHD, в инфракрасном диапазоне и в режиме карты глубины. РЖЯ является одним из официальных языков общения на территории Российской Федерации, его носителями являются свыше 120 тыс. людей в России и сопредельных странах. РЖЯ не обладает системой письменности, недостаточно описан и имеет очень мало электронных ресурсов. Мы сравниваем нашу базу данных с другими корпусами РЖЯ и формулируем основные принципы описания жестовых лексем.

P.2.25: Sámi languages
Authors: Mikkel Rasmus Logje
Country: Norway

Abstract: Sámi languages are defined as a branch of the Uralic language family, and are traditionally spoken in an area stretching from central Norway and Sweden, through northern Norway, Sweden and Finland, to the Kola Peninsula in Russia. Today there are altogether 9 Sámi languages which are more or less mutually unintelligible, especially those that are geographically distant. The traditional boundaries between Sámi languages do not follow the national boundaries. The number of language users varies from one language to another. The largest language is Northern Sámi (est. 20.000–40.000 users). All Sámi languages are minority languages in the respective countries.

Native language: Sámegielat gullet urálalaš gielaide, ja daid árbevirolaš hupmanguovlu gokčá guovlluid Gaska-Ruoŧa ja Gaska-Norgga rájes, Davvi-Norgga, Davvi-Ruoŧa ja Davvi-Suoma bokte, gitta Guoládatnjárgii Ruoššas. Dál gávdnojit oktiibuot 9 sámegiela, ja gielaid gaskka leat unnit eanet erohusat, eandalii daid gielaid gaskka mat leat guhkkálaga. Sámegielaid árbevirolaš hupmanguovllut eai čuovo riikkarájiid. Leat erohusat daid iešguđetge gielaid geavaheddjiid logus. Stuorámus giella lea davvisámegiella (sullii 20.000–40.000 geavaheaddji). Buot sámegielat leat minoritehtagielat dáin guoskevaš riikkain.

P.2.26: LT Data Free for All
Authors: Marko Tadić and Tamás Váradi
Country: Hungary

Abstract: Language technology crucially depends on large amounts of texts. Digitally published text is a natural source for fast production of the fundamental language resources – corpora. However, clean, openly and freely available texts are difficult to come by. Even national languages suffer from scarcity of quality language data. We are presenting a project that can serve as a role model for the collection of large monolingual corpora for under-resourced languages. The approach could be applicable to any linguistic community that publishes legislative texts in their own language in digital form, to quickly build very big corpora

Native language: A nyelvtechnológia számára alapvető fontosságú az óriás mennaiségű szövegek elérhetősége. A digitálisan publikált szövegek kézenfekvő forrásai az alapvető nyelvi erőforrások, a korpuszok gyors előállításának. Azonban a tiszta, nyílt és ingyenesen hozzáférhető szövegeket nehéz beszerezni. Még hivatalos nemzeti nyelvek is szenvednek a jó minőségű szövegek hiányától. Bemutatunk egy olyan projektet, amely mintául szolgálhat arra, hogy miképpen lehet nagyméretű egynyelvű korpuszokat építeni erőforráshiányos nyelveken. A módszer minden olyan nyelvi közösség esetében használható, amely digitális alakban teszi közzé a saját nyelvén a jogszabályokat, melyekből hatékonyan lehet nagyon nagyméretű korpuszokat építeni.

P.2.27: Can we use a spoken Dialogue System to document Endangered Languages?
Authors: Jacqueline Brixey, Seyed Hossein Alavi and David Traum
Country: United States

Abstract: We investigate using a dialogue system to preserve endangered languages, and the viability of a multilingual dialogue system to generate a general use corpus of audio responses in . We introduce DAPEL (Dialogue APP for Endangered Languages). DAPEL elicits responses from speakers of endangered languages by having a conversation with them. We conducted a pilot user study to examine the efficacy of using an automated system like DAPEL versus a human interviewer. We also studied the effects of engaging in small-talk in a different language in between recording prompts for the target language.

P.2.28: Technologies for Endangered Languages: The Case of the Languages of Sardinia

Authors: Adrià Martín-Mor

Country: Andorra

Abstract: This poster shows the impact that technology may have on endangered languages, with a focus on Sardinian, one of the five native languages of Sardinia, according to the regional law. Specifically, it presents an example of how technology can be used to translate online texts, localise digital products and develop language resources. By resorting to free-licensed products, the output of these efforts can be re-used to generate further resources that, in turn, help increase the amount of texts generated.

Native language: Custu poster ammustrat s'impatu chi sa tecnologia podet tènnere in is limbas in perìgulu, e mescamente in sa limba sarda, una de is chimbe limbas nativas de Sardigna segundu s'istatutu regionale. In s'ispetzìficu, si presentat un'esempru de comente sa tecnologia podet èssere impreada pro bortare testos in lìnia, localizare produtos digitales e isvilupare resursas linguìsticas. Tràmite su sèberu de produtos cun lissèntzias lìberas, su resurtadu de custu traballu podet èssere aprofitadu pro generare àteras resursas chi, a su turnu suo, podent agiudare a crèschere sa cantidade de testos generados.

P.3.1: Challenges for language technologies in Ayapaneco
Authors: Jhonnatan Rangel
Country: France

Abstract: There are currently 577 critically endangered languages in the world, making up almost 10% of all languages. These languages are only spoken by a few elder speakers and are technologically low-resourced. Numde ‘oode or Ayapaneco is one of these languages, spoken by less than 11 elders in southern Mexico. Ayapaneco, like other critically endangered languages, poses various fundamental challenges including the annotation bottleneck that limits the scope of its documentation, preservation, reclamation, revitalization and utilization in language technologies. This poster addresses the challenges Ayapaneco confronts as it is vanishing before our eyes.

Native language: En el mundo hay 577 lenguas en muy alto riesgo de desaparición, constituyendo casi el 10% del total. Estas las hablan algunos adultos mayores además de que tienen pocos recursos tecnológicos. Numde ‘oode o ayapaneco, hablada por menos de 11 adultos mayores en el sureste mexicano, es una de estas. El ayapaneco, como otras lenguas en muy alto riesgo de desaparición, plantea retos fundamentales como el cuello de botella de anotación que limita las posibilidades de su documentación, mantenimiento, recuperación y uso en tecnologías del lenguaje. Este poster aborda los retos que enfrenta el ayapaneco al borde de la desaparición.

P.3.2: Mainumby: computer-assisted Spanish-to-Guarani translation
Authors: Michael Gasser
Country: United States

Abstract: Technology plays an important role in the daily work of the modern translator. However, computer-assisted translation (CAT), like machine translation, relies on extensive bilingual corpora, which are only available for a small minority of the world's languages. This poster presents a framework for the development of rudimentary CAT systems for translation into languages with limited resources and the compilation of bilingual corpora as a side-effect of the systems' use by translators. The framework has been implemented in Mainumby, a web application for CAT from Spanish to Guarani, the majority language of Paraguay.

Native language: Tuichaite mba’e niko ñe’ẽasahára ko’ağagua rembiapo pa’ũme mohendaha ha opaichagua apopyre jeporu. Upéicharamo jepe umi pojoaju, oñemboheráva ñe’ẽasa mohendaha ñepytyvõ rupive (ÑMÑ), oñemopyrenda ñe'ẽkõi retépe, ha tete kakuaaitéva ojejapo mbovymi ñe'ẽmente. Ko jehaipyre ohechauka mba’éichapa ojejapokuaa ÑMÑ ñe’ẽnguéra oñemomichĩvape ğuarã, ha omombe'u Mainumby, peteĩ apopyrã ÑMÑ oñemongakuaáva hína ojejapo porãve hağua ñe’ẽasa kastellánogui guaraníme. Upe pojoaju ikatu oipytyvõ porã tapicha ñe’ẽasahárape ha avei ombyaty umi ñe’ẽ ñe'ẽkõi rete.

P.3.3: Baby Quechua robot
Authors: Maximiliano Duran
Country: France

Abstract: A robot using artificial intelligence and a comprehensive set of linguistic resources may help to preserve Quechua. It may help in M.T of scientific, and cultural French documentation into Quechua. Written documentation, is essential to keep this language alive. I have been working on such a robot, for several years. I named it Yachaj/expert. I will show the first stage of this project: Baby Quechua Robot. Its functions are Automatic conjugation, lexical queries of Quechua-FR-SP; elementary spelling checking; and transliteration (alpha version) of texts written in the official spelling of Cuzco, Ecuador or Bolivia to that of Ayacucho and vice-versa.

Native language: Sinchillatañam Peru kitipi runasimi chintiramun kay ñawpaq pachak watallapi. Chaymi nichiwanchik: cheqappunim runasimiqa wañunayaypaq kachkan! Utjayllañam ima kutirichiykunatapas rurananchik ama kay simi wañunampaj. Allin qispichisqa, allin yachachisqa, llapan rikchaq runasimi cheqap-yachaykunawan kikin-ruraqqa yanapakuwanchikmanpunim runasimi unanchaypi. Chaymi ñuqa, kay ñawpaj qanchis watakunapi “yachachichkani” runasimita huk kikin-ruraqta. Paymi yanapakunqa mana-sasa runasimi yachachiykunata. Paytaqmi, allintaña puqurquspaqa. Payqa yanapawananchik punim Fransepi, Castellanopi qellqakunata runasimiman tikraipi, chayna achkallaña runasimipi qellqasqa taqekuna kanampaq. Cheqap-yachaymanta qellqakuna, willakuy-yachaymanta, llimpi-taki-yachaykuna qellqakuna achkallaña runasimipi qellqasqa taqekuna rikurinampaq. Chaynam michasun mana runasimi wañunapaq.

P.3.4: On the development of the Mexican Languages Parallel Corpus
Authors: Cynthia Montaño, Gerardo Sierra Martínez and Gemma Bel-Enguix
Country: Mexico

Abstract: The project we present is called Mexican Languages Parallel Corpus (CPLM) and its main goal is to contribute to development of NLP for low-resources Mexican languages. The CPLM consist of two modules: core module and subcorpus of religious and political texts module. The core module currently comprises 6 linguistics groups from 3 linguistics families; Mayan: Yucatec Maya and Ch’ol; Otomanguean: Mazatec, Zapotec and Otomí; Uto-Aztec: Nahuatl. The STRyP comprises 83 translations of the New Testament and 11 translations of three types of texts. The STRyP comprises a wide range of languages.

Native language: El proyecto que presentamos se llama Corpus Paralelo de Lenguas Mexicanas y su objetivo principal es contribuir al desarrollo de PLN para las lenguas de bajos recursos. El CPLM se compone de dos módulos: el módulo nuclear y el módulo de subcorpus de textos religiosos y políticos (STRyP). El módulo nuclear contiene actualmente seis grupos de tres familias lingüísticas; maya: maya yucateco y ch’ol; otomangue: mazateco, zapoteco y otomí, y yutoazteca: náhuatl. El STRyP se basa en 83 traducciones del nuevo testamento y once traducciones de tres tipos de textos. El STRyP se compone de un amplío rango de lenguas.

P.3.5: Project: Endless Oaxaca Multilingual

Authors: Tajëëw Díaz

Country: Mexico

Abstract: The Endless Oaxaca Multilingual project is an interdisciplinary project to bring computer equipment with the Endless operating system to indigenous communities in Oaxaca Mexico that have diverse content in the indigenous languages spoken in each community. The contents are mainly books for first readers, with the perspective of developing desktop applications in the coming months that help teachers to teach the indigenous language of the community. Rural communities in Oaxaca have very limited internet connectivity, so we plan to focus on content that may be available off line.

Native language: El proyecto de Endless Oaxaca Multilingüe es un proyecto interdisciplinario para llevar equipos de cómputo con el sistema operativo Endless a comunidades indígenas de Oaxaca México que tengan diversos contenidos en las lenguas indígenas que se hablan en cada comunidad. Los contenidos son principalmente libros para primeros lectores, con la perspectiva de desarrollar en los próximos meses aplicaciones de escritorio que ayuden a los profesores a la enseñanza de la lengua indígena de la comunidad. Las comunidades rurales en Oaxaca tienen conectividad a internet muy limitad, por lo que planeamos enfocarnos a contenido que pueda estar disponible sin necesidad de internet.

P.3.6: Large-scale audio-recordings to study infant language acquisition

Authors: Camila Scaff, Marvin Lavechin and Alejandrina Cristia

Country: France

Abstract: Studies of individual and socioeconomic variation in North America suggest that infant-directed speech quantities determine children's language advancement, inspiring interventions to get parents to talk more to their child. In this context, day-long audio-recordings analysed with proprietary software trained on American data are increasingly used to measure children's input and production, but there is little research on how fair this technique is to other languages and cultures. We present results from 10 Tsimane' children and their families (>270h audio, ~5h hand-annotated). Identification Error Rates averaged 62% (range 0-100%), inviting further work on open source diarization solutions that are retrainable.

Native language: Numerosos estudios sobre la variación individual y socioeconómica en América del Norte sugieren que las cantidades de habla dirigida a los bebés determinan el avance del lenguaje de los niños, lo cual ha inspirado intervenciones para que los padres hablen más con sus hijos. En este contexto, las grabaciones de audio de día completo analizadas con un software patentado entrenado en datos estadounidenses se utilizan cada vez más para medir la producción de los niños y cuanto se les habla, pero hay poca investigación sobre cuán justa es esta técnica para otros idiomas y culturas. Presentamos resultados de 10 niños Tsimane' y sus familias (> 270h de audio, ~ 5h anotadas a mano). Las tasas de error de identificación promediaron el 62% (rango 0-100%), invitando soluciones de diarización con código abierto y re-entrenable.

P.3.7: Nierika Red Social para aprender y enseñar una lengua indígena

Authors: Vania Ramírez

Country: Mexico

Abstract: NIERIKA is a niche social network on development, which is founded on the objective to collaborate with and support the preservation of all the Mexican indigenous languages. This unique platform enables users to present and create their publications which contains original indigenous linguistic data and it's closest translation into modern-day Spanish that does justice to the original meaning, this in turn promotes the idea of gathering linguistic data for investigative and research purposes. Nierika intends to create a digital community for members who want to share, learn, study and preserve the long lost linguistic treasures of Mexico.

Native language: NIERIKA es una red social en desarrollo, que se funda en el objetivo de colaborar y apoyar la preservación de todas las lenguas indigenas mexicanas. Esta plataforma única permite a los usuarios presentar y crear publicaciones que contienen datos originales de las lenguas indígenas y su traducción más cercana al español moderno tratando de ser fiel al significado original, esto a su vez permite recopilar datos lingüísticos con fines de investigación y estudio. Nierika tiene la intención de crear una comunidad digital para los miembros que desean compartir, aprender, estudiar y preservar los tesoros lingüísticos perdidos de México.

P.3.8: PRESERVING INDIGENOUS LANGUAGES IN SOUTH AND CENTRAL AMERICA BY LEVERAGING OPEN LICENSING AND TECHNOLOGY

Authors: Purvi Shah

Country: India

Abstract: StoryWeaver is a digital platform with 17,000+ free storybooks in 200+ languages, including 18 South and Central American languages, of which 11 are Indigenous. This poster presentation highlights how StoryWeaver's initiatives and technology catalyse the revitalisation of Indigenous languages in this region: Archiving endangered languages through storybooks - the first books published in Chocholteco in over a decade Building a sustainable self-publishing model, empowering local communities to print books in places like Oaxaca, where content creation is highly regulated Supporting communities of practice and building content repositories in 'bridge' languages like Spanish which aid translations in Indigenous language

Native language: 'StoryWeaver' es una plataforma digital con 16,000 libros gratuitos en 200 idiomas. Incluyen 18 idiomas sur y centroamericanos, de los cuales 11 son indígenas. Esta presentación destaca cómo las iniciativas y la tecnología de StoryWeaver aceleran la revitalización de las lenguas indígenas:

1. Archivar lenguas en peligro de extinción a través de libros y salvarlas de la extinción
2. Construir un modelo de autopublicación que empodera las comunidades locales para imprimir libros en lugares donde la creación de contenido está altamente regulada
3. Apoyar a las comunidades a construir repositorios en idiomas como Español, que ayudan a las traducciones en lenguas indígenas

P.3.9: Comunidad Elotl. Language Technologies for Mexico's Indigenous Languages
Authors: Ximena Gutierrez-Vasques and Victor Mijangos
Country: Mexico

Abstract: Comunidad Elotl gathers a group of enthusiasts that share the interest of generating language technologies for the languages spoken in Mexico. So far, our projects have focused in the recollection of parallel corpora, building accesible web search interfaces for these corpora and in the research and development of Natural Language Processing (NLP) Techniques for building taggers.

