LT4All 2025

Advancing Humanism through Language Technologies

UNESCO Headquarters
24-26 February 2025

Book of Abstracts

LT4All is held in the framework of
the International Decade of Indigenous Languages 2022-2032
and in commemoration of
the Silver Jubilee of International Mother Language Day 2025

We present in this book of Abstracts, all the abstracts of the presentations taking place during the LT4All 2025 conference. We have asked each author to submit an abstract of 100 words, preferably in **English**, and in their **native language** or one of the languages they are working on, strongly advise to provide a version in a **minoritised language**.

However, the status and nature of certain languages means that producing summaries in these languages is not possible.

LT4All2025 Organizing committee

THEME: ACHIEVEMENTS

Session Setting the Scene:

sciencesconf.org:lt4all2025:616875

Internationalized Domain Names and Universal Acceptance for Digital Inclusion

Sarmad Hussain [1]

1 : Internet Corporation for Assigned Names and Numbers (ICANN)

The Domain Name System has evolved, with new top-level domains (TLDs) available in different languages and scripts (e.g., .LONDON, .PHOTOGRAPHY, .இலங்கை, .世界). Universal Acceptance (UA) of all valid domain names and email addresses, formed using these TLDs, by different software applications and systems is crucial for promoting multilingual internet. Achieving UA ensures everyone can navigate and communicate online using a domain name and email address aligning with their interests, business, and culture. Most online software applications and systems are not UA-ready, so all stakeholders need to work towards promoting UA adoption for a more inclusive Internet.

ڈومین ناموں کے آخرے حصے (TLDs) اب مختلف زبانوں اور خطوط میں دستیاب ہیں، جیسے .LONDON .PHOTOGRAPHY . 世界 .இலங்கை۔ان سے ڈومین ناموں کا نظام آگے بڑھا ہے۔ مختلف سافٹ ویئر ایپلی کیشنز اور سسٹمز میں ان TLDs کا استعمال کرتے ہوئے بنائے گئے تمام درست ڈومین ناموں اور ای میل پتوں کی عام قبولیت (UA)، انٹرنیٹ کو مختلف زبانوں میں فروغ دینے کے لیے بہت ضروری ہے۔ UA کا حصول یہ یقینی بناتا ہے کہ ہر کوئی اپنی دلچسپی، کاروبار اور ثقافت کے مطابق ڈومین نام اور ای میل پتے کے ساتھ آن لائن جا سکتا ہے اور بات چیت کر سکتا ہے۔ زیادہ تر آن لائن سافٹ ویئر ایپلی کیشنز اور سسٹمز ان کو استعمال نہیں کر سکتے۔ اس لیے تمام اسٹیک ہولڈرز کو زیادہ جامع انٹرنیٹ کے لیے UA کو فروغ دینے کی ضرورت ہے۔

Redefining Progress in the Margins: Lessons from Big Tech & Small Data

Christopher Morse [1]

1 : Zenter fir d'Lëtzebuerger Sprooch (Center for the Luxembourgish Language) (ZLS)

163 rue du Kiem L-8030 Luxembourg - Luxembourg

Progress in the development of contemporary language technologies has often relied on massive datasets that privilege widely spoken languages, overlooking the nuanced needs of smaller linguistic communities. Luxembourgish, a language traditionally embedded within a rich multilingual setting, provides a compelling counter-narrative. This presentation illustrates how a state administration redefines progress by prioritizing depth over scale in its mission to document the Luxembourgish language. In the absence of vast datasets, typically leveraged by large language models, the Centre for the Luxembourgish Language harnesses small, context-rich corpora, allowing for advancements in computational linguistics for Luxembourgish that are shaped by deep language expertise and cultural embeddedness. This underscores the essential role of cultural intelligence in refining and expanding the applications of language technologies.

Poster Session P1: Language Technology for Language Diversity

sciencesconf.org:lt4all2025:616964

AI Thinking as a Meaning-Centered Framework: Reimagining Language Technologies Through Community Agency

Quesada Jose F [1]

1 : Universidad de Sevilla (Univ. Seville)

Escuela Tecnica Superior de Ingeniería Informática (Universidad de Sevilla) Avda Reina Mercedes, s/n 41012 Sevilla (España) - Espagne

---

While language technologies have advanced significantly, current approaches fail to address the complex sociocultural dimensions of linguistic preservation. AI Thinking proposes a meaning-centered framework that would transform technological development from creating tools FOR communities to co-creating solutions WITH them. This approach recognizes that meaningful solutions emerge through the interplay of cultural understanding, community agency, and technological innovation. The proposal articulates a holistic methodology and a five-layer technological ecosystem where communities maintain control over their linguistic and cultural knowledge representation. This systematic integration of community needs, cultural preservation, and advanced capabilities could revolutionize how we approach linguistic diversity preservation in the digital age.

Aunque las tecnologías del lenguaje han avanzado significativamente, los enfoques actuales no abordan las complejas dimensiones culturales y sociales de la preservación lingüística. AI Thinking propone un marco centrado en el significado que transformaría el desarrollo tecnológico desde la creación de herramientas PARA las comunidades hacia la co-creación de soluciones CON ellas. Este enfoque innovador reconoce que las soluciones significativas emergen de la interacción entre comprensión cultural, participación comunitaria e innovación tecnológica. La propuesta articula una metodología holística y un ecosistema tecnológico donde las comunidades mantienen el control sobre la representación de su conocimiento, integrando sistemáticamente necesidades comunitarias, preservación cultural y capacidades avanzadas.

Digital Tools for Cherokee and the Role of a Morphological Analyzer in Language Empowerment

Vipasha Bansal [1], Wyman Kirk [2], Ryan Mackey [2]

1 : Translation Commons/University of Washington

2 : Cherokee Nation

Cherokee is an endangered Iroquoian language, spoken in Oklahoma and North Carolina (Montgomery-Anderson 2008). The Cherokee community is actively working on revitalization efforts, including speaker training programs at the Cherokee Language Institute (https://language.cherokee.org/). Cherokee is a polysynthetic language, and has a wide range of prefixes and suffixes. These affixes undergo a complex set of vocalic changes based on phonetic context. This can be particularly challenging for language learners. This project presents Morphy, a morphological analyzer that can both parse and generate Cherokee verbs. Morphy is part of a larger effort to build tools to support Cherokee language students.

Le cherokee est une langue iroquoienne en voie de disparition, parlée en Oklahoma et en Caroline du Nord (Montgomery-Anderson, 2008). La communauté Cherokee est très active dans les efforts de revitalisation, y compris des programmes de formation de locuteurs au Cherokee Language Institute (https://language.cherokee.org/). Le cherokee est une langue polysynthétique avec un large éventail d'affixes. Ces affixes subissent un ensemble complexe de changements vocaliques. Cette complexité présente des obstacles pour apprendre la langue. Ce projet présente Morphy, un analyseur morphologique capable à la fois d'analyser et de générer des verbes cherokee. Morphy fait partie d'un effort plus large visant à créer des outils pour soutenir les étudiants en langue cherokee.

Endurance and Adaptation: The Resilience of the Woirata Language in a Changing World

N Nazarudin **1, 2, ***

1 : Universitas Indonesia

2 : Universiteit Leiden = Leiden University

Leiden University | 2300 RA Leiden The Netherlands - Pays-Bas

* : Corresponding author

This paper examines the resilience of the Woirata community, whose language persists amid socio-cultural shifts. Drawing on ten years of fieldwork, it highlights their use of digital platforms, such as Facebook groups, to connect with the diaspora and share cultural knowledge. Intellectuals have formalized the language through a Woirata dictionary, while youth developed a machine translation system (hosted at https://www.yotowawa.com) to promote learning. These innovations reflect proactive strategies to sustain their language against modernization and globalization. The study reveals how tradition and adaptation intertwine in indigenous communities, emphasizing the role of cultural identity in linguistic endurance amidst change.

Artikel ini mengkaji ketahanan komunitas Woirata, yang bahasanya bertahan di tengah pergeseran sosio-budaya. Berdasarkan penelitian lapangan selama sepuluh tahun, artikel ini menyoroti penggunaan platform digital, seperti grup Facebook, untuk terhubung dengan diaspora dan berbagi pengetahuan budaya. Kaum intelektual telah memformalkan bahasa melalui kamus Woirata, sementara generasi muda mengembangkan sistem penerjemahan mesin (yang dapat diakses di https://www.yotowawa.com) untuk mendukung pembelajaran. Inovasi-inovasi ini mencerminkan strategi proaktif untuk mempertahankan bahasa mereka di tengah modernisasi dan globalisasi. Studi ini mengungkap bagaimana tradisi dan adaptasi saling terkait dalam komunitas adat, menekankan peran identitas budaya dalam ketahanan linguistik di tengah perubahan.

Globalisation: Colonising the Space of Flows?

Tunde Adegbola [1, *]

1 : African Languages Technology Initiative (Alt-i)

* : Corresponding author

Sociologist Manuel Castells conceptualised the space of flows, within which the production, transmission and processing of information flows takes place. In contrast, the space of places is a space of physical contiguity and physical exchanges. This dichotomy of spaces offers a valuable framework for engaging globalisation and colonisation. Dispossession of spaces of physical contiguity and material resources characterised colonisation, while globalisation is about the space of flows. Language is foundational to the space of flow.
This study engages colonisation and globalisation from a language technology point of view, towards gaining a level of understanding that can anticipate and prevent some of the avoidable pitfalls of globalisation.

Ọ̀mọ̀wé Manuel Castells pète àáyè ìṣàn fún aáyán àgbédìde, àtagbà ati àtòpọ̀ ìmọ̀. Àáyè alárìídìmú olójúlọ̀rọ́wà wáá jẹ́ bí àáyè ìdàkéjì sí àáyè ìṣàn.
Àwọn àáyè abèjì yì ṣe àǹfàní fún àgbékalè tó péye láti ṣe ìtọpinpin ìbámu tí ó wà láàrin ìsọ-ayé-di-alujára àti ìbónílẹ̀dulẹ̀. Ìbónílẹ̀dulẹ̀ jásí ìpàdánù àáye alárìídìmú, ṣùgbọ́n ìsọ-ayé-di-alujára wà ní ìlànà àáyè ìṣàn. Èdè sì ni ìpìlẹ̀ àáyè ìṣàn.
Àpilẹ̀kọ yìí fi ìkọbiarasí ọgbọ́n-àmúṣe ajẹmédè wo ìbónílẹ̀dulẹ̀ àti ìsọ-ayé-di-alujára, gégé bi atọ́nà àròjinlẹ̀, igi gogoro máà gún mi lójú, fún ìdẹ́kun àwọn àbámọ̀ tí ó lè jáde wá láti inú ìsọ-ayé-di-alujára.

Promoting Equity for Non-Dominant Languages: Enhancing Metadata in Multilingual Digital Libraries

Amel Fraisse [1,*], Znaidi Nassereddine [2], Arjun Sanyal [3,*], Laurence Favier [4,*]

1 : Fraisse

Univ. Lille EA 4073 - GERiiCO

2 : Znaidi

Univ. Lille EA 4073 - GERiiCO

3 : Sanyal

4 : Favier

Univ. Lille EA 4073 - GERiiCO

* : Corresponding author

This paper examines challenges in managing multilingual bibliographic records in digital libraries, focusing on the under-representation of minority languages and metadata gaps in translated works. Using *The Adventures of Huckleberry Finn* as a case study, it highlights disparities in metadata quality and cultural representation. The study proposes solutions based on collaborative knowledge models, such as the ROSETTA project, which integrates multilingual corpora and data visualization tools. By leveraging crowdsourcing, these approaches enhance metadata completeness and accessibility for low-resource languages, promoting cultural diversity and equitable knowledge access in multilingual repositories. The paper discusses implications for language technology in bridging these gaps.

यह शोधपत्र डिजिटल पुस्तकालयों में बहुभाषी ग्रंथसूची अभिलेखों के प्रबंधन में चुनौतियों की जांच करता है, जो अनुवादित कार्यों में अल्पसंख्यक भाषाओं के कम प्रतिनिधित्व और मेटाडेटा अंतराल पर ध्यान केंद्रित करता है। द एडवेंचर्स ऑफ हकलबेरी फिन को केस स्टडी के रूप में उपयोग करते हुए, यह मेटाडेटा गुणवत्ता और सांस्कृतिक प्रतिनिधित्व में असमानताओं को उजागर करता है। अध्ययन सहयोगी ज्ञान मॉडल, जैसे कि ROSETTA परियोजना, जो बहुभाषी कॉर्पोरा और डेटा विज़ुअलाइज़ेशन टूल को एकीकृत करता है, पर आधारित समाधान प्रस्तावित करता है। क्राउडसोर्सिंग का लाभ उठाकर, ये दृष्टिकोण कम संसाधन वाली भाषाओं के लिए मेटाडेटा पूर्णता और पहुंच को बढ़ाते हैं, बहुभाषी रिपॉजिटरी में सांस्कृतिक विविधता और समान ज्ञान पहुंच को बढ़ावा देते हैं। शोधपत्र इन अंतरालों को पाटने में भाषा प्रौद्योगिकी के निहितार्थों पर चर्चा करता है।

---

sciencesconf.org:lt4all2025:617280

Sama-sama: Bringing Southeast Asia together to build better LLMs

Jann Railey Montalan [1, 2], Jian Gang Ngui [1, 2], Yosephine Susanto [1, 2], William Chandra Tjhi Tjhi [1, 2]

1 : AI Singapore (AISG)

AI Singapore innovation 4.0 3 Research Link #02-04 Singapore 117602 - Singapour

2 : National University of Singapore (NUS)

21 Lower Kent Ridge Rd, Singapore 119077 - Singapour

Southeast Asia (SEA) is linguistically and culturally diverse, making language technologies (LT) development using Large Language Models (LLMs) for the region a complex challenge. Despite having 700 million people, there are severe gaps in digital representation, inadequacies in out-of-the-box AI solutions, and scarcity in culturally-specific evaluations. As such, AI Singapore developed the SEA-LION family of open-source LLMs for SEA. We engage with academia, government, and industry under Project SEALD to unlock previously-closed SEA data. We built SEA-HELM, a holistic, multilingual, and multicultural LLM evaluation suite, and strengthened community-driven AI research, evaluations, and applications (e.g. SEACrowd, GlobalMMLU, Sahabat-AI).

Mayaman sa wika at kultura ang Timog-silangang Asya (SEA), kaya malaking hamon ang paggawa ng mga language technology (LT) gamit ang mga Large Language Model (LLM) para sa rehiyon. Bagamat mayroon itong 700 milyong katao, mayroon itong malalaking kakulangan sa digital na representation, sa mga AI solution, at sa mga kultural na pagsusuri.
Binuo ng AI Singapore ang SEA-LION, mga open-source na LLM para sa SEA. Nakikipagtulungan kami kasama ang akademya, gobyerno, at industriya sa ilalim ng Project SEALD upang buksan ang dating saradong datos ng SEA. Binuo din namin ang SEA-HELM, isang holistiko, multilingguwal, at multikultural na pagsusuri ng mga LLM, at pinalakas namin ang mga pananaliksik sa AI, mga pagsusuri, at mga aplikasyan ng iba't ibang komunidad sa SEA (tulad ng SEACrowd, GlobalMMLU, Sahabat-AI).

---

sciencesconf.org:lt4all2025:617116

Toward Human-bound Language Technology, Autonomy, Equality, and Diversity

Ilan Kernerman [1], Kilim Nam [2]

1 : Lexicala by K Dictionaries (Lexicala)

Alumot 10A, Nitsane Oz 4283600 - Israël

2 : Yonsei University

Seoul - Corée du Sud

We aim to harness advanced technology to develop human-driven language resources, tools, applications and skills, aided by industry, academia and organizations worldwide. These efforts will enhance long-term preservation and prosperity of indigenous languages, by designing data structures that reflect their nuances and interoperate with multilingual systems. All languages will be researched and mapped minutely and systemically, linking cross-lingually to create massive data clouds/spaces. Each language set will constitute a Gold Standard for generating synthetic data to train and fine-tune Large/Small/Specialized/Multilingual Language Models. We will apply international regulations and assure open and fair science and access that empower low-resource languages and decrease language discrepancy for all humans.

본 연구는 첨단 기술을 활용하여 인간 중심의 언어 자원, 도구, 응용 프로그램 및 기술 역량을 개발하는 방안을 제시하는 것을 목표로 한다. 이러한 과정은 산업계, 학계, 국제 기관과의 협력에 기반하며, 저자원언어의 장기적 보존과 번영을 도모하는 데 기여할 것이다. 특히, 모국어화자가 가진 각 언어의 고유한 구조와 뉘앙스를 반영하고 다국어 시스템과의 상호운용성을 확보하기 위한 데이터 구조 설계를 중심으로 논의한다. 모든 언어는 정밀하고 체계적인 연구 및 매핑 과정을 거치며, 다국어 간 연계 기술을 통해 대규모 데이터 클라우드 및 데이터 공간에 구축되어야 할 것이다. 각 언어 세트는 합성 데이터 생성을 위한 골드 스탠더드(Gold Standard)로 작용하며, 대형/소형/전문 분야/다국어 언어 모델의 훈련 및 미세 조정에 활용될 예정이다. 또한, 본 연구는 국제 규정을 준수하고 개방적이고 공정한 과학 연구와 접근성을 보장함으로써, 저자원언어의 역량 강화와 언어 불평등 완화를 추구한다. 이는 모든 언어 사용자가 동등하게 언어 자원에 접근할 수 있는 환경을 조성하기 위한 지속적인 노력의 일환으로 진행된다.