In this poster we summarize our main contributions, moreover, we highlight some of the main challenges that arise when dealing with these low-resource languages from a computational perspective.

Native language: (Nahuatl) Nechikol Elotl kichiua tein mo tekipachoa maj onka tajtol in amantekayotl tech Mexiko tajtolmej. Axkan sayo tik sentilia “corpus paralelos”, uan tlatemoa ika in tekit, no tik amatemoa uan tikchiua “Procesamiento del Lenguaje Natural (NLP)“. Itech in amatl tik nextia to tekit, no mo nextia ken oui etoke in tajtolmej itech in tonalmej.

P.3.10: Language and Landscape: Hiking and Documenting the Chatino Language of San Juan Quiahije
Authors: Emiliana Cruz
Country: Mexico

Abstract: A few remaining elder speakers of the San Juan Quiahije Chatino language (Oaxaca, Mexico) have unique command of specialized words, expressions, and grammatical features that relate to local landscapes and nature specific activities. In a moment marked by rapid decline of indigenous languages, this area of language undergoes swifter deterioration than any other. This poster displays a methodology that aids in the study of landscape specific language. Documentation and dissemination of collective knowledge alongside use of specialized place-based expressions highlights scholarly-community collaboration in indigenous language revitalization, and outlines my journey on foot with inhabitants of the San Juan Quiahije landscape.

Native language: waC tiC chiqH qaJ ntenB tqaG xiA tyinJ noA tiC jlyoH riqC neG saA skaI naF sqwiJ neqC xqoF. LaC qaE xqanE naF, qoE neC waC ndyiH snaC ndyiE chaqF jnyaJ. NaqG nyiA qyanH chaqF tiC chiqH qaJ ntenB noK tiC jlyoH riqC naF noJ yqwiJ tiC sqneE. LoA ktyiC reC ktsanH chinqH qwanK noK qneI waG jnyaF chaqF tiC xnyiJ tyqiC ntenB noK tiC jlyoH riqC saA skaI chaqF. tyonC noE ngaJ waG noA qneI waG jnyaF qinF ranF, qneJ waG jnyaF chaqF jaA tyiI chaqF jnyaJ.

P.3.11: Resources and digital materials in Mexico’s indigenous languages
Authors: Luis Flores Martínez
Country: Mexico

Abstract: In Mexico, 68 indigenous languages (LI) are spoken which are at risk of disappearance. The use of ICTs can be an ally for LIs by allowing them to increase their visibility, promoting their use, teaching, and learning. A multidisciplinary group of LI speakers created a platform on Facebook (@lenguasweb) intending to raise awareness of the linguistic richness, as well as teach illiterate speakers the written form of their own language. We hope to be a motive, especially for young people, and contribute to the preservation, dissemination, and revitalization of the minority languages of Mexico and the world.

Native language: Tsakam dhuchlab (Tének de San Luis Potosí, México) Ti Labtóm Tsabál ajyamej óx inik laju waxik i Tének kawintaláb, po axe'chik yab exladh, ani wa'ats i atiklabchik axi yabáts in le' kin eyendha'. Jaxtam jún kubél i atikláb axi i kawnál i Tének káwintal u junkun abal ki ts'ejka' jún i xeklek ti Facebook axi in jamat bij játs @lenguasweb, taná' i tejwa'medhál junchik i káw abal pilchik i atikláb kin exla'chik ti waw i kwentaj, po jayej abal i juntal Tének kin exla' jant'ini' tu dhucháb i káwintal; axé' jayej pel abal ki edhanchij in tsaláp pil i juntal ani ki ela' ti ébtsolom jun i jolataláb abal wawá' ani i káwintal.

P.3.12: Ayöök, México
Authors: Marco Martinez
Country: Mexico

Abstract: Kumoontun was developed out of necessity in order to preserve the Ayook language (a variant of Mixe from the Totontepec area) by taking advantage of digital media and platforms. It is a new and different experience because the Kumoontun App was created is shared directly with the communities by means of workshops for children, youth and the community at large. Today we share the experiences, challenges and achievements of this project only months after its creation.

Kumoontun App for IPhone https://cutt.ly/FeOlmPQ

Kumoontun App for Android https://cutt.ly/peOlTZb

Native language: Yi Kumoontun wä’ä yiwe tsyöönt ïik jats yiwe yi ayöök juu’ yak’kojtsp Anyiköjmtsoj yajk’kojtswijït yak’kojtswa’atsït mëët yi tuköjtsïn, tonpäjtïn jats mëët jomajatsy yi ayöök wya’kxtïk. Pi’k’önïkta, waa’tyëjkta, kiixtë’ëxta jats kajpïn jayïta ananyijoma mëët ëëtse ntun. Xyam ëëtse’n’awanat wintsowe ja winma’ayïn myiijn jats yak’ukwaajny, tyëjxyiwe winköpk jats yiwe yi aa’ayöök ntöönkimtat.

Kumoontun App IPhone https://cutt.ly/FeOlmPQ

Kumoontun App Android https://cutt.ly/peOlTZb

P.4.1: African Wordnet – digital documentation and preservation of indigenous knowledge

Authors: Sonja Bosch and Marissa Griesel

Country: South Africa

Abstract: Indigenous knowledge concepts in isiZulu, collected from a variety of sources such as monolingual and bilingual dictionaries can be transformed from alphabetically ordered entries into a hierarchical wordnet structure as a set of relations. The ensuing synsets can further be enriched and lexical gaps filled with information from other cultural resources to transcend the physical limitations of such traditional sources by including definitions, usage examples, pictures and dialect information. The multilingual African Wordnet can be used as important tool in the globalisation of Africa's indigenous knowledge systems, thereby contributing to language empowerment through revitalization of culture and knowledge

Native language: I-African Wordnet - imibhalo eyidijithali kanye nokugcinwa kolwazi lwendabuko Kulokhu okwethulwayo kuzovezwa ukuthi amagama esiZulu awulwazi lwendabuko, aqoqwe eqhamuka emithonjeni enhlobonhlobo enjengezichazimazwi ezinolimi olulodwa nezinezilimi ezimbili, angaguqulwa asuke emagameni afakwe ahleleka ngokwe-alfabhethi, abe yisakhiwo samagama ahlelwe ngokwamazinga ehlukahlukene saba yiqoqo eliveza ubudlelwane obukhona phakathi kwawo. Amaqoqo amagama amqondofana (synsets) atholakala kulokhu, aphinde athuthukiswe kuvalwe namagebe aphathelene namagama asolimini ngemininingwane eqhamuka kwezinye izinsiza zezamasiko ukuze kugwenywe ukungapheleli kwemithombo leyo ejwayelekile ngokuthi kufakwe izincazelo, izibonelo zokusetshenziswa kwawo kanye nemininingwane yolimi olusetshenzisiwe. I-African Wordnet enezilimi eziningi ingasetshenziswa njengethuluzi elibalulekile ekusebenzeni kumazwe ngamazwe kwezinhlelo zase-Afrika zolwazi lwendabuko, ngalokho iphonse itshe esivivaneni ekunikezweni kwezilimi amandla ngokuvuselelwa kwamasiko nolwazi.

P.4.2: Automated Speech Segmentation: Example of an African Language

Authors: Brigitte BIGI

Country: France

Abstract: Speech segmentation is the process of identifying boundaries between speech units in the speech signal and determining where in time they occur. Linguistic resources of the target language should be defined: a lexicon (the words to be recognized), a word dictionary (their pronunciations as a sequence of phonemes), an acoustic model (a stochastic representation of input waveform patterns per phoneme).

SPPAS software tool implements language-and-task-independent algorithms. This multilingual approach was applied to the african langage Naija (Nigerian pidgin). We developped language resources for a tokenizer, an automatic speech system for predicting the pronunciation of the words and their segmentation.

Native language: La segmentation de la parole consiste à identifier les unités dans le signal de parole et à déterminer où celles-ci se produisent dans le temps. Des ressources linguistiques de la langue cible doivent être définies : un lexique (les mots à reconnaître), un dictionnaire de mots (leurs prononciations en tant que séquence de phonèmes), un modèle acoustique (une représentation stochastique par phonème).

L'outil logiciel SPPAS implémente des algorithmes indépendants du langage et des tâches. Cette approche multilingue a été appliquée au langage africain Naija (pidgin nigérien). Nous avons développé des ressources linguistiques pour un tokenizer, un convertisseur graphème-phonèmes et leur alignement avec le signal.

P.4.3: Establishing Sustainable Infrastructures for African Languages

Authors: Z Steyn

Country: South Africa

Abstract: The South African Centre for Digital Language Resources (SADiLaR) is a new research infrastructure (RI) set up by the Department of Science and Innovation (DSI) forming part of the new South African Research Infrastructure Roadmap (SARIR)

The centre runs two main programmes. A digitisation programme and a Digital Humanities (DH) programme. The focus of the poster will show the different elements of establishing the digitisation programme, which focusses on the creation text, audio and multimodal datasets as well as the development of NLP tools and software for the 11 official languages of South Africa.

Native language: Die Suid Afrikaanse Sentrum vir Digitale Taal Hulpbronne (SADiLaR) is 'n nuwe navorsingsinfrastruktuur befonds deur die Departement van Wetenskap en Innovasie en maak deel uit van die Suid Afrikaanse Navorsingsingsinfrastruktuur Padkaart (SARIR)

Die sentrum huisves twee programme. 'n Digitalisering en 'n Digitale Humaniora-program. Die plakkaataanbieding sal fokus op die digitaliseringsprogram, spesifiek op die samestelling daarvan wat behels die skep van teks, klank en ander multimodale datastelle sowel as die ontwikkeling van NLP programme vir die 11 amptelike tale van Suid Africa.

P.4.4: A South African Corpus of Multilingual Code-switched Soap Opera Speech
Authors: Febe De Wet, Ewald Van der westhuizen and Thomas Niesler
Country: South Africa

Abstract: We introduce a speech corpus containing multilingual code-switching compiled from South African soap operas. The corpus contains monolingual as well as code-switched examples of English, isiZulu, isiXhosa, Setswana and Sesotho speech. The last four are indigenous languages, all belonging to the Southern Bantu family. IsiZulu and isiXhosa are Nguni languages that, while distinct, are to some degree mutually intelligible and linguistically similar. The same applies to Setswana and Sesotho, which are Sotho-Tswana languages. The data contains both inter-sentential and intra-sentential code-switching. Intra-sentential code-switching occurs as alternation, insertion as well as intra-word switches.

Native language: Sethula i-corpus yenkulumo equkethe ukushintshwa kwekhodi yezilimi eziningi ehlanganiswe kuma-soap opera waseNingizimu Afrika. I-corpus iqukethe izibonelo zesiNgisi nesiZulu nesiXhosa nesiTswana nesiSuthu ezinolimi olulodwa kanye nezibonelo ezishintshile ikhodi. Ezine zokugcina ziyizilimi zomdabu, zonke zingabomndeni waseSouthern Bantu. IsiZulu nesiXhosa yizilimi zesiNguni, nakuba zihlukile, ngezinga elithile ziyaqondana futhi zifana ngohlelo. Kwenzeka okufanayo nesiTswana futhi nesiSuthu, eziyizilimi zesiSuthu-Tswana. Idatha iqukethe ukushintshwa kwekhodi okungaphandle kwemisho futhi okungaphakathi kwemisho. Ukushintshwa kwekhodi okungaphakathi kwemisho kwenzeka njengokushintshana (alternation), ukufakwa (insertion) kanye nokushintshwa ngaphakathi kwamagama (intra-word switches).

P.4.5: Corpora Mandeica: text corpora for Mande languages (West Africa)
Authors: Valentin Vydrin
Country: France

Abstract: "Corpora Mandeica" is a set of corpora of annotated written texts in languages of the Mande family, openly accessible in the Internet. All the texts in the corpora are provided with POS tags and French (eventually also English and Russian) glosses. The corpora are partly disambiguated; parallel subcorpora are also being developed. So far, there are corpora for four languages available on line: Bambara (more than 11 million words), Guinean Maninka (about 3,5 million words), Eastern Dan (about 460,000 words), Mwan (47,000 words). The corpora are accompanied by electronic dictionaries and electronic libraries. Further language corpora are planned.

Native language: Проект Corpora Mandeica представляет собой совокупность аннотированных корпусов письменных текстов на языках манде, находящихся в открытом доступе в Интернете. Все тексты аннотированы (снабжены частеречными пометами и французскими глоссами; отчасти также английскими и русскими). Для части текстов проведено снятие омонимии. Создаются также параллельные корпуса. К настоящему моменту доступны для поиска корпуса 4 языков: бамана (11 млн. слов), гвинейский манинка (около 3,5 млн.), восточный дан (около 460 тыс. слов), муан (47 тыс. слов). На корпусных сайтах вывешены электронные словари; имеются также электронные библиотеке для 3 языков. Планируется создание корпусов и для других языков семьи.

P.4.6: Missing link: A centralised digital archive for endangered languages of southern Africa
Authors: Kerry Jones
Country: South Africa

Abstract: Language endangerment and language loss is a worldwide phenomenon. As a result, the scramble to identify, document and preserve indigenous languages using digital technology has gained traction. The challenge we face in southern Africa, is the lack of a centralised digital archive for endangered languages. Currently, efforts are dispersed on various platforms, hosted by universities, non-government organisations or private collections, if digitised at all. In order to provide a holistic description of endangered and extinct languages in southern Africa, an online digital archive could centralise existing efforts, while creating opportunities for the digitisation of historical records and new digitised entries.

Native language: Gowaga ǁō‡oas !aorosasib tsî gowaga ǃaris tsîra ge !hūbaib ǂhabase, harase a ǂansa !nae!khaira. ǁNā-amaga di ge ǃoro ǁan‡gāsaben gowaga nēsi hâ texnoloxib ǂkha ôa‡ui, xoamâi tsi ǁkhaubas di ǁgūbade nēsi ǃgaisase ra ǂoaxa. Afrikab !khawagas !nâ da ra hō!â nausa ǁgoa‡uis ge ǃgui!nâxa digitel‡khanisâuǁgâugu !nuwusiba, ǁō‡oas !aorosasib !nâ mâ gowagu !aroma. Nē ǁaeb ai di ge ǃguilguibe dītsârode !kharaga!nâgu !harodi ai ra hōhe, universiteitdi tawa i ka hâ tama kara io, o ǂhanub !auga hâ ǂnûi‡gādi tawa, tamas ka io, ǃguilguibe khoen tawa - ǁnās ge hâna i ka digitelǃgaub !nâ a hōhe ǁkhā osa. Hoa !hariga !khō‡gā hâ ǃgaub !nâ da ka ǁō‡oas !aorosasib !nâ mâ gowagu tsî ǃnai ge ǁō‡oa gowagu Afrikab !khawagas digu tsîna a xoamâi ǂgao, o i ge kaise nî !gâi online digitel‡khanisâuǁgâuba kurusa. Nēti ī digitel‡khanisâuǁgâub ge ǃnai hâ sîsengu hoaga ǃhaoǃhao tsî ǃgui !khais tawa ǃgui nî !khōǃgara, tsî nēs !nâ-u ǃgūǃgarus !nae!khaidi xoaǃguigu tsî ka hâ ǃasa xoadi tsîna digitaǃgaub !nâ sâus di sîsen-i di daode nî ǁkhowa-am.

P.4.7: Using Citizen Linguistics to Empower Indigenous Communities
Authors: Christopher Cieri and Mark Liberman
Country: United States

Abstract: While Language Technologies promote digital linguistic diversity and community development through access to knowledge, such technologies rely upon datasets absent in most indigenous languages. The LanguageARC Citizen Linguist portal augments traditional sources of language data by empowering indigenous communities to create their own. Via brief, engaging tasks such as picture and video description, vocabulary elicitation and usage surveys that can be completed on a computer or smart phone, indigenous communities collect spoken or written data as appropriate, and augment it by transcribing, translating, judging grammaticality and annotating for use in technology development and language development.

P.4.8: Heuristic guided probabilistic graphic language modelling for morphological segmentation of isiXhosa

Authors: Lulamile Mzamo, Albert Helberg and Sonja Bosch

Country: South Africa

Abstract: The IsiXhosa Heuristics Maximum Likelihood Segmenter (XHMLS), an unsupervised isiXhosa segmenter, is evaluated. The study contributes use of isiXhosa word morphology heuristics as a guide to probabilistic graphical modelling (PGM) the segmentation of isiXhosa. Four guided PGMs with options for modified Kneser-Ney (mKN) smoothing are presented. XHMLS's boundary identification accuracy of 78.7% outperforms the benchmark Morfessor-Baseline's 77.2%, and shows an even better f1-Score, 68.0%, compared to Morfessor-Baseline's 48.9%, when modelled with circumfixing and smoothing. The study shows that better word segmentation performance could be achieved in the unsupervised morphological segmentation of isiXhosa if a representative and smoothed PGM is used.

Native language: I-IsiXhosa Heuristics Maximum Likelihood Segmenter (XHMLS), isicaluli-mbhalo sesiXhosa esingagadwanga, siyavavanywa. Igalelo loluphando kukusetyenziswa kwendlela amagama esiXhosa aguquka ngayo njengesikhokelo somFanekiso-mBoniso-Thuba (FBT) ekucaluleni isiXhosa. Ii-FBT ezikhokelweyo ezine, ezinokukhetha ukusebenzisa ugudiso lwe-Kneser-Ney elungisiweyo (mKN), ziyaboniswa. Inkcaneko yokukhomba imida yezimilo ye-XHMLS eyi-78.8% igqitha eyomgangatho-jikelele oyiMorfessor-Baseline, we-77.2%, kwaye igqithe ngakumbi ngenqaku le-f1, ngo-68.0%, xa ithelekiswa neyeMorfessor-Baseline engu-48.9%, xa inkokhelo izizimi-macala yaye igudiswe nge-mKN. Olu phononongo lubonisa ukuba ucalulo-magama lwesiXhosa olungcono lungafumaneka xa kusetyenziwa isicaluli-magama esingagadwanga se-FBT esisifuziseleyo isiXhosa sibe sigudiswe nge-mKN.

P.4.9: Radio-browsing in support of relief and development work in rural Africa

Authors: Astik Biswas, Febe De Wet, Herman Kamper, Raghav Menon, Thomas Niesler, Armin Saeb, John Quinn, Ewald Van der westhuizen and Emre Yilmaz

Country: South Africa

Abstract: In countries with well-established internet infrastructure, social media has become an accepted platform for voicing opinions. However, in some parts of Africa internet infrastructure is poorly developed, precluding the use of social media to gauge sentiment. Instead, community radio phone-in talk shows are used to voice views and concerns. Our contribution will introduce a radio browsing system that is intended to support relief and developmental programmes by the United Nations (UN). Browsing systems were developed to monitor community radio broadcasts for keywords related to specific topics such as natural disasters, disease outbreaks, or other crises.

Native language: Wadamadda leh kaabayaasha internetka ee sida wanaagsan loo aasaasey, warbaahinta bulshada waxay noqotey meel lagu aqbalo fikradaha codadka la dhiibto. Si kastaba ha ahaatee, qeyb ka mida kaabayaasha internetka ee Afrika ayaa si liidata loo horumariyey, iyadoo la sii saadaalinaayo adeegsiga warbaahinta bulshada si loo qiyaaso dareenka. Taa badal keed waa bandhigyada wada hadalka telifoonka telefishanka bulshada ayaa loo isticmaalaa in lagu dhawaaqo aragtida iyo shirarka. Waxqabadkeena ayaa soo bandhigi doona, nidaam raadiya raadyaha oo loogu tala galey in lagu taageero gargaarka iyo barnaamijyada horumarineed ee ay bixiso Qaramada Midoobey (UN). Nidaam baadhitaan ayaa loo sameeyey si loola socdo raadyaha idaacadda bulshadaee ereyada muhiimka ah ee la xidhiidha mowduucyada gaarka ah, sida masiibooyinka dabiiciga ah, cudurada faafa ama dhibaatooyin kale.

P.4.10: Analysis of Language Relatedness for the Development of Multilingual Automatic Speech Recognition for Ethiopian Languages

Authors: Martha Yifiru Tachbelie, Solomon Teferra Abate and Tanja Schultz

Country: Ethiopia

Abstract: In this poster, we present the analysis of GlobalPhone (GP) and speech corpora of Ethiopian languages (Amharic, Tigrigna, Oromo and Wolaytta). The aim is to select speech data from GP for the development of multilingual Automatic Speech Recognition (ASR) system for the Ethiopian languages. To this end, the phonetic overlaps among GP and Ethiopian languages have been analyzed. Moreover, morphological complexity of the GP and Ethiopian languages, reflected with high out of vocabulary rate and type to token ration, has been analyzed using training transcriptions. We also present baseline ASR performances for each of the GP and four Ethiopian languages.

Native language: በዚህ ፖስተር የምናቀርብላችሁ በግሎባል ፎን እና በአራት የኢትዮጵያ ቋንቋዎች (አማርኛ፣ ትግርኛ፣ ኦሮምኛ እና ወላይትኛ) የድምፅ ዳታ መካከል ያደረግነውን የማነጻጸር ጥናት ነው። የጥናቱ ዋና አላማ ከግሎባል ፎን የድምፅ ዳታ ውስጥ ለብዙ የኢትዮጵያ ቋንቋዎች ንግግርን ወደ ድምፅ የሚቀይር መተግበሪያ ለመስራት ጠቃሚ የሆነ የድምፅ ዳታ መምረጥ ነው። ስለዚህም በግሎባል ፎን እና በአራቱ የኢትዮጵያ ቋንቋዎች መካከል ያለውን የድምፅ መመሳሰል አጥንተናል። በተጨማሪም የቋንቋዎቹን ምእላዳዊ ውስብስብነት ለመረዳት እንዲቻል "Out of Vocabulary" እና "Type to Token Ratio" በማስላት ለማየት ሞክረናል። ለእያንዳንዱ የግሎባል ፎን እና የኢትዮጵያ ቋንቋዎች የተዘጋጁ ንግግርን ወደ ድምፅ የሚቀይሩ መተግበሪያዎችንና አፈፃፀማቸውን አሳይተናል።

P.4.11: Automatic Learning of a Phonological System: a Case Study on the Mboshi Language

Authors: Lucas Ondel and Lukas Burget

Country: Czech Republic

Abstract: Over the last decade, a lot of research focused on automatically learning basic acoustic units, a.k.a "pseudo-phones", for low-resource languages. In this work, we investigate the potential and the limits of this research on a real case scenario for documenting a low-resource language. We performed our experiments on Mboshi, an African language from the Bantu family. Results show that despite some progress, automatic learning from under-resourced languages remain a very challenging task and requires further research.

Native language: Au cours des dernières années, de nombreux travaux de recherche se sont concentrés sur l'apprentissage d'unités acoustiques, appelées "pseudo-phones", pour les langues peu dotées. Ce travail explore le potentiel et les limites de cette recherche dans un scénario réaliste de documentation d'une langue peu dotée. Nous avons mené nos expériences sur le Mbochi, une langue Africaine de la famille Bantoue. Les résultats montrent que, en dépit de progrès indéniables, l'apprentissage automatique à partir d'une langue peu dotée reste une tâche difficile et nécessite de plus ample recherches.

P.4.12: Current Status, Issues, and Future Directions for Ethiopian Natural Language Processing (NLP) Research

Authors: Seid Yimam and Chris Biemann

Country: Germany

Abstract: These days, the generation of resources (mainly text and speech) for many languages is dramatically increasing. However, high-resource languages such as English and low-resource languages such as Amharic, greatly differ on the amount of NLP components, tools and applications. In this poster, we will briefly discuss the state-of-the-art NLP research for Ethiopian languages. Then, the main bottlenecks that hinder the development of the required resources will be reviewed. Finally, we will point out best practices to solve current issues and indicate appropriate tools and models that can be easily adapted for low-resource NLP research, particularly for Ethiopian languages.

Native language: በአሁኑ ግዜ፣ ለብዙ ቋንቋዎች መረጃዎችን (በዋነኛነት የጽሑፍ እና የንግግር) ማግኘት እጅግ በጣም ቀላል እየሆነ እየሆነ መጥቷል። ሆኖም ግን እንደ እንግሊዝኛ በብዛት መተግበሪያ ያላቸውና እንደ አማርኛ ያሉ እጅግ በጣም ዝቅተኛ መተግበሪያ ያላቸው ቋንቋዎች፣ በተፈጥሯዊ የቋንቋ ቴክኖሎጂ (ተ.ቋ.ቴ - NLP)ግብዓቶች፤ መሳሪያዎች እና መተግበሪያዎች መጠን ሰፊ ልዩነት አላቸው። በዚህ ፖስተር ጽሁፍ፤ በመጀመሪያ የኢትዮያ ቋንቋዎች በተ.ቋ.ቴ ምርምር አሁን ያሉበት ደረጃ በጥልቀት ይብራራሉ። በመቀጠል ለኢዮጵያ ተ.ቋ.ቴ ምርምር እደገት እንቅፋት የሆኑ ዋና ዋና ክፍተቶች ይገመገማሉ። በመጨረሻም ፣ ወቅታዊ እና ተያያዥ ተግዳሮቶችን እንደት መፍታት እንደሚቻል፤ አሁን ላይ የሚገኙ የሌሎች የበለፀጉ የቋንቋ መተግበሪያዎችን እና ሞደሎችን የመተግበሪያ እጥረት ላለባቸው ቋንቋቶች(በተለይም ለኢትዮጵያን ቋንቋዎች) እንደት ማላመድና መጠቀም እንደሚቻል ይጠቁማል።

P.4.13: ACALAN: Platform for African Language Empowerment (PALE)

Authors: Martin Benjamin

Country: Switzerland

Abstract: The African Academy of Languages (ACALAN) is finalizing a proposal for a comprehensive platform for African languages in Cyberspace. The platform will serve four functions:

1. Information about ACALAN and African language policies and commissions
2. Information about African language research and characteristics
3. A hub for growth and dissemination of African linguistic data
4. A communications center for research and development on African languages

The goal is to produce an ever-growing central resource for scholars, policy-makers, students, and the public to learn about, contribute to, and benefit from knowledge regarding all African languages.

P.4.14: SCAnnAL – An Automatic Speech Corpus Annotator for African Speech Corpora

Authors: Moses Ekpenyong, Eno-Abasi Urua and Aniefon Akpan

Country: Nigeria

Abstract: Today, thousands of annotated speech corpora exist worldwide and demand for richly annotated corpora is fast growing, but the process accompanying the segmentation and labeling of corpora has slowed research progress for African languages due to the limitations of current annotation Toolkits to satisfy the challenges African speech systems present. We introduce SCAnnAL, a Toolkit for automatic speech annotation that automates the annotation process by accepting raw audio files, segments the waveforms and finally dumps labels into created segments. SCAnnAL is currently being refined for accuracy and is certain to put an end to the intractable procedure of speech annotation.

Native language: Mfịn ami, mme tọsịn adianañkpadia ikọ eba ke ekondo, ñko enekke eyem mme adianañkpadia ikọ ntom ke ise ikọ eti eti. Daña esaña esep enyVñ ewet anyịñ, anam nduuñọ ke mme usem Afrika anyọnyọọñ sia se ekama enam utom ado anana akeene ñkpọ ñnyan ubọk aabañake mme usem Afrika. Imiben SCAnnAL, anamidem akebe ñkpọ usep ñnyVñ ñwet anyịñ ikọ ndọñ ke ise ikọ iwọt. SCAnnAL akeme adidat utatañ ikọ, asep, anyVñ awet anyịñ ke mbaak ikọ. EsVk enanam utom ke SCAnnAL ma akan anọ nneke ibọọrọ anyVñ ayọ mfịna aasañake ke adisep ikọ ndon ke ikpeghe.

P.4.15: NTeALan - Artificial Intelligence, Development and Promotion of African National Languages

Authors: Elvis Mboning and Damien Nouvel

Country: France

Abstract: Among the emergent challenges that the African continent is currently facing is the problem of safeguarding and enhancing its cultural and linguistic heritage. Created in 2017, the NTeALan association has been working since then to implement intelligent technological tools for the digitization, promotion, development and teaching of African national languages. NTeALan wants to make these languages the pillars of social and technological development in Africa.

Native language: Í kété mìtíík mí mám má ńsòmblà íɲǔ hólós áfríkà í mà ŋgɛ́ŋ máná dì gwě, màhɔ̀l má má ntágbɛ́nɛ́ í máhɔ́p més nì bìɓoŋól gwéṣ, íɲǔ hàlà ɲɛ́n Ntealaṇ (ìɓòdòl 2017), tòhálá kìí à má sál ŋgándàk mú í ndzɛ́l ī, à ŋgí sálák nì láná lɛ́ɛ: à ńtí bínóŋól bí mɔ́ndɔ́ bí bí ńhólá lɛ̀è dí nìígá, nì hólɔ̃s màhɔ́p més lòŋnì ndzɛ́l ì bìɓòŋól bí mɔ̀ndɔ́. hálà à gáhólá í nììgà nì hólós màhɔ́p més . NTeALan à ńsòmból ɓɔ̀ŋ lɛ̀è màhɔ́p má áfríkà má ɓá ŋ́ŋɛ́m ú máhɔ̀l má bílɔ̀n gwés gwɔ́ bísóná.

P.4.16: Grappling with Opportunities and Challenges in developing Language Technologies for Under-resourced African Languages

Authors: Sunday Ojo

Country: South Africa

Abstract: Multilingualism and Diversity is the hallmark of African societal heritage, which can be exploited to foster unity in diversity, towards sustainable development. But challenges and opportunities abound in exploiting Language Technology (LT) potentials. Their under-resourcedness and rich contextual semantics features constitute a challenge, given that most of the LTs are validated using richly-resourced languages, such as English as test beds. There are also contextual issues bordering on paucity of internal capacity for exploiting the LT potentials. The poster draws from an on-going experience in working on LTs for some West and Southern African under-resourced languages, eliciting these challenges and opportunities.