---

sciencesconf.org:lt4all2025:618022

UniDive: A COST Action on Universality, Diversity and Idiosyncrasy in Language Technology

Agata Savary [1,*], Daniel Zeman [2], Verginica Barbu Mititelu [3], Anabela Barreiro [4], Olesea Caftanatov, Marie-Catherine De Marneffe [5], Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli [6], Bruno Guillaume [7], Olha Kanishcheva, Stella Markantonatou, Nurit Melnik [8], Joakim Nivre, Atul Kumar Ojha [9], Carlos Ramisch, Beata Wójtowicz, Alina Wróblewska

1 : Laboratoire Interdisciplinaire des Sciences du Numérique (LISN)

Université Paris-Saclay CNRS

2 : ÚFAL MFF, Charles University,

3 : Romanian Academy Research Institute for Artificial Intelligence

4 : INESC-ID Lisboa

5 : UCLouvain (ILC)

6 : ATHENA - Research and Innovation Center in Information, Communication and Knowledge Technologies

7 : Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

L'Institut National de Recherche en Informatique et e n Automatique (INRIA), CNRS : UMR7503, Univesité de Lorraine

8 : The Open University of Israel (OUI)

9 : University of Galway

* : Corresponding author

In this paper, we present the objectives, organisation, and activities of the UniDive COST Action, a scientific network dedicated to universality, diversity, and idiosyncrasy in language technology. We describe the aims and structure of this initiative, the individuals involved, the working groups, and the ongoing tasks and activities. We also report the extension of UD treebanks and MWE datasets, especially for low-resource and/or endangered languages. Finally, this paper also serves as an open call for participation from new members and countries to join the UniDive Cost action.

Sa pháipéar seo, cuirimid i láthair cuspóirí, eagrúchán agus gníomhaíochtaí Gníomh COST UniDive, líonra eolaíoch atá tiomnaithe don uilíocht, don éagsúlacht, agus don leithleas i dteicneolaíocht teanga.Déanaimid cur síos ar aidhmeanna agus ar struchtúr an tionscnaimh seo, ar na daoine aonair atá i gceist, ar na grúpaí oibre, agus ar na tascanna agus gníomhaíochtaí leanúnacha. Tuairiscímid freisin go bhfuil síneadh curtha le bainc crann UD agus tacair sonraí MWE, go háirithe do theangacha íseal-acmhainne agus/nó i mbaol. Ar deireadh, feidhmíonn an páipéar seo mar ghlao oscailte freisin ar rannpháirtíocht ó chomhaltaí agus ó thíortha nua chun páirt a ghlacadh i ngníomh Costas UniDive.

---

sciencesconf.org:lt4all2025:615879

Unified Language Interface to address Linguistic Diversity through technology for Indigenous Digital Inclusion in The Context of Viksit Bharat @ 2047

Kotthireddy Mallareddy [1, *]

1 : Associate Professor of Telugu Government Degree College Huzurabad (GDC)

Huzurabad Karimnagar Dist. Telangana, India PIN: 505468 - Inde

* : Corresponding author

India has 121 major languages, 270 mother tongues and 373 classified mother tongues and 1474 unclassified mother tongues. In these critical multilingual circumstances Digital Inclusion is crucial for inclusive equitable quality education to reach transformative AGENDA of Vision India Viksit Bharat 2047. India must harness its demographic dividend by integrating regional languages into AI models to address the digital gaps and to bridge the language barriers. The proposed study aims to promote Indigenous digital inclusion by creating accessible digital platforms in Indigenous languages, facilitating online communications and collaborating with Indigenous communities. India's Unified Payment Interface (UPI) revolutionized the digital payment ecosystem, similarly efforts must be made to create Unified Language Interface (ULI) with new AI models for the inclusive development of Indigenous and regional languages.

121 ప్రధాన భాషలు, 270 మాతృభాషలు ఉన్నాయి మరియు ఒక్కొక్కటి 10,000 కంటే తక్కువ జనాభాన కలిగిన 373 వర్గీకృత మాతృభాషలు, 1474 వర్గీకరించని మాతృభాషలు ఉన్నాయి. ఈలాంటి బహుభాషా పరిస్థితులలో వికశిత్ భారత్ 2047 సురక్షితమైన డిజిటల్ ఇన్‌క్లూజన్, డిజిటల్ లెక్కాలజీ ఎంతో కీలకం. డిజిటల్ అంతరాలను పరిష్కరించడానికి, భాషా అడ్డంకులను తగ్గించడానికి కృత్రిమ మేధా రూపాలలో ప్రాంతీయ భాషలను సమగ్రపరచడం ద్వారా భారతదేశం జనాభా వైవిద్యాన్ని సద్వినియోగపరుచుకోవాలి. ప్రతిపాదిత అధ్యయనం డిజిటల్ ప్లాట్‌ఫారమ్‌లను సృష్టించడం, ఆన్‌లైన్ కమ్యూనికేషన్‌లను సులభతరం చేయడం, అనువాదం కోసం కృత్రిమ మేధాని ఉపయోగించడం మరియు స్వదేశీ కమ్యూనిటీలతో సహకరించడం ద్వారా దేశీయ డిజిటల్ చేరికను ప్రోత్సహించడం లక్ష్యంగా పెట్టుకుంది. భారతదేశ యూనిఫైడ్ పేమెంట్ ఇంటర్ ఫేస్(యు.పి.ఐ)డిజిటల్ చెల్లింపు వ్యవస్థను విప్లవాత్మకంగా మార్చింది, అదేవిధంగా దేశీయ మరియు ప్రాంతీయ భాషల సమగ్ర అభివృద్ధికి కొత్త కృత్రిమ మేధా నమూనాలతో ఏకీకృత భాషా ఇంటర్ ఫేస్(యు.ఎల్.ఐ) రూపకల్పనకు కృషి చేయాలి.

Yezh Ar Vro - the language of the land: community-driven data collection for speech recognition

Loïc Grobol [1,*], Alice Millour, Mélanie Jouitteau, Jean-Yves Antoine

1 : Université Paris Nanterre - Département Sciences du Langage

Université Paris Nanterre

* : Corresponding author

En contexte de technologisation accélérée des rapports humains, les langues pour lesquelles les nouveaux outils ne peuvent être déployés risquent des baisses de pratique potentiellement fatales. Construire des ressources numériques utilisables en TAL devient une tâche essentielle de préservation de la diversité linguistique humaine. Il existe des solutions logicielles pour l'acquisition de données, mais elles rencontrent une appropriation insuffisante par les communautés parlantes. Nous présentons un projet pilote d'envergure déjà significative visant à valider l'hypothèse de recherche qu'une collaboration interdisciplinaire précoce avec les communautés parlantes pour la conception des outils d'acquisition des données augmente significativement leur appropriation et donc leur efficacité. Concrètement, le projet YAR [Yezh Ar vRo - la langue du pays] propose de développer deux outils d'acquisition de données:
- Une application de collecte de parole géolocalisée
- Une plateforme de transcription participative
Dans un premier temps, nous rassemblons des corpus oraux préexistants, et les enrichissons en métadonnées, dont la géolocalisation. Nous développons l'application mobile de collecte de son géolocalisé. Dans un second temps, nous outillons la transcription des corpus oraux. Nous transcrivons un fond d'amorçage pour pré-peupler la carte en données de proximité. Nous développons la plateforme de transcription participative. Les outils pédagogiques dont l'objet est la transcription sont alimentés par la collecte de données et fournissent des corpus alignés qui constituent la ressource TAL. Nous explorons une solution de pré-transcription assistée par reconnaissance vocale automatique.

Session O1: LT Advancements in Enabling Linguistic Diversity and Multilingualism

Linguistic Diversity and Multilingualism for Humans and AI

A. Seza Doğruöz [1]

1 : Universiteit Gent

Millions of speakers and users (e.g., Europe, Africa, Southeast Asia) around the world speak more than one language/dialect in daily communication and they are multilingual. There is extensive research on multilingual communication at the individual and societal levels for different language groups and locations. Language technologies are generally built with monolingual assumptions. However, there is still a need for more research to understand how current AI systems are built (e.g., data types, collection and processing procedures) and to what extent they address the needs and preferences of multilingual users across languages/dialects.

Bridging Scripts and Standards: The work of the Script Encoding Initiative

Anushah Hossain [1],

1 : University of California [Berkeley] (UC Berkeley)

Berkeley, CA - États-Unis

The **Script Encoding Initiative (SEI)** at the University of California, Berkeley has been instrumental in ensuring that the world's diverse writing systems are represented in Unicode, contributing to the successful encoding of over **120 of the 168** scripts currently included. As an intermediary between linguistic experts, community members, and international standards bodies, SEI bridges technical and cultural gaps in script digitization. This talk will highlight SEI's ongoing efforts to encode more scripts, enhance public understanding of script digitization, and explore new methods for digitally supporting writing systems.

La **Script Encoding Initiative (SEI)** de l'Université de Californie à Berkeley joue un rôle essentiel dans la représentation des systèmes d'écriture du monde au sein d'Unicode, ayant contribué à l'encodage de plus de **120 des 168** écritures actuellement incluses. En tant qu'intermédiaire entre experts linguistiques, membres des communautés et organismes internationaux de normalisation, la SEI comble les écarts techniques et culturels liés à la numérisation des écritures. Cette présentation mettra en lumière les efforts continus de la SEI pour encoder davantage d'écritures, améliorer la compréhension publique de la numérisation des systèmes d'écriture et explorer de nouvelles approches pour leur soutien numérique.

From Media Practitioner to Guardian of Language Resources: My Commitment to Protecting and Promoting Linguistic Diversity' 中文版: 标题：从传播者到守护者——我的语保情怀

Han Wang [1]

1 : Hunan Broadcasting System Media Group

Hunan Broadcasting System Media Group, Changsha City, Hunan Province, China - Chine

---

Abstract: Drawing on his media background, Wang Han, who has been serving as an advisor to the Center for the Protection and Research of Language Resources of China, will share his journey in safeguarding linguistic diversity in the face of modernization and the declining use of regional language resources in China. He highlights the efforts in producing TV programs aimed at promoting China's linguistic diversity, initiating the Regional Language Resources Survey "Hsiang Accent Project", and establishing a comprehensive language resources database in Hunan province of China. Emphasizing the importance of linguistic diversity, he outlines future plans, such as to call greater public and academic engagement in language protection, as well as to increase public awareness to promote the linguistic diversity and multilingualism via cyberspace for younger generation.

提要：凭借出色的媒体背景，汪涵先生将分享他在现代化进程中保护语言多样性以及保护中国语言资源的个人工作经历。作为中国语言资源保护与研究中心顾问，他将重点介绍他在主持以中国区域性语言为主题的电视类节目、发起并由他资助的语言资源保护公益项目-“響應”计划”，即利用现代化科技手段建立湖南省区域性语言资源有声数据库等其他方面的努力。他强调语言资源保护的重要性，并概述了未来的计划，其中包括呼吁更多社会公众参与语言保护工作，以及通过互联网空间提升公众意识，促进年轻一代的语言多样性和多语言在网络空间的使用。

Towards linguistically and culturally inclusive LLMs

Partha Pratim Talukdar [1]

1 : Google DeepMind and IISc Bangalore

(English) Large Language Models (LLMs) have seen tremendous progress over the last few years with increasing adoption across the globe. In addition to languages, LLMs need to be knowledgeable about cultural nuances and local norms. These raise interesting research challenges in language and culture which are also the focus of our research in the area of inclusive LLMs. I shall talk about Project Vaani where the goal is to capture the speech landscape of India using a unique geo-anchored approach, the CUBE benchmark to evaluate cultural knowledge of LLMs, and the SMOL dataset of professional translation across 115 very low data-resource languages.

(Assamese) যোৱা কেইবছৰমানৰ পৰা LLMৰ মানদণ্ড উন্নত হৈছে আৰু সমগ্ৰ বিশ্বতে এই মডেলবোৰৰ ব্যৱহাৰ ক্ৰমান্বয়ে বৃদ্ধি পাইছে। কেৱল ভাষা বুজি পোৱাৰ উপৰিও, LLMবোৰে বিভিন্ন সংস্কৃতি আৰু স্থানীয় ৰীতি-নীতিও জানিব লাগে। এইটো এটা কঠিন সমস্যা। এই বক্তৃতাত মই এই ক্ষেত্ৰত আমাৰ গৱেষণাৰ বিষয়ে আলোচনা কৰিম। মই প্ৰজেক্ট ভানিৰ বিষয়ে ক'ম যিয়ে ভাৰতত কথিত ভাষাৰ বৈচিত্ৰ্য সংগ্ৰহ কৰাৰ লক্ষ্য ৰাখিছে, আৰু কেনেকৈ CUBE বেঞ্চমাৰ্কে LLMএ বিভিন্ন সংস্কৃতি বুজি পায় নেকি পৰীক্ষা কৰে।মই SMOL ডাটাছেটৰ বিষয়েও কম যিটো ১১৫টা অতি কম ডাটা-সম্পদ ভাষাৰ অনুবাদেৰে গঠিত।

Towards a truly universal translator?

Alessandro Fusacchia [1]

1 : Translated

Via Indonesia 23, 00144 Rome, Italy - Italie

Linguistic barriers are not overcome through homogenization or lingua francas, but by ensuring that everyone can understand and be understood in their own language. Technology is getting close to achieving this goal (the "universal translator"), although it will arrive for some languages sooner than for others. To prevent this from increasing inequalities, we need to understand: how translation technologies can support all languages, including minority languages, especially in this new era of generative AI; and what needs, old and new, of people and society we will be able to address thanks to the availability of fast, extremely affordable, and high-quality translation.

Verso un traduttore veramente universale?
Le barriere linguistiche non si superano con l'omologazione o lingue franche, ma facendo in modo che ognuno possa capire e farsi capire nella propria lingua. La tecnologia è vicina a farci raggiungere questo punto (il “traduttore universale”) anche se arriverà per alcune lingue prima che per altre. Per evitare che questo porti all'aumento delle disuguaglianze, dobbiamo capire: come le tecnologie applicate alle traduzioni possano aiutare tutte le lingue, comprese le lingue minoritarie, a maggior ragione in questa nuova era di AI generativa; e quali bisogni vecchi e nuovi delle persone e della società potremo affrontare grazie alla disponibilità di una traduzione veloce, estremamente economica e di grande qualità.

Poster Session P2: Language Technology for Education, Inclusion, Innovation

Bridging the digital divide: Where do we stand?

Daniel Wilson [1,*], Matthew Hernandez [2,*], Moises Coronel [2,*], Jennifer Haliewicz [2,*], Maxim Kulik [2,*], Vanessa Nguyen [2,*], Shawna Birnbaum [2,*], Ki Woong Moon [2,*], Haley Logan [2,*]

1 : XRI Global

2 : University of Arizona

* : Corresponding author

As we enter the decade of indigenous languages, it is important to track how much progress is being made in bridging the digital divide. To this end XRI Global and students from the University of Arizona have joined efforts to take inventory of the current data and models that exist for the world's low-resource languages. By cataloging all the data and models on huggingface, GitHub, Common Voice, and other well-known hubs for training data and AI models, our team has produced a map which shows the current level of support for a large number of low-resource languages.
À l'aube de la décennie des langues autochtones, il est important de suivre les progrès réalisés pour combler la fracture numérique. À cette fin, XRI Global et des étudiants de l'Université d'Arizona ont uni leurs efforts pour dresser l'inventaire des données et des modèles actuels qui existent pour les langues à faibles ressources du monde. En cataloguant toutes les données et tous les modèles sur huggingface, GitHub, Common Voice et d'autres plateformes bien connues pour les données de formation et les modèles d'IA, notre équipe a produit une carte qui montre le niveau actuel de prise en charge d'un grand nombre de langues à faibles ressources.

---

sciencesconf.org:lt4all2025:616308

Can machine translation really help minority languages in Europe? An analysis with value scenarios

Sergi Alvarez-Vidal [1]

1 : Universitat Autònoma de Barcelona (UAB)

Machine translation (MT) has improved significantly over the past decade, becoming widely accessible through neural MT (NMT) and large language models (LLMs) like GPT. However, most MT systems are English-centric, performing well only for languages with abundant data. Minority languages face challenges beyond data scarcity, including systemic differences in language communities. We explore the impact of MT on Catalan and Karelian using value scenarios to assess potential harms and challenges. We aim to highlight key issues in MT for minority languages and propose guidelines for future research and applications to enhance their usability and fairness.

La traducció automàtica (TA) ha millorat significativament durant l'última dècada i ara és fàcilment accessible gràcies a la TA neuronal (TAN) i als models massius de llenguatge (MML) com GPT. Tanmateix, la majoria de sistemes de TA són anglocèntrics i només ofereixen bons resultats per a llengües amb grans volums de dades. Les llengües minoritàries afronten reptes més enllà de l'escassetat de dades, incloent-hi diferències sistèmiques entre comunitats lingüístiques. Analitzem l'impacte de la TA en el català i el carelià mitjançant escenaris de valor per avaluar desafiaments i riscos. L'objectiu és identificar problemes clau i proposar directrius per a futures recerques i aplicacions que millorin la usabilitat i l'equitat.

---

sciencesconf.org:lt4all2025:620629

Challenges in Lemmatization: what still remains to be solved?

Olia Toporkov [1], Rodrigo Agerri

1 : HiTZ Center - Ixa, University of the Basque Country UPV/EHU (HiTZ-Ixa, UPV/EHU)

Lemmatization is a basic NLP task that consists of producing, from a given inflected word, its canonical form or lemma, and is especially challenging for languages with rich inflection. We empirically investigate, how crucial is the role of morphology in contextual lemmatization, as well as which approach in obtaining the minimum edit distance between the word and its lemma is more convenient. Along with in-domain evaluation we evaluate lemmatizers in out-of-domain setting. The results of our study show that providing lemmatizers with fine-grained morphological information during training is not that beneficial, not even for agglutinative languages. We also identify the features that contribute to the best approach for obtaining the minimum amount of edits between the word and its lemma. Lastly, we demonstrate that current evaluation practices for lemmatization are not adequate to clearly discriminate between models.