Native language: Orisirisi ede ati Oniruuru jẹ aami pataki fun ohun-ini awujọpọ ti Ilu Afirika, eyiti a le lo lati ṣe isọdi iṣọkan ni oniruuru, si idagbasoke alagbero. Ṣugbọn awọn italaya ati awọn anfani pọ si ni ilokulo awọn agbara Imọ-ẹrọ ede. Wọn labẹ abinibi ati awọn ẹya Semantics ọrọ ọrọ ọlọrọ jẹ ipenija, fun pe ọpọlọpọ awọn Imọ-ẹrọ ede ni a fọwọsi pẹlu lilo awọn ede ti o lọpọlọpọ, bii Gẹẹsi bii awọn akete idanwo. Awọn ọran ti o tọ tun wa ni ilaja lori agbara ti agbara inu fun ilo awọn agbara Imọ-ẹrọ ede. Atẹjade naa fa lati iriri lilọ kiri ni sisẹ lori Imọ-ẹrọ ede fun diẹ ninu awọn ede ti Iwọ-oorun ati Gusu Afirika labẹ itusilẹ, ṣiṣe awọn italaya ati awọn aye wọnyi.

P.4.17: Towards the First Low-Resource French-Somali Machine Translation System

Authors: Houssein Ahmed Assowe and Fatiha Sadat

Country: Canada

Abstract: As part of an ongoing research on NLP for under-resourced languages, we have created several linguistic tools such as a French-Somali machine translation system dedicated to the journalistic domain, allowing to obtain quality translations from a bilingual corpus built by post-editing GOOGLE TRANSLATE results. The final users being the Somali and non-French speaking populations of the Horn of Africa. Furthermore, an aligned parallel corpus was constructed and evaluations were conducted by nine bilingual annotators who assigned a quality score to each segment of the corpus and corrected the post-editing results. We are following up with a neural French-Somali MT system.

Native language: Iyada oo qayb ka ah cilmibaadhista hadda jirta ee NLP-da oo loogu talagalay luqadaha aan aad uu kombuyutargaraynsanayn, waxaan abuurnay dhowr aalado luqadeed oo kala duwan sida nidaamka turjumaad mashiinka ee afka faransiiska iyo soomaliga kaas oo loo qoondeeyay aagga warbaahinta, isago oo oggolaanayno in la helo tarjumaad tayo leh oo laga sameeyay xogta laba-luqadlaha ah oo aan ka dhisnay dib-u-habeynta natiijooyinki turjumaadi uu horeysay ee GOOGLE TRANSLATE. Adeegsadayaasha aaladan ayaa ah dadka Soomaalida ee aan ku hadlin afka Faransiiska ee Geeska Afrika. Intaa waxaa sii dheer, xog laba-luqadle oo isbarbar socda ayaa la dhisay , iyo qiimeynna waxa ku sameeyay xogta sagaal tilmaameyaal laba luuqadoodla ah kuwaas oo siiyay dhibco tayo leh weer kasta oo ka mid ah xogka isla markaana saxay natiijooyinka dib-u-habeynta. Waxaan kale oo sameynay nidaamka turjumaad ee afka Faransiiska iyo Soomaaliga oo ku dhisan shabakadaha neerfaha ee aan dabiici ahayn .

P.4.18: Automatic Dialect Detection in Arabic Broadcast Media

Authors: Bilal Belainine and Fatiha Sadat

Country: Canada

Abstract: We describe our participation at the Arabic Multi-Genre Broadcast (MGB) Challenge for ASRU 2017. In this shared task, we used several supervised learning methods that emphasize only on labelled data, involving Voting Ensemble, Multi-Layer perceptron and Logistic classifiers, and their combinations, to discriminate and distinguish between four major Arabic dialects: Egyptian, Levantine, North African, Gulf and Modern Standard Arabic (MSA). Our best results were obtained during this shared task on closed submission using the Voting Ensemble with an overall accuracy of 56.10, followed by the simple logistic and Multi-Layer perceptron with an overall accuracy of 54.56 and 53.55, respectively.

Native language: وصفنا مشاركتنا في تحدي البث متعدد الأنواع العربية (MGB) لـ ASRU 2017. في هذه المهمة المشتركة ، استخدمنا العديد من أساليب التعلم التي تخضع للإشراف والتي تركز فقط على البيانات المصنفة ، بما في ذلك مجموعة التصويت ، المصنفات متعددة الطبقات والمصنفات اللوجستية ، ومجموعاتها ، للتمييز والتمييز بين أربع لهجات عربية رئيسية: مصر ، بلاد الشام ، شمال إفريقيا ، الخليج والعربية الحديثة المعاصرة (MSA). تم الحصول على أفضل النتائج خلال هذه المهمة المشتركة عند التقديم المغلق باستخدام فريق التصويت مع دقة إجمالية قدرها 56.10 ، تليها الإدراك البسيط لوجستي ومتعدد الطبقات مع الشاملة دقة 54.56 و 53.55 ، على التوالي.

P.4.19: The ADLaM Story

Authors: Abdoulaye Barry and Ibrahima Barry

Country: United States

Abstract: Thirty years ago, my brother and I created ADLaM to accurately transcribe all the sounds of the Fulfulde language spoken by more than 60 million people in Africa. The alphabet contains 28 letters and it was encoded in 2016 in Unicode 9.0. Since then our goal has been to try to get ADLaM working on every smartphone and computer. ADLaM is currently supported by Google through Android 8 and Chrome, and also by Microsoft through Windows 10. The script is now used from Gambia to Sudan by people who are very excited to use their language in their everyday life.

Native language: [illegible]

P.5.1: Linguistic Linked Open Data for All
Authors: John McCrae and Thierry Declerck
Country: Germany

Abstract: In this poster we show how to increase the uptake of language technologies for all by exploiting the combination of linked data and language technologies, that is Linguistic Linked Open Data (LLOD), to create ready-to-use multilingual data, also for low-resourced languages. The project Prêt-à-LLOD develops tools for the transformation and linking of datasets and apply these to both data and metadata in order to provide multi-portal access to heterogeneous data repositories. Prêt-à-LLOD implements a new methodology for building data value chains applicable to all types of language resources and language technologies that can be integrated by means of semantic technologies.

Native language: Nous montrons comment accroître l'adoption des technologies langagières pour tous en exploitant la combinaison de données liées (Linked Data) et de technologies langagières, c'est-à-dire l'infrastructure du Linguistic Linked Open Data (LLOD), pour créer des données multilingues prêtes à l'usage même pour les langues à faibles ressources. Le projet Prêt-à-LLOD développe des outils de transformation et de mise en relation des ensembles de données linguistiques et les applique à la fois aux données et aux métadonnées afin de fournir un accès multi-portail à des référentiels de données langagières hétérogènes. Prêt-à-LLOD met en œuvre une nouvelle méthodologie de construction de chaînes de valeur de données applicables à tous les types de ressources et de technologies langagières pouvant être intégrées par le biais des technologies sémantiques.

P.5.2: Bangla Text and Spoken Language Technology
Authors: Professor Dr. Mohammad Nurul Huda
Country: Bangladesh

Abstract: The poster will primarily focus on the following technologies developed by my team in NLP field. In addition, the major problems and obstacles faced in the development process will be addressed and analyzed. Some Developed NLP Tools are: Spell Checker, Question Answering System, Bangla TTS, Bangla to IPA Converter, Machine Translator, Bangla Sentiment Analyzer, Bangla-English Parallel Corpus. Moreover, in near future, we will develop Sign Language Tool and Screen Reader Tool for vision impaired people.

Native language: পোস্টারটি প্রাথমিকভাবে এনএলপি ক্ষেত্রে আমার দল দ্বারা নির্মিত নিম্নলিখিত প্রযুক্তিগুলিতে ফোকাস করবে। এছাড়াও, তৈরির প্রক্রিয়ায় যে বড় সমস্যা ও বাধার মুখোমুখি হয়েছিল সেগুলি সমাধান করে বিশ্লেষণ করা হবে। কয়েকটি তৈরিকৃত এনএলপি টুলস হল: বানান পরীক্ষক, প্রশ্ন উত্তর সিস্টেম, বাংলা টিটিএস, বাংলা থেকে আইপিএ রূপান্তরকারী, মেশিন অনুবাদক, বাংলা সেন্টিমেন্ট বিশ্লেষক, বাংলা-ইংলিশ প্যারালাল কর্পাস। তদুপরি, অদূর ভবিষ্যতে, আমরা দৃষ্টি প্রতিবন্ধী ব্যক্তিদের জন্য সাইন ল্যাঙ্গুয়েজ টুল এবং স্ক্রিন রিডার সরঞ্জাম তৈরি করব।

P.5.3: Envisioning a Trilingual Machine Translation System for the Language Pairs –<Tamang –English –Nepali>
Authors: Bal Krishna Bal, Amrit Yonjan Tamang and Lasang Jimba Tamang
Country: Nepal

Abstract: A Machine Translation system is a system that provides a gist or tentative translation in any target language for any given text in the source language. In the context of the degrading number of Tamang speakers among the younger generations of the Tamang community and in other contexts where the knowledge of English and/or Nepali may not necessarily be good among Tamang speakers, technology like MT system can help establish a strong stake or identity from a language empowerment or enabling perspective.

Native language: कुनै पनि यान्त्रिक अनुवाद प्रणालीले श्रोत भाषामा रहेको पाठलाई लक्षित भाषामा ठोस वा भावानुवाद गर्दछ। यस्तो प्रणाली दुई वा सोभन्दा बढी भाषी समुदायका बिच भाषिक तथा डिजिटल खाल्डो (समस्या) निवारण गर्न ठूलो भूमिका खेल्दछ। तामाङ समुदायको युवा पिंढीमाँझ आफ्नो तामाङ भाषा सिक्ने सन्दर्भमा देखिएको उदासिनताका साथसाथै कतिपय सन्दर्भमा अङ्ग्रेजी र नेपाली भाषाको ज्ञान उनीहरूमा कम भएको परिस्थितिलाई मध्यनजर राख्ने हो भने यान्त्रिक अनुवाद जस्तो प्रणालीले भाषाको पहिचान स्थापित गर्नका साथै भाषाको प्रबर्धनमा महत प्रुयाउँन सक्छ। यस पोस्टरमा यस्तो त्रिभाषीय यान्त्रिक प्रणाली <तामाङ – अङ्ग्रेजी – नेपाली> को निर्माण सम्बन्धमा उच्च स्तरीय योजना प्रस्तुत गरिएको छ। उक्त योजनाअन्तर्गत भाषिक सामग्रीहरूको साथै न्युरल यान्त्रिक प्रणाली निर्माणबारे विस्तृत जानकारी समावेश छ। यस्तो प्रणालीको सफल विकासपश्चात समाजका विभिन्न तह तप्कामा प्रयोगमा ल्याई स्थानीय भाषा, संस्कृति र परम्पराको प्रबर्धन र विकास गर्न सकिने सम्भावना बोक्छ।

P.5.4: Keyman: High Fidelity Text Input for All Languages
Authors: Marc Durdin, Sok Makara, Joshua Horton and Ty Rasmey
Country: Australia

Abstract: While there are established conventions for typing Khmer using the Unicode Standard, existing systems provide little assistance to users in following the conventions which are thus often ignored. When typing Khmer text, users find that words can be constructed in multiple ways, all of which look 'correct' on-screen. Furthermore, some aspects of Khmer as implemented by common operating systems deviate from the Unicode Standard. This leads to a number of negative outcomes, including phishing and spoofing security risks, poor searchability and complications with natural language processing. Keyman provides a solution for these problems for Khmer and other languages.

Native language: ទោះបីជាមានគោលការណ៍សម្រាប់ការវាយអក្សរខ្មែរដោយប្រើ “យូនីកូដស្តង់ដារ” ក៏ដោយ ក៏ជំនួយដែលបានផ្តល់ឱ្យអ្នកប្រើប្រាស់ក្នុងការអនុវត្តតាមគោលការណ៍ទាំងនោះ នៅមានតិចតួចនៅឡើយ។ នេះជាហេតុនាំឱ្យអ្នកប្រើប្រាស់មិនសូវចាប់អារម្មណ៍អើពើ។ ពេលវាយអត្ថបទជាភាសាខ្មែរ អ្នកប្រើប្រាស់យល់ថាគេអាចវាយពាក្យបានច្រើនរបៀប ហើយវាមើលទៅត្រឹមត្រូវដូចគ្នានៅលើអេក្រង់។ ជាងនេះទៅទៀត ផ្នែកខ្លះនៃភាសាខ្មែរដែលត្រូវបានអនុវត្តដោយប្រព័ន្ធប្រតិបត្តិការពេញនិយម មានលក្ខណៈល្អៀងពីការកំណត់របស់ “យូនីកូដស្តង់ដារ”។ បញ្ហានេះនាំឱ្យមានលទ្ធផលអវិជ្ជមានជាច្រើន ដូចជា៖ ហានិភ័យផ្នែកសុវត្ថិភាពដោយសារការឆបោកនិងភ្លែងបន្លំ លទ្ធភាពក្នុងការស្វែងរកមានកម្រិតទាប និង ភាពស្មុគស្មាញក្នុងដំណើរការភាសាបែបធម្មជាតិ។ ការស្រាវជ្រាវនេះរកឱ្យឃើញនូវបញ្ហានានាក្នុងការអ៊ិនកូដនៃភាសាខ្មែរដោយ “យូនីកូដស្តង់ដារ”។ ឃីមែន (Keyman) មានដំណោះស្រាយចំពោះបញ្ហាទាំងនេះ សម្រាប់ភាសាខ្មែរនិងភាសាផ្សេងទៀតផងដែរ។

P.5.5: Digital archiving and museum for language documentation and revitalization in Japan
Authors: Natsuko Nakagawa, Masahiro Yamada, Kenan Celik, Nobuko Kibe and Yukinori Takubo
Country: Japan

Abstract: There are eight (UNESCO), twelve (Ethnologue), or more (intelligibility) endangered languages/dialects in Japan. We present a database and digital archiving space that NINJAL (National Institute for Japanese Language and Linguistics) is developing for all of these languages where individual researchers or language communities can deposit their field data, language documentation, or audio-visual recordings. Two major features of the database/archiving space include (i) that it is a Japanese-mediated database/archive and thus virtually everyone in Japan can use it, and (ii) that it comes with an online exhibition space so that archiving is tightly connected to public use of the deposited items.

Native language: 日本には8（ユネスコ）、12（エスノローグ）、もしくはそれ以上（相互理解性）の消滅の危機に瀕した言語・方言が存在する。本発表は国立国語研究所がこれらの言語・方言のために開発中の、個別の研究者や言語コミュニティが利用可能なデータベースおよび電子的アーカイブスペースについて報告する。データベース・アーカイブスペースは以下の二つの特徴を持つ。(i) 日本語によるデータベース・アーカイブスペースであり、日本に住む誰もが利用可能である。(ii) オンライン展示スペースが付随し、アーカイブされるデータが社会一般に対する公開と密接に結びついている。

P.5.6: Project Mélange: Speech and Language Technologies for Code-switching
Authors: Sunayana Sitaram, Monojit Choudhury and Kalika Bali
Country: India

Abstract: Code-switching is the use of two or more languages in the same utterance or conversation, and is common in multilingual communities across the world. Project Mélange aims to process, understand and generate code-switched speech and text, so that technologies that interact with multilinguals can be natural and effective. In this poster, we present an overview of our research in the following areas 1. Data collection and generation 2. Core NLP and speech technologies (Language ID, Part of Speech tagging, Language Modeling, Speech Recognition and Synthesis) 3. sociolinguistics and pragmatics using Twitter data 4. user studies on dialogue and discourse

Native language: एक ही वार्तालाप में दो या दो से अधिक भाषाओं के उपयोग को कोड-स्विचिंग कहा जाता है, जो कि दुनिया भर के बहुभाषी समुदायों में आम है। प्रोजेक्ट मिलांज का उद्देश्य कोड-स्विच किए गए भाषण और पाठ को संसाधित करना, समझना और रचना करना है, ताकि बहुभाषियों के साथ बातचीत करने वाली प्रौद्योगिकियां प्राकृतिक और प्रभावी हो सकें। इस पोस्टर में हम निम्नलिखित क्षेत्रों में अपने शोध का एक अवलोकन प्रस्तुत करते हैं: 1. डेटा संग्रह और संश्लेषण 2. मौलिक भाषण प्रौद्योगिकियां (भाषा निर्धारण, शब्द के भेद निर्णय, भाषा मॉडलिंग, भाषण प्रतिलेखन और संश्लेषण) 3. सामाजिक और व्यावहारिक भाषाविज्ञान 4. संवाद और संभाषण पर उपयोगकर्ता अध्ययन।

P.5.7: Providing smart, open fonts for the world's language communities

Authors: Martin Raymond and Peter Martin

Country: United Kingdom

Abstract: SIL's Language Technology team has designed over thirty families of fonts, covering twenty different scripts, many of them Asian scripts, and supporting thousands of languages. All our fonts are released under the SIL Open Font License, enabling them to be distributed and modified freely. They also use font technologies, Graphite and OpenType, to handle complex writing systems.