Lemmatizazioa oinarrizko NLP ataza bat da, hitz flexionatua jakin batetik abiatuta bere forma kanonikoa edo lemma sortzea, eta bereziki erronka da inflexio aberatsa duten hizkuntzentzat. Enpirikoki ikertzen dugu, zein erabakigarria den morfologiaren papera testuinguruko lematizazioan, baita zein hurbilketa den komenigarriagoa hitzaren eta lemmaren arteko edizio-distantzia minimoa lortzeko. Domeinuaren ebaluazioarekin batera, domeinutik kanpoko konfigurazioko lematizatzaileak ebaluatzen ditugu. Gure ikerketaren emaitzek erakusten dute lemmatizatzaileei morfologia ematea ez dela hain onuragarria, ezta hizkuntza aglutinatzaileentzat ere. Halaber, jakin dugu hitzaren eta bere lemaren arteko edizio-distantzia minimoa lortzeko hurbilketarik onena zein ezaugarrik egiten duen ere jakiten dugu. Azkenik, frogatzen dugu lematizaziorako egungo ebaluazio-praktikak ez direla egokiak ereduak argi eta garbi bereizteko.

sciencesconf.org:lt4all2025:618427

Cross-Language Crossword Generation: Harnessing Large Language Models for Multilingual Educational Tools

Kamyar Zeinalipour [1]

1 : Università degli Studi di Siena (UNISI)

Via Roma, 56, 53100 Siena - Italie

This framework uses advanced LLMs (GPT-4, GPT-3.5, GPT-3, Mistral-7B, Llama3-8B) to automatically generate educational crosswords in multiple languages, supporting memory, vocabulary, and problem-solving. By compiling multilingual clue–answer datasets (English, Arabic, Turkish, Italian), we produce text-specific or keyword-based clues. Fine-tuned or zero/few-shot approaches yield context-rich puzzles that enhance active, student-focused learning. This pioneering method merges gamification with interactive instruction, illustrating how LLM-driven solutions can reshape traditional education.

آموزشی متقاطع کلمات هایجدول (GPT-4، GPT-3.5، GPT-3، Mistral-7B، Llama3-8B) پیشرفته زبانی هایمدل از گیریبهره با پژوهش دراین پاسخ–سرنخ چندزبانه هایداده گردآوری با .سازدمی تقویت را مسئله حل و واژگان ،حافظه و اند شده تولید خودکار صورتبه مختلف هایزبان در را محوریبافت معماهای ،کم نمونه با/بدون یا زبانی های مدل تربیت با .کنیممی ایجاد کلیدواژه یا متن با مرتبط هایسرنخ ،(ایتالیایی ،ترکی ،عربی ،انگلیسی) راهکارهای چگونه دهدمی نشان ،تعاملی آموزش و سازیبازی تلفیق با نوآور روش این .دهندمی ارتقا را آموزمحوردانش و فعال یادگیری که آورندمی پدید را .کنند متحول را سنتی تعلیم توانندمی زبانی هایمدل بر مبتنی

sciencesconf.org:lt4all2025:618578

Developing Reading Technologies for Filipino Language

Rhandley Cajote [1], Michael Gringo Angelo Bayona [2]

1 : University of the Philippines Diliman (UP Diliman)

Quezon City, 1101 Philippines - Philippines

2 : Trinity College Dublin

College Green, Dublin 2, Ireland - Irlande

According to a study conducted by the Program for International Student Assessment (PISA) in 2022 out of the 81 countries, the Philippines ranks 76^{th} and among the bottom ten (10) in reading comprehension, mathematics and science. To address this, our proposed solution is to develop a computer-based reading tutor called “Tanglaw” that will help the teachers in teaching children how to read, reducing the need for a one-to-one reading session with the students. Tanglaw uses an automatic speech recognition system in the Filipino language specifically trained to recognize children's speech. A miscue detection system is also implemented to detect reading miscues.

Ayon sa pag aaral na isinagawa ng Program for International Student Assessment (PISA) noong 2022 mula sa 81 bansa, ang Pilipinas ay nasa ika-76 at kabilang sa sampu (10) mula sa ilalim sa pagbasa, matematika at agham. Upang matugunan ito, ang aming panukalang solusyon ay ang pagbuo ng “reading tutor” na nakabatay sa computer at tinatawag na "Tanglaw" na makakatulong sa mga guro sa pagtuturo sa mga bata kung paano magbasa. Gamit ang Tanglaw ay makakabawas ang pangangailangan ng indibidwal na sesyon ng pagbabasa sa mga mag-aaral. Gumagamit si Tanglaw ng automatic speech recognition sa wikang Filipino na partikular sa boses ng mga bata. Ipinatupad din ang miscue detection system upang matukoy ang mga kamalian sa pagbabasa.

---

sciencesconf.org:lt4all2025:617483

Empowering mother tongue education with AI: EdTech solutions for Marma in Bangladesh

Alp Öktem [1,*], Milena Haykowska [1,*], Aong Marma [1], Rani Marma [2], Ushing Marma, Oyessorzo Chowdhury, Chingthowiu Marma, Xiomara Hurni-Cranston, Kuri Chisim [3], Jason Symons [3], Lisa Reiners [3], Amélie Tremelo-Thomas [1], Alyssa Boularés [1], Aimee Ansari [1]

1 : CLEAR Global (CLEAR)

2 : Massachusetts Institute of Technology (MIT)

77 Massachusetts Ave, Cambridge, MA 02139 - États-Unis

3 : CLEAR Global (CLEAR)

* : Corresponding author

CLEAR Global, in partnership with a2i and funded by DFAT, is developing AI-driven tools to support mother tongue education for Marma speakers in Bangladesh's Chittagong Hill Tracts. Our baseline research and user studies highlighted key challenges: low Marma literacy among teachers, limited digital resources, and disparities in access to resources between rural and urban schools. Engaging teachers, community leaders, linguists and education experts, we explored AI solutions, including text-to-speech technology, to enhance Marma mother tongue learning. User testing with Marma teachers showed strong support for interactive features such as writing and pronunciation exercises, gamification, and text-to-speech. However, accessibility challenges—including limited internet, digital literacy gaps, and the need for offline functionality—highlighted areas for further refinement. Future work will focus on improving AI models, expanding teacher training, and integrating the tool with existing educational resources to ensure long-term impact.

CLEAR Global, en partenariat avec a2i et financé par le DFAT, développe des outils d'IA pour soutenir l'enseignement en langue maternelle des locuteurs de Marma au Bangladesh. Nos recherches ont montré des défis majeurs : faible maîtrise du Marma chez les enseignants, ressources numériques limitées et disparités d'accès aux ressources entre écoles rurales et urbaines. Avec des enseignants, responsables communautaires, linguistes et experts en éducation, nous avons exploré des solutions d'IA, dont la synthèse vocale, pour améliorer l'apprentissage. Les enseignants ont plébiscité les exercices interactifs et la ludification, mais des défis subsistent : accès limité à Internet, culture numérique insuffisante et nécessité de fonctionner hors ligne. Les travaux futurs porteront sur l'amélioration des modèles d'IA, la formation des enseignants et l'intégration aux ressources pédagogiques existantes afin de garantir un impact durable.

---

sciencesconf.org:lt4all2025:616216

Enhancing Well-being through Language Technology Innovations

Ethel Ong [1], Jackylyn Beredo [2], Elaine Marie Aranda [2], Joel Navarez [2]

1 : De La Salle University

2401 Taft Avenue, Manila - Philippines

2 : De La Salle University

2401 Taft Avenue, Manila - Philippines

Rising cases of psychological distress are evident among the youth who experience developmental changes and social challenges. Studies reported the benefits of technology in treating mental health illnesses but these assume the presence of disorders and are primarily concerned with diagnosis and treatment of mental health conditions. Stories form the core of every human conversation where sharing of daily events may shed insights on life stressors and actions that show resilience. We leverage conversational interfaces anchored on stories to help the youth practice self-care and well-being management where they can feel comfortable in disclosing their goals, concerns, thoughts and challenges.

Tumataas ang mga kaso ng sikolohikal na pagkabalisa sa kabataan na nakakaranas ng maramaing pagbabago at hamon sa lipunan. Iniulat ng mga pag-aaral ang benepisyo ng teknolohiya sa pagpapagamot ng mga sakit sa kalusugan ng isip ngunit ipinapalagay ng mga ito ang pagkakaroon ng karamdaman at paggamot sa kondisyon na ito. Ang kuwento ang bumubuo sa ubod ng bawat pag-uusap ng tao kung saan ang pagbabahagi ng mga pang-araw-araw na kaganapan ay maaaring magbigay linaw sa mga pangyayaring nakaaabala sa ating isipan at mga aksyon na nagpapakita ng katatagan. Ginamit namin ang chatbot na naka-angkla sa mga kuwento upang matulungan ang mga kabataan na magsanay ng pangangalaga sa sarili at pamamahala sa kapakanan kung saan maaari silang maging komportable sa paglalahad ng kanilang mga layunin, alalahanin, iniisip at hamon.

---

sciencesconf.org:lt4all2025:617652

From Technology to Application - AI Empowering Barrier-Free Communication in Native Languages

Chen Cheng [1,*], Wang Yihan [1]

1 : iFLYTEK Co.,Ltd

* : Corresponding author

From Technology to Application: Advancing AI for Barrier-Free Communication in Native Languages This report focuses on the cutting-edge development of iFLYTEK's AI speech and language technologies, providing an in-depth analysis of its core technologies and product ecosystem. By exploring innovative applications of technology across various domains such as daily life and workplace scenarios,demonstrate how AI enables seamless communication in multiple languages and contexts, supporting the preservation and exchange of native languages while fostering sustainable development in a multicultural world.