The poster will display samples of several fonts, illustrating where in the world they are used, and listing some of the languages they support. Our Andika font was specifically designed for literacy use, and some of its graphic features will be illustrated.

Native language: L'équipe Language Technology de SIL a conçu plus de trente fontes (polices de caractères) dans vingt alphabets différents (beaucoup sont d'origine asiatique) pour assurer une bonne prise en charge dans plusieurs milliers de langues. Toutes nos fontes sont publiées sous « SIL Open Font License », ce qui permet à chacun de les diffuser et de les modifier librement. Elles tirent parti des technologies « Graphite » et « OpenType » pour gérer correctement plusieurs systèmes d'écriture complexes.

Le poster montrera une sélection de ces fontes, illustrera où elles sont actuellement utilisées à travers le monde et dans quelles langues on peut écrire grâce à elles. Andika a été tout spécialement conçue pour l'alphabétisation et certaines de ses caractéristiques graphiques seront aussi détaillées.

P.5.8: InaNLP: Indonesian Natural Language Processing Tools API

Authors: Ayu Purwarianti, Dessi Puji Lestari and Teguh Eko Budiarto

Country: Indonesia

Abstract: We've developed InaNLP, an Indonesian Natural Language Processing Tools API, which consists of several NLP tools that are easily integrated into a text processing module. InaNLP consists of lexical, syntactical and text classification modules, such as POS Tagger, named entity tagger, dependency parser, constituent parser, word normalizer, quotation extraction, document level and concept level sentiment analysis, and topic classification. These modules were built using deep learning algorithms with our own annotated data. The data annotation process was conducted by Indonesian linguists. In this poster, we will show the performance score of several InaNLP modules.

Native language: Kami mengembangkan InaNLP, API Kakas Pemrosesan Bahasa Alami Indonesia, yang terdiri dari beberapa kakas NLP yang mudah diintegrasikan ke dalam modul pemrosesan teks. InaNLP terdiri dari modul klasifikasi leksikal, sintaksis dan teks, seperti POS Tagger, entitas nama tagger , parser dependensi, parser konstituen, normalisasi kata, ekstraksi kutipan, analisis sentimen level dokumen, analisis sentimen level konsep, dan klasifikasi topik. Modul-modul ini dibangun menggunakan algoritma pembelajaran yang mendalam (deep learning) dengan data yang dianotasi sendiri. Proses anotasi data dilakukan oleh ahli bahasa Indonesia dan terdiri dari beberapa langkah seperti persiapan pedoman anotasi, pelabelan data dan pengecekan kualitas. Dalam poster ini, kami akan menunjukkan skor kinerja beberapa modul InaNLP.

P.5.9: The Pangloss Collection: an open archive of under-documented languages designed with Natural Language Processing in view

Authors: Séverine Guillaume, Balthazar Do Nascimento and Alexis MICHAUD

Country: France

Abstract: The Pangloss Collection was created by the research centre langues et civilisations à tradition orale (LACITO) in the 1990s, as a natural extension of traditional methods in linguistic fieldwork. As of 2019, the Pangloss Collection hosts about 170 languages, with 1900 hours of recordings (about 70% are transcribed and annotated). The resources in the Pangloss Collection benefit from long-term archiving services. Almost all resources are open access, so they are available for a variety of uses, for specialists but also for the general public and, last but not least for research in Natural Language Processing.

Native language: La Collection Pangloss : une archive ouverte de langues peu documentées conçue pour faciliter des emplois en Traitement Automatique des Langues

La Collection Pangloss a été créée par le laboratoire de langues et civilisations à tradition orale (LACITO), dans les années 90, dans le prolongement des méthodes classiques d'enquête et d'analyse de la linguistique de terrain. En 2019, la Collection Pangloss regroupe environ 170 langues, avec 1900h d'enregistrements (dont environ 70% transcrit et annoté). Les ressources de la Collection Pangloss bénéficient de services d'archivage pérenne. La quasi-totalité des ressources est en accès libre, elles sont donc disponibles pour divers usages : découverte, enseignement, recherche. Recherche en linguistique et anthropologie mais aussi, grâce au numérique, recherche dans le traitement automatique des langues.

P.5.10: Multi-lingual Support in Connective Learning Scheme for Refining and Connecting the Open Educational Videos
Authors: Virach Sornlertlamvanich, Nannam Aksorn and Thatsanee Charoenporn
Country: Japan

Abstract: Tons of educational videos are available online. It is a big burden for learners to figure out the videos they need in the preferred time and language. Not all videos are suitable for learning according the length and presentation components. According to the Sweller's cognitive load theory, the working memory in learning process is very limited, the learner must be selective to what information from sensory memory to pay attention. In the connective learning, we effectively apply NLP approach to refine the video subtitle in archiving, translating, summarizing, classifying, and labelling the relevant keywords to create the multi-lingual learner-friendly environment.

Native language: ปัจจุบันมีวิดีโอเพื่อการศึกษาที่เผยแพร่ออนไลน์มากมาย จึงเป็นการไม่สะดวกสำหรับผู้เรียนในการหาวิดีโอที่ต้องการได้ ซึ่งส่วนใหญ่ก็ต้องเลือกดูบางส่วนก่อนเพื่อให้ทราบเนื้อหา และวิดีโอส่วนใหญ่ก็เป็นภาษาอังกฤษหรือภาษาอื่นๆ ที่ผู้เรียนไม่สันทัดมากนัก จากทฤษฎีการเรียนรู้ (cognitive load theory) ของ Sweller ที่ได้กล่าวไว้ว่าในกระบวนการเรียนรู้ผู้เรียนจำเป็นต้องอาศัยหน่วยความจำชั่วคราว (working memory) ซึ่งมีพื้นที่จำกัด ดังนั้นเพื่อให้การเรียนรู้มีประสิทธิภาพสูงสุด งานวิจัยนี้ได้นำเสนอการใช้การประมวลผลภาษาธรรมชาติเพื่อช่วยในการจัดเก็บคำบรรยายประกอบ แปลคำบรรยาย ย่อความ จำแนก และสกัดคำสำคัญสำหรับการนำเสนอบทเรียนด้วยภาษาที่ต้องการและปรับแต่งให้เป็นวิดีโอที่เหมาะสมตามทฤษฎีการเรียนรู้

P.5.11: Promoting and Preserving Philippine Culture and Languages through Language Technologies
Authors: Ethel Ong, Nathalie Rose Lim-Cheng, Charibeth Cheng and Edward Tighe
Country: Philippines

Abstract: Advances in language technologies enabled the computational representation, processing and generation of human languages that gave computers the ability to analyze varying text and participate in human conversations. Digital resources such as lexicons and textual corpora led to the development of intelligent agents that can perform tasks in language translation and generation, sentiment analysis, fake news detection, and text mining. In this poster, we present our work in preserving and promoting the Philippine culture and language through computer-generated stories and descriptions of museum artifacts, and the analysis of bilingual social media posts to detect public sentiments and monitor public health.

Native language: Ang pagsulong sa teknolohiyang pang-wika na nagsasagawa ng representasyon, pagproseso at automatikong pagsulat ang nagbigay sa kompyuter ng kakayahang pag-aralan ang iba't ibang teksto at makilahok sa pakikipag-usap sa tao. Ang mga yamang wika tulad ng talasalitaan at mga sulatin ang nagpalaganap sa Artificial Intelligence upang magsagawa ng pagsalin at pagsulat ng wika, pagsuri ng sentimento, pagtuklas ng huwad na balita, at pagmina ng teksto. Nais naming ipakita sa poster na ito ang iba't ibang mga gawain sa pagpapanatili at pagtataguyod ng kultura at wika ng Pilipinas sa pamamagitan ng mga kwento na nilikha ng kompyuter at mga paglalarawan ng mga bagay sa museo, at pagsusuri ng mga sulatin ng publiko sa social media upang makita ang mga damdamin ng publiko at subaybayan ang kalusugan ng publiko.

P.5.12: Improvement of Thai NER and the Corpus
Authors: Thatsanee Charoenporn and Virach Sornlertlamvanich
Country: Japan

Abstract: Thai named entity (NE) corpus is rarely found though the named entity recognition (NER) task can make a big contribution in processing the huge amount of available texts. We propose an iterative NER refinement method using BiLSTM-CNN-CRF model with word, part-of-speech, and character cluster embedding to clean up the existing NE tagged corpus due to its inconsistent and disjointed annotation. As a result, in the newly generated corpus, we obtain 639,335 NE tags, much larger than the original size of 172,232 NE tags. The generated model by the newly generated corpus also improves the NER F1-score 16.21% to mark 89.22%.

Native language: การพัฒนาคลังข้อความภาษาไทยสำหรับการประมวลผลภาษาธรรมชาตินั้น มีประเภทและปริมาณเพิ่มมากขึ้น แต่คลังข้อความชื่อเฉพาะภาษาไทย หรือ Thai Name Entity Corpus ยังคงมีทั้งจำนวนที่จำกัด แม้ว่างานวิจัยด้านการรู้จำชื่อเฉพาะ (Name Entity Recognition: NER) จะส่งผลต่อความถูกต้องของการประมวลผลข้อความเป็นอย่างมากก็ตาม งานวิจัยนี้ เสนอวิธีการปรับแต่ง NER แบบวนซ้ำโดยใช้แบบจำลอง BiLSTM-CNN-CRF ประกอบกับคำแวดล้อม หน้าที่ของคำ และกลุ่มอักขระข้างเคียง เพื่อปรับปรุงคลังข้อความชื่อเฉพาะภาษาไทย จากเดิมจำนวน 172,232 ชื่อ ให้มีความถูกต้อง แม่นยำ และสอดคล้องกัน ผลการวิจัยพบว่า คลังข้อความชื่อเฉพาะภาษาไทยที่ปรับปรุงขึ้น ประกอบด้วยคำและป้ายระบุชื่อเฉพาะ (Tags) จำนวนถึง 639,335 ชื่อ ทั้งนี้ ผลการปรับปรุงคลังข้อความชื่อเฉพาะด้วยแบบจำลองที่นำเสนอนี้สามารถกำกับชื่อเฉพาะภาษาไทยมีความถูกต้อง ที่วัดด้วยค่า F1-score ได้ที่ 89.22 เปอร์เซ็นต์ ซึ่งให้ผลที่ดีกว่าแบบจำลองที่สร้างด้วยคลังข้อความเดิมถึง 16.21 เปอร์เซ็นต์

P.5.13: Deploying Language Technologies for Underserved Communities
Authors: Kalika Bali, Monojit Choudhury, Sunayana Sitaram and Sebastin Santy
Country: India

Abstract: Gondi, a South-Central Dravidian language in the vulnerable category on UNESCO's Atlas of the Worlds Languages in Danger (Moseley,2010). Spoken by nearly 3 million people (India, 2011) it is also one of the least resourced languages in India, with little available data and technology. In this poster, we will present Adivasi Radio, a mobile application that provides the Gond tribal community with access to information. We believe that by focusing on Gondi, we will not only empower the Gondi community but also help understand and create a framework to serve as a guide for introducing language technologies in under-served communities.

Native language: गोंडी, यूनेस्को के एटलस ऑफ़ द वल्र्ड्स लैंग्वेजेस ऑफ़ डेंजर (मोसेली, 2010) में एक दक्षिण-मध्य द्रविड़ भाषा की असुरक्षित श्रेणी में है। लगभग ३0 लाख लोगों (भारत जनगणना, २०११) द्वारा बोली जाने वाली यह भारत की सबसे कम पुनर्जीवित भाषाओं में से एक है, जिसमें बहुत कम डेटा और तकनीक उपलब्ध है। इस पोस्टर में, हम आदिवासी रेडियो पेश करेंगे, जो एक मोबाइल एप्लिकेशन है जो गोंड आदिवासी समुदाय को सूचना तक पहुंच प्रदान करती है। हमारा मानना है कि गोंडी पर ध्यान केंद्रित करने से, हम न केवल गोंडी समुदाय को सशक्त बनाएंगे, बल्कि पिछड़े समुदायों में भाषा प्रौद्योगिकी को शुरू करने के लिए एक मार्गदर्शक के रूप में काम करने में मदद करेंगे।

P.5.14: Language Technologies at the University of the Philippines DIliman
Authors: Angelina Aquino and Rhandley Cajote
Country: Philippines

Abstract: Research initiatives on language technologies at the University of the Philippines Diliman are the result of collaborations between the disciplines of engineering and linguistics. These efforts can be categorized into database development, speech processing, text processing, and linguistic analysis. We highlight some of the recent developments in speech processing, speech synthesis, and natural language understanding for the Filipino language and other major languages in the Philippines. We demonstrate some of these technologies in applications such as closed-captioning, automated reading tutors, and literacy assessment. We also discuss efforts towards documentation, preservation, and historical linguistics for indigenous Philippine languages.

Native language: Ang mga pananaliksik sa pangwikang teknolohiya sa Unibersidad ng Pilipinas Diliman ay bunsod ng pakikipagtulungan ng mga disiplina ng inhinyeriya at linggwistiks. Ang mga pagsisikap na ito ay maaaring maiuri sa pagbuo ng mga korpus, pagproseso ng pagsasalita, pagproseso ng teksto, at pagsusuri ng wika. Binibigyang-diin namin ang ilan sa mga kamakailang pag-unlad sa pagproseso ng pagsasalita, sintesis ng wika, at pag-unawa ng wikang Filipino at iba pang mga pangunahing wika sa Pilipinas. Ipinapakita namin ang ilan sa mga gamit ng teknolohiyang ito tulad ng closed-captioning, gabay sa pagbasa, at pagsusuri ng kasanayan sa pagbasa. Itinatatalakay din namin ang mga gawain tungo sa dokumentasyon, pagpapanatili, at pangkasaysayang linggwistiks para sa mga katutubong wika ng Pilipinas.

P.5.15: Languages and Technology in Bhutan
Authors: Tenzin Namgyel
Country: Bhutan

Abstract: Dzongkha is the national language of Bhutan. Officially, there are 19 languages in Bhutan according to the survey carried out in 1991. Dzongkha efficiently serves as the official language of Bhutan while Chökê serves as the language of Dharma and liturgy; and English is apparently used as the necessary foreign language while mother tongues are used at the grass-roots level. Though we are working on development of technology for Bhutanese languages, we are acutely challenged with the lack of expertise and funding.