Portable Articulatory Measurements Using Optopalatography

João Menezes [1], Arne-Lukas Fietkau [1], Jihyeon Yun [1], Peter Birkholz [1]

1 : Technische Universität Dresden = Dresden University of Technology (TU Dresden)

TU Dresden 01062 Dresden - Allemagne

Optopalatography (OPG) is a sensing technique generally used for measuring speech articulation, consisting of optical sensors positioned on the palate of the speaker aiming to measure the position of articulators such as the tongue.
This work presents a OPG device with 15 sensors to measure tongue and lip position with a 100 Hz sampling frequency.
Its main features are: 1) its recording setup, portable enough to serve for field recordings, and 2) the possibility of a personalized version, mounted on an artificial palate, and a non-personalized version, which can be reused for various speakers.

Optopalatografia (OPG) é uma técnica de medição geralmente utilizada para a articulação da fala, composta por sensores ópticos posicionados no palato da(o) falante com o intuito de medir a posição de articuladores da fala, como a língua.
Esse trabalho apresenta um dispositivo OPG com 15 sensores para medir a posição da língua e dos lábios com uma frequência de amostragem de 100 Hz.
Suas características principais são: 1) sua configuração para gravação, suficientemente portátil para ser usada em campo, e 2) a possibilidade de uma versão personalizada, montada num palato artificial, e uma versão não-personalizada, que pode ser reutilizada para várias(os) falantes.

---

sciencesconf.org:lt4all2025:620464

Sign Language Technology and Digital Deaf Studies: Sign Languages, vitalization and technology: a cross-continental perspective

Nargess Asghari [1], Victoria Nyst [1,2], Leslie Okyere *, Peter Van Der Putten [3,*]

1 : Leiden University Centre for Linguistics (LUCL)

2 : HANDS!Lab, Leiden University

3 : Leiden Institute of Advanced Computer Science [Leiden] (LIACS)

Niels Bohrweg 1 2333 CA Leiden - Pays-Bas

* : Corresponding author

(For abstracts in Ghanaian SL and French, see attached files at sciencesconf.org:lt4all2025:620464)

The use of signing is a resilient feature of Deaf communities around the world. In many societies, the vitality of sign languages and their transmission over time is compromised by ideologies and attitudes regarding the relative values of signing and speaking and regarding variations in hearing status. *Oralism* (i.e. stigmatization of signing) and *audism* (stigmatization of being deaf) interact with more general developments in society, politics, science, and technology, with fluctuating impacts on sign language vitality both at very local and at very international levels. As a result, deaf communities, especially in Western Educated Industrialized Rich Democracies (WEIRD) countries, have a long history of sign language activism, at times out in the open, and at times forced to go underground. In this talk, we will introduce some efforts of deaf signers and signing communities to maintain and vitalize their sign language. This includes a discussion of ways in which signing communities with indigenous sign languages respond to the introduction of sign languages from foreign countries via deaf education in many Southern countries.

---

sciencesconf.org:lt4all2025:617410

The National EFFT_Mod Project: An investigation into Sustainable Technology Practices, Approaches and Expectations in Italian HE

Giulia Staggini [1]

1 : Università degli Studi di Siena = University of Siena (UNISI)

Via Roma 56 - 53100 Siena - Italie

The “Eco-Friendly Flexible Teaching Model for Languages” (EFFT_Mod) project is a National Interest Research Project focused on exploring technology-driven strategies and educational initiatives to promote a more inclusive and sustainable teaching of language-related subjects, such as Linguistics and Literature. Based on data about current Higher Education practices, existing initiatives, and teachers' perceptions and concerns regarding technology, the research team has developed open-source materials and adaptable teaching models designed to be flexible, user-friendly, and accessible. These resources aim to make teaching more sustainable, optimize existing resources, and support diverse learners—particularly those often marginalized in HE, such as working students, learners with disabilities, and individuals facing economic hardship.

Il progetto PRIN “Eco-Friendly Flexible Teaching Model for Languages” (EFFT_Mod) ha l'obiettivo di esplorare strategie didattiche e strumenti tecnologi che promuovano un insegnamento più inclusivo e sostenibile delle discipline linguistiche in ambito universitario. Sulla base di un'indagine esplorativa inerente alle attitudini, percezioni e preoccupazioni dei docenti di lingua e letteratura inglese e tedesca riguardo all'applicazione della tecnologia alla didattica, il gruppo di ricerca ha sviluppato materiali open-source e modelli didattici flessibili da adattare e integrare in diversi contesti disciplinari. Tali risorse mirano a rendere l'insegnamento più sostenibile, a ottimizzare le risorse esistenti e a supportare una platea eterogenea di studenti, sopratutto coloro che spesso si trovano in situazioni di marginalità nel contesto accademico, come studenti lavoratori, con disabilità oppure in difficoltà economica.

Session O2: LT Innovations for Inclusive Communication Across All Ages and Diverse Abilities

Language Technology Innovations for Inclusive Communication Across All Ages and Diverse Abilities

Sainimili Tawake [1, *]

1 : Pacific Disability Forum

* : Corresponding author

Inclusive communication and accessible information are essential for fostering social inclusion and participation among people of all ages and abilities. People with disabilities continue to face significant challenges in accessing communication technology. A new report published by WHO and UNICEF in 2022 reveals that more than 2.5 billion people need one or more assistive products, which include apps that support communication and cognition.

Many people with disabilities face significant barriers including limited access to quality education and decent employment, primarily due to the lack of accessible communication technology. This presentation will highlight both the challenges faced by people with disabilities in the Pacific regarding accessible communication. It will explore recent innovations in language technology, including AI-driven speech recognition systems and assistive devices that enhance accessibility, promote inclusion, and support independence

The presentation will showcase studies on the use of assistive and accessible technology to enhance education and employment of people with disabilities. It will also identify key challenges faced by people with disabilities in the Pacific due to lack of accessible communication.

Finally, the presentation will illustrate how communication technologies, through strategic partnerships and development initiatives, can empower people with disabilities in selected Pacific countries.

Breaking Barriers: Next-Gen Voice Activated Companion for Inclusive Mobility

Pooja Gambhir [1], Amita Dev [2], Poonam Bansal [3], Shyam Sunder Agrawal [4]

1 : Amity Institute of Information Technology, Amity University (AIIT, Amity University Noida)

2 : Director General, Vivekananda Institute of Professional Studies. School of engineering and technology (VIPS)

3 : Indira Gandhi Delhi Technical University for Women (IGDTUW)

4 : KIIT GROUP OF COLLEGES Gurgaon (KIIT)

KIIT Campus, Sohna Road, Near Bhondsi, Gurgaon, Haryana - Inde

---

Language and Assistive Technologies (AT) can transform accessibility for individuals with diverse abilities. Be-Vocal Smart Crutch is an AI-powered, voice-enabled mobility aid, featuring obstacle detection, real-time navigation, and speech recognition to enhance independence and safety. To extend its utility, Sanrakshak, an AI-driven virtual caretaker, supports the elderly and individuals with speech and vision impairments by leveraging AIoT and audio processing to address challenges related to spatial awareness and mental health. The system incorporates multilingual voice assistance to enhance accessibility. With 2.2 billion visually impaired individuals and the AT market projected to grow from $6.11bn in 2024 to $7.03bn in 2025, this research aligns with the increasing demand for AI-driven accessibility solutions, fostering inclusivity to improve lives globally.

भाषा और सहायक प्रौद्योगिकियां (एटी) विविध क्षमताओं वाले व्यक्तियों के लिए पहुंच को बदल सकती हैं। बी-वोकल स्मार्ट क्रच एक एआई-संचालित, आवाज-सक्षम गतिशीलता सहायता है, जिसमें स्वतंत्रता और सुरक्षा बढ़ाने के लिए बाधा का पता लगाने, रीयल-टाइम नेविगेशन और भाषण मान्यता की विशेषता है। इसकी उपयोगिता का विस्तार करने के लिए, संरक्षक, एक एआई-संचालित वर्चुअल केयरटेकर, स्थानिक जागरूकता और मानसिक स्वास्थ्य से संबंधित चुनौतियों का समाधान करने के लिए एआईओटी और ऑडियो प्रोसेसिंग का लाभ उठाकर बुजुर्गों और भाषण और दृष्टि हानि वाले व्यक्तियों का समर्थन करता है। इस प्रणाली में सुलभता बढ़ाने के लिए बहुभाषी आवाज सहायता शामिल है। 2.2 बिलियन दृष्टिबाधित व्यक्तियों और एटी बाजार के 2024 में 6.11 बिलियन डॉलर से बढ़कर 2025 में 7.03 बिलियन डॉलर होने का अनुमान है, यह शोध एआई-संचालित एक्सेसिबिलिटी समाधानों की बढ़ती मांग के साथ संरेखित होता है, जो विश्व स्तर पर जीवन को बेहतर बनाने के लिए समावेशिता को बढ़ावा देता है।

---

sciencesconf.org:lt4all2025:616883

The Silent Revolution: AI-Powered Sign Language for a Fully Accessible World / La revolución silenciosa: Lengua de signos con IA para un mundo plenamente accesible

Martin Curzio [1]

1 : eldes

21 de septiembre 2690 - Uruguay

Sign language is the mother tongue of nearly **500 million Deaf people** worldwide, yet accessibility remains a major challenge in both digital and physical spaces. While AI has transformed spoken and written communication, **sign language lacks the structured data needed for true technological integration, until now**! Eldes is pioneering **AI-powered sign language education and interaction**, breaking down communication barriers and enabling inclusion at an unprecedented scale. In this talk, we will explore **how AI can bridge the accessibility gap**, the potential for **sign language Large Language Models (LLMs)**, and **real-world applications** that are already reshaping the future of accessibility worldwide.