Native language: རྫོང་ཁ་འདི་འབྲུག་གི་རྒྱལ་ཡོངས་སྐད་ཡིག་ཨིན། སྤྱི་ཟླ་༡༩༩༡ ལུ་འབད་མི་ཞིབ་འཚུག་དང་འཛིལ་སྡུད་ འབྲུག་འབྲེལ་དུ་ འབྲུག་ལུ་སྐད་ཡིག་ ༡༩ ཡོད། རྫོང་ཁ་འདི་གཞུང་འབྲེལ་ཀྱི་སྐད་ཡིག་ཞིན་དང་ གཡུས་སྐད་གཞན་ཀུ་སྤྱིར་བཏང་འབྲུག་མི་ཚུ་གིས་རང་རྐྱངམ་དོ་ཁ་ཤས་ཀྱི་ཡོད་པ་ཨིན། དེ་ལས་ ཨིང་སྐད་འདི་མེད་ཐབས་མེད་པའི་ཐྱི་རྒྱལ་ཀྱི་ཞི་ཐུབ་ཨིན་དང་རོམ་སྐད་འདི་རོམ་ཀྱི་དོན་ལུ་ལག་ལེན་འཐབ་དོ་ཡོད་པ་ཨིན། ང་བཅས་ཀྱིས་ འབྲུག་པའི་ཁ་སྐད་ཀྱི་འཛུམ་རིག་གོང་འཕེལ་གཏང་ནའི་ལུ་ཚུ་ད་འབད་དོ་ཡོད་པ་ཨིན་རུང་ ལྷོགས་གྲུབ་ཅན་ཀྱི་མི་རྟོག་དང་མ་དཏུལ་ཚུ་མེད་པའི་དཀའ་ངལ་སྤྲོོག་ར་འབྱུང་དེ་ཡོད་པ་ཨིན།

P.5.16: Language technology at MILE Lab, Indian Institute of Science
Authors: Ramakrishnan AngaraiGanesan
Country: India

Abstract: Medical Intelligence and Language Engineering Laboratory at the Department of Electrical Engineering, Indian Institute of Science is engaged in research, development and deployment of technologies that facilitate individuals, organisations and governments to easily carry out their daily operations, amidst the existence of multiple, living Indian languages. Using the OCR and TTS developed, hundreds of books have been converted to Braille and audio books, which are being used by blind students. Also developed are online handwriting recognition, browser-based tools to instantly convert text across different Indian scripts that enable people to read text in other languages through their own script.

Native language: இந்தியாவில் பெங்களூரிலுள்ள இந்திய அறிவியல் பயிலகத்தில் மின் பொறியியல் துறையிலுள்ளது மருத்துவ நுண்ணறிவு மற்றும் மொழிப் பொறியியல் ஆய்வகம், இந்த ஆய்வகம், மொழித் தொழில்நுட்பங்களின் ஆராய்ச்சி, உருவாக்கம் மற்றும் பயனாக்கம் ஆகியவற்றில் ஈடுபட்டுள்ளது. இதனால், பல்வேறு வழங்கு மொழிகள் இந்தியாவில் இருப்பினும், தனிநபர்கள், நிறுவனங்கள் மற்றும் அரசாங்கங்கள் தங்கள் அன்றாட நடவடிக்கைகளை எளிதில் மேற்கொள்ள இயலும்.

இங்கு உருவாக்கப்பட்ட அச்சிடப்பட்ட உரை எழுத்து உணர்வி மற்றும் பேச்சு உருவாக்கி ஆகிய மென்பொருட்களைப் பயன்படுத்தி, அறுநூற்றுக்கும் மேற்பட்ட பள்ளி, மற்றும் கல்லூரிப் புத்தகங்கள் பிரெய்ல் மற்றும் பேசும் புத்தகங்களாக மாற்றப்பட்டுள்ளன. அவை பார்வையற்ற மாணவர்களால் பயன்படுத்தப்படுகின்றன. டேப்லெட்களில் எழுதப்பட்ட உரையை உணரும் கருவியும் உருவாக்கப்பட்டுள்ளது. மற்றும் வெவ்வேறு இந்திய லிபிகளிலுள்ள உரையை உடனடியாக பிற மொழிகளில் மாற்றுவதற்கான உலாவி அடிப்படையிலான மென்பொருளால் தங்கள் சொந்த எழுத்துரு மூலமே படிக்க முடியும்.

P.5.17: MULTILINGUAL PROFILE OF INDIA
Authors: Udaya Narayana Singh
Country: India

Abstract: South Asia is the largest repository of multi-lingual and pluri-ethnic civilization, dating back to 2600 BC. Beginning from the Indus Valley Civilization, traces of which were found in India, Pakistan, and Afghanistan, it had two writing systems - Brāhmī (5th CBC) & Kharoṣṭhī (4thCBC) both widely used in many Asian countries. With its 1,576 rationalized mother-tongues and 1,796 other mother-tongues, India has 122 major languages and 200 minor ones. Here, 96% speak only 20-odd IA & Dravidian languages included in the 8th Schedule of the Constitution, and 50 were found to have a high vitality and use.

Native language: দক্ষিণ এশিয়া হলো বিশ্বের বহুভাষিক ও বহুসাংস্কৃতিক সভ্যতা গুলির মধ্যে একটি অন্যতম শ্রেষ্ঠ উদাহরণ. খ্রিস্ট পূর্ব ২৬০০ সালে এর ভাষিক মিলন ও মিশ্রণ প্রক্রিয়া শুরু হয়েছিল. সিন্ধু সভ্যতার সময় থেকেই এর বহু উদাহরণ আধুনিক ভারত, পাকিস্থান ও আফগানিস্থানে পাওয়া যায়. লিখিত রূপে উত্তরণ দুটি লিপির ব্যবহার দিয়ে শুরু হয়. একটি হলো 'ব্রাহ্মী' লিপি (খ্রিস্ট পূর্ব পঞ্চম শতক থেকে), অন্যটি খরোষ্ঠী (খ্রিস্ট পূর্ব চতুর্থ শতক). এই দুটিই এশিয়ার বহু দেশে ছড়িয়ে পড়েছিল. এখন প্রায় ১৫৭৬ নথিভুক্ত মাতৃভাষা ও ১৭৯৬ অন্য মাতৃভাষা সমেত ১২২ টি প্রমুখ ভাষা আছে ভারতে যার অতিরিক্ত রয়েছে শ' দুয়েক ছোট ছোট ভাষা সম্প্রদায়. শতকরা ৯৬ জন এর মধ্যে কুড়িটি ভারোপিয় ও দ্রাবিড় ভাষায় কথা বলেন যেগুলি ভারতীয়সংবিধানের অষ্টম অনুসুচিতে স্থান পেয়েছে. এগুলির মধ্যে পঞ্চাশটি ভাষা এমন যাতে প্রচুর লেখা ও ছাপা হচ্ছে.

P.5.18: Mediating Multilingualism
Authors: Udaya Narayana Singh, Esha Jainiti, Rusha Mudgal and Anwita Maiti
Country: India

Abstract: "Mediating Multilingualism" is a collaborative project for language revival and survival between Amity University Haryana and the University of Highlands & Islands, UK. It encourages video ethnographic documentation of some marginalized speech communities, with special focus on the Indian context based on work done earlier in Scotland and Ireland. Colloquiums, seminars and workshops organized focused on protecting and promoting the endangered languages of India. The primary aim is building capacity and networking with other academic partners, take up interdisciplinary research from both sociolinguistic and Communication technology angle. It will be followed by a desk-based monitoring, assessment and theoretical work.

Native language: “মেডিয়েটিং মাল্টিলিঙ্গুয়ালিজম” অ্যামিটি বিশ্ববিদ্যালয় হরিয়ানা (AUH) এবং ইউনিভার্সিটি অফ হাইল্যান্ডস এবং আইল্যান্ডস (UHI), স্কটল্যান্ড, ইউ-কে--র মধ্যে ভাষা পুনরুজ্জীবন এবং ভাষা সংরক্ষণ বিষয়ক আকাদেমিক মধ্যস্থতা করার সম্ভাবনা খুঁটিয়ে দেখার জন্য একটি সহযোগী প্রকল্প - যে ফোরামটি কিছু প্রান্তিক ভাষা ও সংস্কৃতি বিষয়ে সেই সব সম্প্রদায়ের আলোচনা এবং ভিডিও-র মাধ্যমে নৃতাত্ত্বিক ডকুমেন্টেশনকে উৎসাহিত করবে, এবং এর আগে স্কটল্যান্ড এবং আয়ারল্যান্ডে করা কাজের ভিত্তিতে ভারতীয় প্রসঙ্গে সেই অভিজ্ঞতা কতটা কাজে লাগানো যায় সেবিষয়ে বিশেষ মনোযোগ দেবে. কথোপকথন, সেমিনার এবং ওয়ার্কশপগুলি ভারতের আদিবাসী ও বিপন্ন ভাষা সুরক্ষিত ও প্রচারের দিকে দৃষ্টি নিবদ্ধ করবে । প্রাথমিক লক্ষ্য দক্ষতা তৈরি করা এবং অন্যান্য একাডেমিক অংশীদারদের সাথে নেটওয়ার্কিং করা, অন্তর্বিভাগীয় গবেষণা কার্যক্রম গ্রহণ করা, ভাষাতত্ত্ব ও যোগাযোগ প্রযুক্তি দুই দিক থেকেই দক্ষতা অর্জন করা। এটি একটি পর্যবেক্ষণ, মূল্যায়ন এবং তাত্ত্বিক কাজ অনুসরণ করবে।

P.5.19: Technology Development for Indian Languages
Authors: Vijay Kumar and Dr S K Srivastava
Country: India

Abstract: Technology Development for Indian Languages (TDIL) Programme of Government of India has been sponsoring projects for development of Linguistic Resources, Standards and Technologies like Fonts, Unicode Typing Tool, Localized Open Source Software, Machine Translation Systems, Speech Technologies (TTS, ASR), Optical Character Recognition, etc. The developed prototypes are accessible through http://www.tdil-dc.in. A new initiative viz. “Natural Language Translation Mission” is being started with an objective to build a speech to speech translation system for major Indian languages in the domains like Science & Technology, Education, Healthcare, Law& Justice and Governance.

Native language: भारत सरकार का भारतीय भाषाओं के लिए प्रौद्योगिकी विकास (टीडीआईएल) कार्यक्रम भाषाई संसाधनों के विकास , मानकों और प्रौद्योगिकी जैसे फोंट, यूनिकोड टाइपिंग टूल, स्थानीयकृत ओपन सोर्स सॉफ्टवेयर, मशीन ट्रांसलेशन सिस्टम्स, स्पीच टेक्नोलॉजीज (टीटीएस, एएसआर), ओसीआर आदि के विकास के लिए परियोजनाओं को प्रायोजित कर रही है। विकसित किए गए प्रोटोटाइप को http://www.tdil-dc.in के माध्यम से देखा जा सकता है। विज्ञान और प्रौद्योगिकी, शिक्षा, स्वास्थ्य, कानून और न्याय , शासन जैसे क्षेत्रों में प्रमुख भारतीय भाषाओं के लिए स्पीच टु स्पीच ट्रांसलेशन सिस्टम का निर्माण करने के उद्देश्य से एक नई पहल की जा रही है ।

P.5.20: CREATING ACCESS TO OPENLY LICENSED EARLY READING RESOURCES IN ASIA'S INDIGENOUS LANGUAGES
Authors: Purvi Shah
Country: India

Abstract: StoryWeaver is an open-access digital platform with 17,000+ children’s storybooks in 200+ languages. 78 of these are Asian, including 16 Indigenous languages. The platform contains tools to read, translate, publish, download, print books – all for free. This poster presentation highlights how StoryWeaver uses technology and fosters communities to build a large repository of local language reading resources: from creating bilingual books to aid a child's transition to the language of instruction in school, to facilitating the self-expression of Indigenous groups by publishing books in languages without writing systems, and finally creating sustainable self-publishing models by leveraging open licences.

Native language: स्टोरीवीवर बाल कहानियों का खुला मंच है जहाँ 200 से अधिक भाषाओं में 16000 से अधिक कहानियाँ मौजूद हैं। इनमें 78 एशियायी और 16 भारतीय भाषाएँ शामिल हैं। यहाँ कहानियाँ पढ़ी, अनुवाद, प्रकाशित, डाउनलोड और मुद्रित की जा सकती हैं- बिल्कुल निशुल्क। यह प्रस्तुति बताती है कि स्थानीय भाषाओं में पठन संसाधनों का विकास करने में स्टोरीवीवर किस प्रकार सहायक बना- द्विभाषी पुस्तकों के द्वारा स्कूली और मातृभाषाओं के बीच पुल बनके, क्षेत्रीय समूहों द्वारा अपनी बिना लिपि वाली भाषाओं में स्वाभिव्यक्ति का साधन बन, अंततः ओपन लाइसेंस का लाभ उठाकर कारगर स्वप्रकाशन मॉडल तैयार करके।

P.5.21: Conversational Bot for Eyesight Testing Automation
Authors: Ari Yanase, Thatsanee Charoenporn and Virach Sornlertlamvanich
Country: Japan

Abstract: The people’s visual acuity can be examined using a standard Snellen eye chart by determining a relationship between the sizes of certain letters viewed at certain distances, or a broken wheel vision chart. In general, we examine our eyesight by reading the letter “A” with 88 mm in height at 20 feet or 6 meters in distance. To lessen the burden of the ophthalmologist, we equip an NLP based chatbot with basic eyesight testing knowledge in a mobile app. With the natural conversation, the bot recognizes the response from the subject, and returns the prescriptions for eyeglasses measured in diopters.

Native language: 人々の視力測定には、文字のサイズなどを元に作成されたSnellenのモデルと壊れたタイヤにまつわるモデルが使われています。一般的には、６メートルもしくは２０フィート離れた距離から、高さ８８ミリで書かれたＡの文字が読めれば、１.０の視力を持っている、と言われています。私達が視力測定の基本的知識を備えた自然言語処理を元に作成されたチャットボットを備える、作成することで、眼科医の負担になる仕事を減らしました。また、自然言語を用いて、眼鏡の度を予測したりもできます。

P.5.22: Dictionary 4.0: Alternative Presentations for Indonesian Multilingual Dictionaries
Authors: Arbi Haza Nasution and Totok Suhardijanto
Country: Indonesia

Abstract: Building a multilingual dictionary for 719 languages in Indonesia is a challenging task. We have developed application to create the Leipzig-Jakarta list database for all indigenous languages in Indonesia. The database can be used to generate lexical similarity or lexical distance matrix between languages by comparing the word list. For starter, we covered 11 languages: Indonesian, Javanese, Sundanese, Madurese, Bima, Ternate, Tidore, Palembang Malay, Mandailing Batak, Malay, and Minangkabau. The application has two main features: exploring the existing translations and adding translations to a new language or editing existing translations through crowdsourcing. User acceptance test showed 3.48 / 4 score.

Native language: Membangun kamus multibahasa untuk 719 bahasa di Indonesia adalah tugas yang berat. Kami telah mengembangkan aplikasi untuk membuat pangkalan data daftar Leipzig-Jakarta untuk semua bahasa daerah di Indonesia. Pangkalan data tersebut dapat digunakan untuk menghasilkan kesamaan leksikal atau matriks jarak leksikal antar bahasa dengan membandingkan daftar kata tersebut. Sebagai permulaan, aplikasi ini mencakup 11 bahasa: Indonesia, Jawa, Sunda, Madura, Bima, Ternate, Tidore, Melayu Palembang, Batak Mandailing, Melayu, dan Minangkabau. Aplikasi ini memiliki dua fitur utama: menjelajahi terjemahan yang ada dan menambahkan terjemahan ke bahasa baru atau mengedit terjemahan yang ada melalui mekanisme urun daya dari masyarakat. Uji keberterimaan pengguna menunjukkan skor 3,48 / 4.

P.5.23: Speech Technology in three tonal languages of North-East India
Authors: Viyazonuo Terhiija, Samudra Vijaya and Priyankoo Sarmah
Country: India

Abstract: An account of speech systems implemented in three tonal languages of North East India is presented here. Angami, Ao, and Mizo are under-resourced languages of the Tibeto-Burman language family, spoken in North-East of India. The distribution of tones across the three languages vary from three to four tones. Automatic Speech Recognition systems were developed for Angami and Mizo languages using Kaldi toolkit. The performances of various versions of the system using different types of acoustic models (GMM-HMM, SGMM and DNN) are reported. In Ao, a dialect identification system was built that distinguishes two Ao dialects, namely, Changki and Mongsen.

Native language: Diepu kesezho nu North East India nu chiiwaketa die pfhephra se pie kepushiizhie. Angami, Ao mu Mizo seyakezha dieko Tibeto - Burman die kikru nu ba. Die hako puo pfhephra kezakeshii ki pfhephra se mu dia mese kekrei ba. Automatic speech recognition systems-e Angami mu Mizo die la Kaldi Toolkit se di chii kehielie. Die puo zho kekreikecii mhathokoe acoustic models (GMM-HMM, SGMM and DNN) kekreikecii se di chiikelie silie. Ao mia-e die si kekreilie kevi zho puo chiilie di uko die kenie Changki mu Mongsen si kekreilie kevi chiilie.

P.5.24: Bringing Zero-resourced Languages of Myanmar to the Digital World
Authors: Win Pa Pa
Country: Myanmar

Abstract: Accessing Technology in one's own language is promoting personnel, social and economic life of that one and also preserving the language. Myanmar language(Burmese) is one of low-resourced language although it is spoken by 33 millions people as their first language. Data scarcity of Myanmar language is a big challenge for language processing. Recent technologies and resources of Machine Translation, Automatic Speech Recognition and Speech synthesis are presented in the poster to promote Myanmar languages and its dialects' language processing.

Native language: နည်းပညာအသုံးပြုမှုကို မိမိတို့ ကိုယ်ပိုင် ဘာသာစကားဖြင့် ပြုလုပ်နိုင်ခြင်းသည် မိမိတို့၏ ကိုယ်ပိုင်ဘဝ၊ လူမှုပတ်ဝန်းကျင် နှင့် စီးပွားရေး တိုးကို မြင့်တင်ပေးရာ ရောက်ပြီး ထို ကိုယ်ပိုင်စကားကို ထိန်းသိမ်းရာလည်း ရောက်ပေသည်။ မြန်မာဘာသာစကားသည် မိခင်ဘာသာစကား အဖြစ် အသုံးပြုသူ ဒုရုသန်း ရှိသော်လည်း အရင်းအမြစ် နည်းပါးသော ဘာသာစကားများတွင် တစ်ခု အပါအဝင်ဖြစ်သည်။ မြန်မာဘာသာ အချက်အလက် ရှားပါးခြင်း သည် ဘာသာစကားနည်းပညာ အတွက် ကြီးမားသော စိန်ခေါ်မှု တစ်ခုဖြစ်သည်။ ပို့ကောတွင် မြန်မာဘာသာစကားအတွက် စက်ဖြင့်ဘာသာပြန်ခြင်း၊ အသံမှ စာသား သို့ပြောင်းခြင်း နှင့် စာသားမှ အသံသို့ပြောင်းခြင်း လုပ်ငန်း၏ လက်ရှိ နည်းပညာများ ကို ဖော်ပြထားပြီး မြန်မာ နှင့် တိုင်းရင်းသား ဘာသာစကားများ နှင့် ဒေသိယ စကားများ အတွက် နည်းပညာများ ကို လူအများ၊ စီတင်စားမှု မြင့်တင်ရန် ရည်ရွယ်ပါသည်။

P.5.25: Building Corpora for Under-Resourced Languages in Indonesia
Authors: Totok Suhardijanto and Arawinda Dinakaramani
Country: Indonesia

Abstract: Indonesia has known as the second most linguistically-diverse country, but ironically also known as a country with many under-resourced languages. In this poster, we present our attempt to develop language resources in Indonesian indigenous languages for linguistic research purposes. For the first stage, we developed corpora for Javanese, Sundanese, Malay/Indonesian, and Minangkabau which are chosen because of the number of speakers. This poster discusses the drawbacks and opportunities in our attempt to build Indonesian language corpora that are publicly accessible. The corpus application is still under development, but it is a good step to start compiling language corpora in Indonesia.

Native language: Indonesia dikenal sebagai negara paling kaya kedua dalam hal bahasa, namun ironisnya juga dikenal dengan negara paling sedikit sumber daya bahasanya. Poster ini menyajikan upaya kami dalam membangun korpora untuk bahasa-bahasa di Indonesia untuk kepentingan kajian bahasa. Pada tahap awal, dipilih bahasa Jawa, Sunda, Melayu/Indonesia, dan Minangkabau sebagai konten korpora karena bahasa-bahasa itu mempunyai jumlah penutur banyak. Poster ini akan membahas kendala dan kesempatan dalam menyusun korpora bahasa-bahasa di Indonesia. Aplikasi korpusnya pun masih dalam tahap pengembangan, namun ini merupakan langkah baik untuk mengembangkan korpus untuk bahasa-bahasa di Indonesia.

P.5.26: Language Resources and Technology Development Efforts for some Lesser-known Indian Languages
Authors: Ritesh Kumar, bornini lahiri, Atul Kr. Ojha, Mayank Jain and Deepak Alok
Country: India

Abstract: For the last few years, we have been involved in the development of language technologies and resources for some of the lesser-known Indian langauges viz. Magahi, Bhojpuri, Awadhi and Braj Bhasha. These languages (among others) have been largely marginalised and ignored and have low prestige and negative attitude towards them because of these being considered ‘illiterate’ and ‘rural’ varieties of Hindi. Our poster will showcase different kinds of corpora as well as basic technologies like part-of-speech tagger, morphological analyser as well as some applications like machine translation systems that have been developed for these languages.

Native language: पिछला कै साल से, हमनी मगही, भोजपुरी, अवधी आउ ब्रज भासा जईसन भासा, जेकरा पर बहुते कम काम होल हई, ओकर प्रौद्योगिकि आउ संसाधन के विकास में लगल ही । इ सब भासा (औ अइसन बड़ीमनी) के नजर अंदाज और हासिया पर कर देल गेल हे आउ कम परतिस्ठा आउ नीचा दृस्टि से देखल जा हे काहे की इ सब हिंदी भासा के देहाती आउ असीछित बोली के प्रकार समझल जा हे। हमनी के पोस्टर बिभिन्न प्रकार के कॉर्पोरा के साथ-साथ बुनयादी प्रौद्योगिकि जैसे की सब्द भेद टैगर (पार्ट ऑफ़ स्पीच टैगर), रूप सम्बन्धी बिश्लेसक (मोरफोलॉजिकल अनलाइज़र) आउ कुछ अनुप्रयोग सॉफ्टवेयर जैसे की मसीनी अनुवाद पद्धति/सिस्टम जे इसब भासा ला विकिसित कइल गइल हे ओकरा परदरसीत करत।

P.5.27: A 1000-language Collaborative Universal Dictionary and Universal Translator

Authors: David Yarowsky, Arya D. McCarthy, Garrett Nicolai, Winston Wu, Aaron Mueller, Dylan Lewis, Yingqi Ding, Abhinav Nigam, Emre Ozgu, Debanik Purkayastha, James Scharf and Kenneth Zheng

Country: United States

Abstract: We present JHU's Universal Dictionary and Universal Translator, covering 1000+ world languages in a broadly-accessible Android/IOS mobile phone and web-browser app, with 1,000,000+ planet-wide language pairs and 100's of under-resourced languages which have never had access to a substantial dictionary or machine translation capability. In addition to providing immediate access to a base vocabulary of 1500-20000 core vocabulary lemmas in all 1000+ languages, this novel app actively engages its users to contribute collaboratively to the universal dictionary in an easy-to-use and efficient way, with automatic suggestion of possible translations based on sound-shift transductions from related languages and pan-linguistic compositional constructions.

Native language: Presentamos el Diccionario Universal y el Traductor Universal de JHU, que cubren más de 1000 idiomas del mundo en una aplicación de navegador web y teléfono móvil Android / IOS de amplio acceso, con más de 1,000,000 de pares de idiomas en todo el planeta y cientos de idiomas con recursos insuficientes que nunca han tenido acceso a un diccionario sustancial o capacidad de traducción automática. Además de proporcionar acceso inmediato a un vocabulario base de lemas de vocabulario básico de 1500-20000 en los más de 1000 idiomas, esta nueva aplicación involucra activamente a sus usuarios para que contribuyan colaborativamente al diccionario universal de una manera fácil de usar y eficiente, con sugerencia de posibles traducciones basadas en transducciones de cambio de sonido de lenguajes relacionados y construcciones composicionales pan-lingüísticas.

P.5.28: Tagalog-English Code-Switching: Challenges for Automatic Detection

Authors: Nathaniel Oco

Country: Philippines

Abstract: In this poster, I will give a brief overview of the Philippines - a country in Southeast Asia - and discuss Tagalog-English code-switching (TECS). TECS is the use of both Tagalog and English in a discourse. I will also detail existing works to automatically detect TECS and the challenges ahead (e.g., intra-word code-switching and interlingual homographs).

Native language: Sa poster na ito, ako'y magbibigay ng maikling panimula sa Pilipinas - isang bansa sa Timog-silangang Asya - at aking tatalakayin ang Taglish. Ang Taglish ang paggamit sa parehas Tagalog at Ingles sa isang diskurso. Ako rin ay magbabahagi ng mga kaugnay na pananaliksik hingil sa awtomatikong pagsusuri sa Taglish at ang mga hamong kinahaharap (gaya ng panlalapi sa mga salitang Ingles at ang paggamit ng mga salitang parehas makikita sa Tagalog at Ingles).

P.5.29: How a low-resource named entities recognition and transliteration framework for Vietnamese can improve the automatic machine translation ?

Authors: Tan Ngoc Le and Fatiha Sadat

Country: Canada

Abstract: This presentation focuses on the low-resource pair of languages, French-Vietnamese, in order to develop a powerful machine translation system while focusing on the recognition of named entities and their transliterations. In addition to statistical approaches, we used a deep learning approach within our different systems to further improve the quality and efficiency of automatic translation of named entities and to reduce the rate of words outside vocabularies, untranslated and / or incorrectly translated words, but also to improve the quality of the machine translation system.

Native language: Bài trình bày này tập trung vào cặp ngôn ngữ tài nguyên thấp, tiếng Pháp-tiếng Việt, để phát triển một hệ thống dịch máy mạnh mẽ trong khi tập trung vào việc công nhận các thực thể được đặt tên và phiên âm của chúng. Ngoài các phương pháp thống kê, chúng tôi đã sử dụng phương pháp học sâu trong các hệ thống khác nhau của mình để cải thiện hơn nữa chất lượng và hiệu quả của dịch tự động các thực thể được đặt tên và để giảm tỷ lệ các từ bên ngoài từ vựng, từ chưa được dịch và / hoặc dịch sai, nhưng cũng để nâng cao chất lượng hệ thống dịch máy.

P.5.30: SITUATION AND CHALLENGES OF TECHNOLOGIES FOR INDIGENOUS LANGUAGES OF INDIA

Authors: Shweta Sinha and Shyam Sundar Agrawal

Country: India

Abstract: India is a country with huge linguistic diversity. Out of 900 languages spoken in the country, only a few have witnessed the digital world. This poster presents in detail the Indian languages situation in terms of resources,and technologies . It highlights the relative need , opportunities, barriers and complexities specific to the Indian Languages technologies. The aim is to study their influence on the adoption and adaptation of digital technology vis a vis Technological achievements/ fallout's of Indian languages relating to the world languages and to identify the gap and the need to take up future projects for technological advancements.

Native language: भारत एक विशाल भाषाई विविधता वाला देश है। देश में बोली जाने वाली 900 भाषाओं में से कुछ ही डिजिटल दुनिया में देखी गई हैं। यह पोस्टर, संसाधनों और प्रौद्योगिकियों के संदर्भ में भारतीय भाषाओं की स्थिति को विस्तार से प्रस्तुत करता है। यह भारतीय भाषाओं की प्रौद्योगिकियों के लिए विशिष्ट आवश्यकताओं, अवसरों, बाधाओं और जटिलताओं को उजागर करता है। इसका उद्देश्य डिजिटल प्रौद्योगिकी को अपनाने और विश्व भाषाओं से संबंधित भारतीय भाषाओं की एक तकनीकी उपलब्धियों को अपनाने, अंतर को पहचानने और तकनीकी प्रगति के लिए भविष्य की परियोजनाओं की आवश्यकता पर उनके प्रभाव का अध्ययन करना है।

P.5.31: Unicode for Indigenous Languages - Standards and technology for getting online
Authors: Craig Cornelius
Country: United States

Abstract: Three basic technologies are essential for a community to use its language with digital systems:

1. standardization defining digital codes for the writing system, 2.fonts and rendering technology for display and print
2. tools to create new text such as virtual keyboards and input methods

In addition, community leadership and engagement.

The poster discuss the Unicode Standard and associated technologies that are available today for use by indigenous communities. It will also outline steps that groups to make digital technology appropriate for their community needs.

P.5.32: CASS-LING's Linguistic Infrastructure: Resources, Platforms and Services
Authors: Wei Wang, Aijun Li and Danqing LIU
Country: China

Abstract: As China's highest academic institution of linguistic research, the Institute of Linguistics of Chinese Academy of Social Sciences (CASS-LING) has created language resources and platforms to provide varied services, which include: Contemporary Chinese Dictionary and Xinhua Dictionary with a Guinness World Record as the world's most popular dictionary; the Dictionaries and Speech Archives of Contemporary Chinese Dialects; an online system for dictionary compilation; the Database of the Grammars of the Chinese Dialects; a benchmark Putonghua pronunciation model; a visual 3D pronunciation model for English phonetic learning; the state government's examination and standardization of the pronunciation of characters and words.

Native language: 作为中华人民共和国语言学研究的最高学术机构，中国社会科学院语言研究所近年来搜集、整理大量语言资源，建设完成专业平台，提供以扎实学术研究为基础的社会服务。这些资源、平台、服务包括：《现代汉语词典》和《新华字典》（发行量吉尼斯世界纪录）；41卷本《现代汉语方言大词典》和40册《现代汉语方言音库》；词典编撰的支撑系统；已经搜集了10种方言数据的方言语法数据库；“九州音集”方言语音数据采集和分享微信平台；在调查4000多名儿童发音基础上推出的1.5~6岁普通话儿童语音发音常模；基于1万多小时不同方言区和少数民族地区英语学习者发音材料、面向英语语音教学的可视化3D发音模型；主导普通话审音工作。

P.5.33: Including Linguistic Knowledge in an Auxiliary Classifier CycleGAN for Corrective Feedback Generation in Korean Speech
Authors: Seung Hee Yang and Minhwa Chung
Country: Korea

Abstract: This work introduces a methodology to inject linguistic knowledge into an auxiliary classifier Cycle-consistent Generative Adversarial Network (CycleGAN), a machine learning method that retains domain knowledge. A related work used CycleGAN to generate a native version of a language learner's accented input speech in Korean. A linguistically-motivated auxiliary classifier is proposed in this work that enables generator-student interaction. This additional two-layer CNN learns to ensure the discriminability between the generated samples, and distinguishes three error types, “segmental,” “suprasegmental,” and “no correction” of the generated speech so that the learners will receive corrective feedback together with linguistic information.