La lengua de signos es la lengua materna de casi 500 millones de personas sordas en todo el mundo, pero su accesibilidad sigue siendo un gran reto tanto en los espacios digitales como en los físicos. Mientras que la IA ha transformado la comunicación oral y escrita, la lengua de signos carece de los datos estructurados necesarios para una verdadera integración tecnológica, hasta ahora! Eldes es pionera en la educación e interacción en lengua de signos impulsada por la IA, rompiendo las barreras de la comunicación y permitiendo la inclusión a una escala sin precedentes. Exploraremos cómo la IA puede salvar la brecha de la accesibilidad, el potencial de los LLM en lengua de signos y las aplicaciones reales que ya están cambiando el futuro de la accesibilidad mundialmente.

Inclusive Voices; Advancing Language Technology for People with Impaired Speech Communicating in Local Languages.

Gifty Ayoka [1], Richard Cave, Giulia Barbareschi, Katrin Tomanek, Isaac Wiafe, Catherine Holloway, @

1 : Global Disability Innovation Hub (GDI Hub)

The Centre for Digital Language Inclusion led by Global Disability Innovation Hub is pioneering a project in Ghana to collect local languages from individuals with impaired speech. We are developing automated speech recognition models to create applications that facilitate communication for these individuals. The need for communication aids is greatest in low resource language settings. Our goal is to enable culturally relevant communication in native languages with both people and devices. This initiative is the first of its kind in Africa, designed specifically for African needs. We aim to make all methods open-source, inviting global participation in our movement towards communication equality.

Center for Digital Language Inclusion a Global Disability Innovation Hub di wɔn anim no reyɛ akwampaefoɔ wɔ adwuma bi mu wɔ Ghana de aboaboa mpɔtam hɔ kasa ano afiri ankorankoro a wɔn kasa anyɛ yie hɔ. Yɛreyɛ kasa a wɔhunu ho nhwɛsoɔ a wɔde afiri yɛ de ayɛ dwumadie ahodoɔ a ɛbɛma nkitahodiɛ ayɛ mmerɛ ama saa ankorankoro yi. Nkitahodi mmoa ho hia kɛse wɔ kasa tebea horow a ɛho nhia pii mu.Yɛn botae ne sɛ yɛbɛma nkitahodi a ɛfa amammerɛ ho wɔ ɔman kasa mu a ɛne nnipa ne mfiri nyinaa betumi adi dwuma. Saa nhyehyɛe yi ne nea edi kan wɔ Afrika, a wɔayɛ ama Afrikafo ahiade titiriw. Yɛde asi yɛn ani so sɛ yɛbɛma akwan nyinaa ayɛ nea wɔabue ano, na yɛato nsa afrɛ wiase nyinaa sɛ wɔmfa wɔn ho nhyɛ yɛn dwumadi a ɛkɔ nkitahodi mu pɛyɛ mu no mu.

Detecting Cognitive Decline in Natural Speech: Focus on Words and Voice

Luwen Cao [1, *], Zhiming Bao [2, *]

1 : National Univerisity of Singapore (NUS)

Lower Kent Ridge Rd, Singapore - Singapour

2 : Faculty of Arts and Social Sciences, National Univerisity of Singapore (NUS)

Lower Kent Ridge Rd, Singapore - Singapour

* : Corresponding author

This presentation discusses our research using natural speech for early detection of cognitive decline. We analyzed spontaneous speech from older adults with MCI and cognitively healthy controls. Our findings revealed significant differences in noun usage . Individuals with amnestic MCI produce fewer and more abstract nouns. To further enhance our understanding, we are currently investigating phonetic and phonological patterns in the dataset. Natural speech analysis can be a valuable, non-invasive tool for early detection of cognitive decline. Language technology can be adapted to meet critical healthcare challenges.

Cette présentation porte sur notre recherche utilisant la parole naturelle pour la détection précoce du déclin cognitif. Nous avons analysé la parole spontanée d'adultes âgés atteints de troubles cognitifs légers (TCL) et de témoins cognitifs sains. Nos résultats ont révélé des différences significatives dans l'utilisation des noms. Les personnes atteintes de TCL amnésique produisent moins de noms et des noms plus abstraits. Pour approfondir notre compréhension, nous étudions actuellement les schémas phonétiques et phonologiques dans l'ensemble de données. L'analyse de la parole naturelle peut être un outil précieux et non invasif pour la détection précoce du déclin cognitif. La technologie du langage peut être adaptée pour répondre à des défis critiques en matière de soins de santé.

sciencesconf.org:lt4all2025:616205

Speech and Language Technologies for Mental Health and Wellbeing across Lifespan

Shrikanth Narayanan [1]

1 : USC Viterbi School of Engineering

Converging technological advances, from multimodal sensing and signal processing to machine learning, are enabling new possibilities for advancing speech science and in creating technologies supporting mental health across human lifespan. This includes understanding human speech and language communication and social behavior with wide-ranging applications in screening, diagnostics and treatment across varied domains of clinical and quality of life significance. This talk will highlight some advances, opportunities, and challenges using examples from domains such as distressed relationships, depression, autism spectrum disorder and dementia, and underscore the need for approaches that are inclusive, robust, safe and secure.

Abstract translated into Tamil (using Google translate)

மல்டிமோடல் சென்சிங் மற்றும் சிக்னல் ப்ராசஸிங் முதல் மெஷின் லேர்னிங் வரை ஒன்றி ணைந்த தொழில்நுட்ப முன்னேற்றங்கள், பேச்சு அறிவியலை மேம்படுத்துவதற்கும் மனித ஆயுட்காலம் முழுவதும் மன ஆரோக்கியத்தை ஆதரிக்கும் தொழில்நுட்பங்களை உருவாக் குவதற்கும் புதிய சாத்தியங்களை செயல்படுத்துகிறது. இதில் மனித பேச்சு மற்றும் மொழி த் தொடர்பு மற்றும் சமூக நடத்தை ஆகியவற்றைப் புரிந்துகொள்வது, மருத்துவ மற்றும் வா ழ்க்கைத் தரம் முக்கியத்துவம் வாய்ந்த பல்வேறு களங்களில் திரையிடல், கண்டறிதல் மற்று ம் சிகிச்சை ஆகியவற்றில் பரவலான பயன்பாடுகளுடன் உள்ளது. இந்த பேச்சு, சில முன்னே ற்றங்கள், வாய்ப்புகள் மற்றும் சவால்களை எடுத்துக்காட்டுகிறது, அதாவது மன உளைச்சலி ல் உள்ள உறவுகள், மனச்சோர்வு, ஆட்டிசம் ஸ்பெக்ட்ரம் கோளாறு மற்றும் டிமென்ஷியா போன்ற களங்களின் எடுத்துக்காட்டுகளைப் பயன்படுத்தி, உள்ளடக்கிய, வலுவான, பாது காப்பான மற்றும் பாதுகாப்பான அணுகுமுறைகளின் அவசியத்தை அடிக்கோடிட்டுக் காட் டும்.

Poster Session P3: Language Technology Solutions in a Multilingual World

(L)LMs for Multilingual and Cross-lingual Harmful Speech Mitigation

Daryna Dementieva [1]

1 : Technical University of Munich (TUM)

(EN)
Language Models (LMs) have shown impressive performance across various natural language processing tasks. However, their application in multilingual and multicultural contexts often exposes challenges, particularly in effectively addressing toxicity across different languages. In this poster, we present our study on toxicity mitigation techniques for nine languages, highlighting various cultural nuances and exploring if toxicity knowledge can be cross-lingualy transferred. Additionally, we propose strategies to enhance multilingual NLP solutions, with a focus on developing more proactive, accurate, and culturally aware approaches for moderating harmful speech using modern NLP advancements.
(UKR)
Мовні моделі ШІ продемонстрували вражаючу продуктивність у різних завданнях обробки природної мови. Однак їхнє застосування в багатомовних і мультикультурних контекстах часто викликає труднощі, зокрема в ефективному вирішенні проблеми токсичності в різних мовах. У цій доповіді ми презентуємо наше дослідження методів детекції токсичності для дев'яти мов, висвітлюючи різні культурні нюанси та досліджуючи, чи можна переносити знання про токсичність між мовами. Крім того, ми пропонуємо стратегії для вдосконалення багатомовних рішень НЛП, з акцентом на розробці більш проактивних, точних і культурно-орієнтованих підходів для модерації шкідливого мовлення з використанням сучасних досягнень НЛП.