Native language: 본 논문에서 제안하는 방법은 보조 분류기로 증강된 순환적 일관성 손실 함수를 사용한 생성적 적대 신경망 (Cycle-consistent Generative Adversarial Network; CycleGAN) 모델이며, 언어학 도메인 지식을 신경망 학습에 활용한다. 이전의 CycleGAN은 학습자의 목소리를 원어민 발음으로 변환하여 들려준다. 그러나, 이러한 피드백 방법은 어떠한 유형의 오류가 있었는지에 대한 정보를 제공하지 못한다는 데에 한계가 있다. 본 연구는 생성 결과 간의 언어학적 차별성을 학습한 보조 분류기를 기존의 CycleGAN에 추가하였다. 먼저 “분절음,” “초분절음,” “오류 없음” 유형별 CycleGAN 학습이 이루어지고, 모든 출력 결과는 하나의 분류기에 다시 입력된다. 최종적으로 교정 피드백은 교정 피드백 음성을 생성함과 동시에 해당 피드백이 어떤 오류인지 함께 제공한다.

P.6.1: Machine Translation 4 All: Developing informed and critical users through a program of machine translation literacy
Authors: Lynne Bowker
Country: Canada

Abstract: Machine translation is easy to use: copy, paste, click. However, just because users know HOW to use this tool doesn't mean that they are able to use it appropriately. In this age of free, online machine translation, we need for a new type of digital literacy: machine translation literacy. Literacy is a cognitive issue, rather than a techno-procedural issue. Machine translation literacy involves knowing not just how, but also whether, when and why to use machine translation. In Canada, we have developed guidelines that can help people who are not trained language professionals to become savvy users of machine translation.

Native language: La traduction automatique pour tous : Former des utilisateurs avertis et critiques avec une approche raisonnée de la traduction automatique

Les outils de traduction automatique sont faciles à utiliser : copier, coller, cliquer. Cependant, ce n'est pas parce que les utilisateurs savent COMMENT utiliser cet outil qu'ils sont bien équipés pour l'utiliser de façon optimale. À l'ère de la traduction automatique gratuite et en ligne, un nouveau type de culture numérique est nécessaire : il faut cultiver une « approche raisonner » de la traduction automatique. Développer une approche raisonnée, c'est une question cognitive plutôt qu'une question techno-procédurale. Une approche raisonnée de la traduction automatique implique de savoir non seulement comment, mais aussi si, quand et pourquoi utiliser la traduction automatique. Au Canada, nous avons élaboré des lignes directrices qui peuvent aider les personnes qui ne sont pas des professionnels langagières à devenir des utilisateurs avertis et critiques de la traduction automatique.

P.6.2: Building a common Digital Infrastructure to sustain Algonquian Languages
Authors: Marie-Odile Junker and Delasie Torkornoo
Country: Canada

Abstract: Our project includes Algonquian languages and communities of speakers, teachers and learners, at different degree of language vitality or endangerment (www.atlas-ling.ca, resources.atlas-ling.ca). Using a collaborative, participatory action research framework, we focus on dictionaries and integrated language resources. Our long-term goal is sustainability. Our design choices include:

- most affordable web server environments
- web frameworks that are reliable and diverse
- matured active open source frameworks
- systems capable of synchronizing online and offline data sources
- integration of the various applications We also raise questions about multimedia components, open-source, data stewardship, and long-term maintenance of not-for-profit resources.

Native language: Notre projet comprend des langues de la famille algonquienne présentant différents degrés de vitalité (www.atlas-ling.ca, resources.atlas-ling.ca). Dans un cadre de recherche particip-action, nous développons des dictionnaires et ressources linguistiques intégrées. Notre objectif est la durabilité. Nos choix de design incluent:

• environnements de serveur Web abordables

• cadres Web fiables et diversifiés

• infrastructures Open-Source matures et actives

• systèmes capables de synchroniser des sources de données en ligne et hors ligne

• intégration de différentes applications Nous soulevons des questions sur les composants multimédias, les logiciels Open-Source, la gestion responsable des données et la maintenance à long terme de ressources à but non lucratif.

P.6.3: On the promise and pitfalls of repurposing existing language technologies for endangered language documentation
Authors: Emily Prud'hommeaux, Robert Jimerson, Richard Hatcher, Raymond Ptucha and Karin Michelson
Country: United States

Abstract: Like many indigenous languages of North America, the Iroquoian language Seneca is endangered, with fewer than fifty living native speakers. Although descriptive grammars of Seneca exist, there are few texts and recordings available to support immersion programs and other revitalization efforts. In a collaboration between university researchers and Seneca community members, we are working to produce textual and audio documentation of the Seneca language using both existing toolkits and custom architectures. We find that while some toolkits yield promising results, the morphological complexity of Seneca and the variable quality of the available recordings present challenges for deploying one-size-fits-all solutions.

Native language: Dah ne:' dih tša'deyo'dëh oya' yei' niyöëdza'geh niodiwënö'dës koh neh onödowa'ga:' gawënö' agaiwahdö't so'jih gao' wis niwashëh niënödih ahsöh deodišnye'öh onödowa'ga:' gawënö'. Gwaheh ha'deyöh gayadö' niyo'dëh onödowá'ga:' gawënö' koh neh dohga:'ah niyoh gawënöhdas ogwenyöh aonödesdë' adiyë'he't onödowá'ga:' gawënö'. Dah ne:' hae'gwah dogëh dwadade'gë:' tënödeyësdahgwa'geh hënöjëönyanih koh neh onödowa'geonö' deodiyenö'. Dah ne:' hae'gwah hodihšöniaje' gayadöshäse:' watšowih na'ot gawënöhdas gayë' hadiyä'ta' dejaöh yesta' koh neh gadogëh neh nigayeëh gahšönih. Dah ne:' wa'agwaiwaho' neh dohga:'ah niyoh yesta' agwas wadesta' gwaheh so'jih ha'deyöh gayë' sgawënot koh neh dewenö:' niyo'dëh gawënohdas da'aöh ahsehšöni' gwisdë' neh ogwenyöh agaya'daei' gagwegöh koh neh agaiwaeis.

P.6.4: 7000 Languages: Free Language-Learning Software for Language Reclamation
Authors: Alexa Little, Kayleigh Jeannette and Kelsey Riggs
Country: United States

Abstract: 7000 Languages is a US non-profit organization that creates free language-learning software in partnership with Indigenous and endangered language communities. As of 2019, we have created 36 courses for 19 different languages, primarily Indigenous languages of North America. In this poster, we will show examples of our work and explain the process of creating a free online language course for your language.

Native language: 7000 Languages (7000 Lenguas) es una organización sin ánimo de lucra estadounidense que colabora con las comunidades indígenas y de lenguas en peligro de extinción para crear programas usadas para aprender esas lenguas. A partir del año 2019, hemos creado 36 cursos para 19 lenguas, que son principalmente de América del Norte. En este póster, vamos a mostrar ejemplos de nuestro trabajo y explicar el proceso de crear un curso libre en la red para la lengua suya.

P.6.5: National Research Council Canada Indigenous Language Technology Project
Authors: Aidan Pine, Nathan Brinklow, Heather Souter and Delaney Lothian
Country: Canada

Abstract: The Canadian federal budget released in 2017 invested $89.9M over three years to support Indigenous languages and cultures. The Canadian National Research Council (NRC) was granted $6M of this funding. Under the guidance of an Indigenous advisory committee, the NRC is researching and developing speech- and text-based technologies which aim to assist the stabilization, revitalization and reclamation of Indigenous languages.

We argue that successful technological solutions must support grass-roots language revitalization efforts and not overshadow them. This poster session describes three separate challenges faced by Indigenous language communities in Canada and highlights three corresponding NRC solutions.

Native language: Ne Canada kakó:ra óntka'we ne 2017 ohwísta, $89.9M nè:ne áhsen niiohserà:ke nika'shén:nes nè:ne aiakoié:nawa'se' ne Onkwehón:we raotiwenna'shón:'a tánon tsi nihotirihó:ten. Ne Canadian National Research Council (NRC) $6M tahonwén:non e'tho tonterá:ko. Onkwehón:we rotitióhkwaien ne shakonaten'nikonhrá:wis ne enhonten'nikón:raren, ne NRC enhatirihó'kwate tánon enhonnón:ni ne tsi ní:ioht ne aiontá:ti tánon tsi ní:ioht tsi akahiatón:hake, á:se nahò:ten ronaterien'tatshén:rion tsi ní:ioht tsi akonwaié:nawa'se' kwah tokén:en aiorihwahnírha, aonsontónhnhete tánon taontá:we ne Onkwehonwehnéha owenna'shón:'a. Iakwá:ton ne aioiá:neren'ne' tsi nahatí:iere ne á:se ronaterien'tatshén:rion tsi nakaié:ren wentá:'on ki' kwah tsi tka'nónhkwaien ne owén:na tsi sewatónhnhet, tóhsa ó:ia' nahò:ten ohén:ton watohétstakw. Kí enkahiatonhseroha:rake áhsen niiorì:wake enwentó:ren'ne' ne Onkwehonwehnéha owén:na aorihwà:ke ne tsi nón:we nikanakerahseraké:ron ne Canada, tánon ó:ni ronne'néstha ne áhsen niiorì:wake tsi nahatí:iere' ne NRC.

P.6.6: St. Lawrence Island Language Technology for Documentation & Revitalization
Authors: Lane Schwartz, Emily Chen, Hayley Park, Sylvia Schreiner and Benjamin Hunt
Country: United States

Abstract: St. Lawrence Island Yupik is an endangered language indigenous to St. Lawrence Island in Alaska and the Chukotka Peninsula in Russia. This poster presents joint work in language documentation and computational linguistic research, designed to promote language revitalization efforts underway on St. Lawrence Island.

P.6.7: Building a Language Model of Nehiyawewin (Cree, Y-dialect)
Authors: Delaney Lothian, Daniela Teodorescu, Denilson Barbosa and Carrie Demmans Epp
Country: Canada

Abstract: Cree (Nehiyawewin) is a polysynthetic language Indigenous to the Canadian prairies. We are building a system intended to identifying morphemes using an N-gram language model of Cree. This research is motivated by focus groups that explored the needs and expectations of Indigenous people of Western Canada with respect to learning their ancestral language through technology. A lack of exposure to practice and limited resources are among the barriers they face. Our findings demonstrate the importance of including language structure as well as cultural meanings in technological tools to support learning ancestral languages instead of focusing only on vocabulary.

Native language: Le cri (Nehiyawewin) est une langue polysynthétique autochtone dans les Prairies canadiennes. Nous construisons un système destiné à identifier des morphèmes en utilisant un modèle de langage N-gramme du cri. Cette recherche est motivée par des groupes de discussion qui ont exploré les besoins et les attentes des peuples autochtones de l'Ouest canadien en ce qui concerne l'apprentissage de leur langue ancestrale par la technologie. Un manque d'exposition à la pratique et des ressources limitées font partie des obstacles auxquels ils sont confrontés. Nos résultats démontrent l'importance d'inclure la structure linguistique ainsi que les significations culturelles dans les outils technologiques afin de soutenir l'apprentissage des langues ancestrales au lieu de se concentrer uniquement sur le vocabulaire.

P.6.8: From Talking Leaves to Pixels: The Evolution of the Cherokee Syllabary
Authors: Roy Boney
Country: United States

Abstract: The Cherokee syllabary has been adapted to new technologies since its invention in 1821. It evolved from handwriting, to the printing press, to manual and electric typewriters, and now is part of the Unicode standard. This poster is a summary of the evolution of the Cherokee syllabary through the 21st century.

Native language: ᎾᏍᎩᏃ ᏣᎳᎩ ᏧᏃᏴᎩ ᎾᏳ 1821 ᏧᏕᎳᏴᏌᏗᏒ ᏗᎪᎷᏩᏛᏓ ᎾᏍᎩᎾᏃ ᏧᏩᏗᏔᏅᏒ ᎢᏤ ᏗᏕᏙᏯᏒ ᏗᏙᎳᎩ ᏂᏧᎵᏍᏔᏅᏍᏔᏅᎢ. ᎾᏍᎩᎾ ᎪᏪᎳᏅᎯ ᎤᏓᏁᏟᏴᏒ, ᎾᎢ ᏗᏕᏠᏴᏙᏗ ᏗᏥᏍᏙᏗᎢ, ᎾᎢ ᏱᏩᏥ ᎠᏍᏕᏍᏗ ᎠᏠ ᎠᎾᏕᎵᏍᎩ ᏗᏥᏅᏍᏙᏗᏱ, ᎠᏠ ᏄᏬ ᏅᏳᏍᏆᏍᏓ ᎾᏍᎩ ᎠᏰᎵ ᏚᏓᏂᏴᏛ ᏗᎪᏪᎵ ᎠᏟᎶᏍᏗ. ᎯᎠᏃ ᏕᏫᏒᏙᏗ ᏫᏃᎮᏍᎩ ᎾᏍᎩ ᏣᎳᎩ ᏧᏃᏴᎩ ᏚᏓᏁᏟᏴᏍᏔᏅ ᏧᎶᏒ ᎾᏍᎩ ᏥᏦᏁ ᏍᎪᎯᏧᏈ ᏧᏕᏘᏴᏌᏓ.

P.6.9: ChoCo: A multimodal corpus for the Choctaw language

Authors: Jacqueline Brixey

Country: United States

Abstract: ChoCo is a general use corpus for Choctaw, an American indigenous language (ISO 639-2: cho, endonym: Chahta). The corpus contains audio, video, and text resources, with many texts also translated in English. The Oklahoma Choctaw and the Mississippi Choctaw variants of the language are represented in the corpus. The data set provides documentation support for the threatened language, and allows researchers and language teachers access to a diverse collection of resources.

Native language: ChoCo yvt chahta anumpa, miliki asha anumpa, ma i kanomma ish ia hinla. ChoCo yvt na hakll, na holbatoba apisa micha tali holisso anumpa ishi, awant anumpa lawa ho na hullo tosholi. ChoCo yvt Okla Homma micha Mississippi chahta im anumpa alhpesa holissochi chika ahobachi. ChoCo yvt anumpa mosholi ma holisso apela atahli micha na hoyo micha anumpa ikhanauchi ma tali holisso lawa atahli.

P.6.10: Issues and challenges of NLP in relation to Canada's Aboriginal languages

Authors: Fatiha Sadat, Tan Ngoc Le and David Huggins Daines

Country: Canada

Abstract: Natural Language Processing is a multidisciplinary field that aims to create tools and linguistic resources for various applications. These resources include emotion and sentiment analysis, speech analysis, machine translation, information extraction, prediction tools, and more. Through this presentation, we would like to present the issues and challenges of the NLP to endangered and / or poorly endowed languages such as Aboriginal languages. Also, we would like to present reflections on a multi-disciplinary project involving Aboriginal languages and cultures of Canada to build linguistic resources for machine translation and for learning and teaching Aboriginal languages and cultures.

Native language: هو مجال متعدد التخصصات يشمل اللغويات وعلوم الكمبيوتر والعلوم المعرفية. ويهدف إلى إنشاء الأدوات (NLP) لغة المعالجة الطبيعية والموارد اللغوية لمختلف التطبيقات. تتضمن هذه الموارد تحليل العاطفة والمشاعر ، وتحليل الكلام ، والترجمة الآلية ، واستخراج المعلومات ، وأدوات التنبؤ ، وأكثر من ذلك. من خلال هذا العرض التقديمي ، نود أن نعرض قضايا وتحديات البرمجة اللغوية العصبية على اللغات المهددة بالانقراض و / أو ذات اللغات الضعيفة مثل لغات السكان الأصليين. ونود أيضًا تقديم أفكار حول مشروع متعدد التخصصات يتضمن لغات وثقافات السكان الأصليين في كندا لبناء موارد لغوية للترجمة الآلية وللتعلم وتعليم لغات وثقافات السكان الأصليين.