---

sciencesconf.org:lt4all2025:617744

Bridging Cultures Through AI: Applying LLMs and RAG for Cross-Cultural Communication in Tourism

Winda Monika [1,*], Arbi Haza Nasution [2]

1 : Universitas Lancang Kuning (UNILAK)

Jl. Yos Sudarso No.KM. 8, Umban Sari, Kec. Rumbai, Kota Pekanbaru, Riau 28266 - Indonésie

2 : Universitas Islam Riau (UIR)

Jl. Kaharuddin Nst No.113, Simpang Tiga, Kec. Bukit Raya, Kota Pekanbaru, Riau 28284 - Indonésie

* : Corresponding author

Cross-cultural communication is a challenge in the hospitality industry, particularly for hotel receptionists who are responsible for interacting with international guests. This study develops an AI-powered chatbot using Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) to address language barriers. The chatbot incorporates RAG-enhanced tourism information retrieval for real-time recommendations and LLM-based machine translation for multilingual interactions. It evaluates multiple LLMs (Meta Llama, Google Gemma, Microsoft Phi, and Mistral), with Llama 3.1 70B scoring the highest (0.3712 BLEU), and is implemented using Ollama and Open WebUI. RAG-based responses were perceived as more informative by 87% of respondents. The potential of LLMs and RAG to enhance hospitality services and improve visitor interactions is demonstrated by the findings.

Komunikasi lintas budaya merupakan tantangan dalam industri perhotelan, terutama bagi resepsionis hotel yang bertanggung jawab berinteraksi dengan tamu internasional. Studi ini mengembangkan chatbot berbasis AI menggunakan Large Language Models (LLMs) dan Retrieval-Augmented Generation (RAG) untuk mengatasi hambatan bahasa. Chatbot ini mengintegrasikan RAG untuk pencarian informasi wisata secara real-time dan terjemahan mesin berbasis LLM untuk interaksi multibahasa. Sistem ini mengevaluasi beberapa LLM (Meta Llama, Google Gemma, Microsoft Phi, dan Mistral), dengan Llama 3.1 70B memperoleh skor tertinggi (0.3712 BLEU), serta diimplementasikan menggunakan Ollama dan Open WebUI. 87% responden menilai respons berbasis RAG lebih informatif. Temuan ini menunjukkan potensi LLMs dan RAG dalam meningkatkan layanan perhotelan dan interaksi dengan pengunjung.

sciencesconf.org:lt4all2025:617768

Bridging the Gap in Low-Resource Language Translation: Pivot-based Hybrid Machine Translation Using LLMs and Bilingual Dictionaries

Arbi Haza Nasution [1,*], Winda Monika [2], Panji Rachmat Setiawan [1], Rizdqi Akbar Ramadhan [1]

1 : Universitas Islam Riau (UIR)

Jl. Kaharuddin Nst No.113, Simpang Tiga, Kec. Bukit Raya, Kota Pekanbaru, Riau 28284 - Indonésie

2 : Universitas Lancang Kuning (UNILAK)

Jl. Yos Sudarso No.KM. 8, Umban Sari, Kec. Rumbai, Kota Pekanbaru, Riau 28266 - Indonésie

* : Corresponding author

Bridging the gap in low-resource language translation is challenging due to scarce high-quality bilingual corpora. This study integrates Large Language Models (LLMs), bilingual dictionaries, and client-server technology to enhance real-time translation in seminar settings. The system enables speech-to-text translation from high-resource languages (e.g., English, French) into Indonesian using LLMs, followed by word-to-word translation into Minangkabau via a bilingual dictionary. The translation is broadcasted parallel to client devices, allowing participants to access text and real-time audio playback. Usability evaluation (4.25 average score) confirms its effectiveness in multilingual events, ensuring seamless communication and improved accessibility for low-resource languages. The adequacy and fluency evaluations yielded average scores of 3.86 and 3.82, respectively, further validating the system's linguistic performance.

Manjanjang rintang dalam panarjemahan bahasa nan sakik sumberdayo masih tantangan gadang karano kurangnyo korpus dwibahaso nan barupo tinggi. Kajian iko manggabuangkan Large Language Models (LLMs), kamus dwibahaso, jo teknologi client-server untuak maningkekkan panarjemahan real-time dalam lingkungan seminar. Sistim nan diusulkaan mamungkinkan panarjemahan suara ka teks dari bahasa nan banyak sumberdayo (contohnya, Inggris, Prancis) ka Bahasa Indonesia manggunoan LLMs, lalu diterjemahkan sacara word-to-word ka Minangkabau manggunoan kamus dwibahaso. Hasil tarjemahan iko lalu disebarkan sacara paralel ka parangkat klien, mamungkinkan hadirin manjadiakan teks tarjemahan atau mandanga tarjemahan malalui audio real-time. Hasil evaluasi (skor 4.25) manunjukkan efektivitasnyo dalam acara multibahaso, manjamin komunikasi nan lancar jo maningkekkan aksesibilitas untuak bahasa nan sakik sumberdayo. Evaluasi adequacy jo fluency mancapai skor rata-rata 3.86 jo 3.82, nan semakin memvalidasi kinerja linguistik sistem iko.

---

sciencesconf.org:lt4all2025:617835

Converging to a Lingua Franca: Evolution of Linguistic Regions and Semantics Alignment in Multilingual Large Language Models

Hongchuan Zeng [1], Senyu Han [1], Lu Chen [1,*], Kai Yu [1,*]

1 : Shanghai Jiao Tong University [Shanghai]

800 Dongchuan Road, Shanghai, 200240 - Chine

* : Corresponding author

Large language models (LLMs) have demonstrated remarkable performance, particularly in multilingual contexts. While recent studies suggest that LLMs can transfer skills learned in one language to others, the internal mechanisms behind this ability remain unclear. We observed that the neuron activation patterns of LLMs exhibit similarities when processing the same language, revealing the existence and location of key linguistic regions. Additionally, we found that neuron activation patterns are similar when processing sentences with the same semantic meaning in different languages. This indicates that LLMs map semantically identical inputs from different languages into a "Lingua Franca", a common semantic latent space that allows for consistent processing across languages. This semantic alignment becomes more pronounced with training and increased model size, resulting in a more language-agnostic activation pattern. Moreover, we found that key linguistic neurons are concentrated in the first and last layers of LLMs, becoming denser in the first layers as training progresses. Experiments on BLOOM and LLaMA2 support these findings, highlighting the structural evolution of multilingual LLMs during training and scaling up. This paper provides insights into the internal workings of LLMs, offering a foundation for future improvements in their cross-lingual capabilities.

大语言模型（LLMs）在多语言环境中展现出了卓越的性能。尽管近期研究表明，LLMs 能够将一种语言中学到的技能迁移到其他语言，但这种能力背后的内部机制仍不清楚。我们观察到，LLMs 在处理相同语言时的神经元激活模式存在相似性，这揭示了关键语言区域的存在及其位置。此外，我们发现，当处理不同语言中语义相同的句子时，LLMs 的神经元激活模式也表现出相似性。这表明，LLMs 能够将不同语言中语义相同的输入映射到一个“通用语”（Lingua Franca），即一个通用的语义潜在空间，从而实现跨语言的一致处理。这种语义对齐随着训练的进行和模型规模的增加而更加显著，最终表现为更加语言无关的激活模式。此外，我们发现关键语言神经元主要集中在 LLMs 的前几层和后几层，并且随着训练的推进，关键神经元在前几层的分布更加密集。在 BLOOM 和 LLaMA2 上进行的实验支持了这些发现，揭示了多语言 LLMs 在训练和参数扩展过程中结构的演化。本文为理解 LLMs 的内部工作机制提供了新的视角，并为其跨语言能力的进一步提升奠定了基础。

sciencesconf.org:lt4all2025:617223

Counter-narrative generation against hate speech

Arturo Montejo-Ráez [1] , Helena Bonaldi [2] , María Estrella Vallecillo-Rodríguez [1] , Irune Zubiaga [3] , Aitor Soroa [3] , María Teresa Martín-Valdivia [4] , Marco Guerini [2] , Rodrigo Agerri [3]

1 : Universidad de Jaén (UJA)

Las Lagunillas s/n, 23071 - Jaén - Espagne

2 : Fondazione Bruno Kessler

3 : Universidad del País Vasco [Espainia] / Euskal Herriko Unibertsitatea [España] = University of the Basque Country [Spain] = Université du pays basque [Espagne] (UPV / EHU)

Barrio Sarriena s/n, 48940 Leioa, BizkaiaCampus d'Álava : Vice-Rectorado, San Antonio 41, 01005 Vitoria ; campus de Biscaye et services centraux : Apdo 1397, 48080 Bilbao ; campus de Guipúzcoa : Vice-Rectorado, Fuenterrabia 13-1° , 20006 San Sebastián - Espagne

4 : Universidad de Jaén (UJA)

Campus Las Lagunillas, s/n, 23071 Jaén - Espagne

This work presents the shared task of the First Workshop on Multilingual Counterspeech Generation at COLING 2025, focused on the generation of counter-narratives in Basque, English, Italian and Spanish to combat hate speech. It addresses challenges such as external knowledge integration, evaluation and automatic text generation in low-resource languages. Evaluation methods include BERTScore, ROUGE-L, BLEU and JudgeLM classification. The study emphasizes the need to define quality counter-narratives (persuasiveness, length, aggressiveness, correctness), improve evaluation methods and generation systems in low-resource languages. With 100 submissions, it advances counter-narrative generation and promotes safer online discussions.

Este trabajo presenta la tarea compartida del First Workshop on Multilingual Counterspeech Generation en COLING 2025, centrada en la generación de contranarrativas en euskera, inglés, italiano y español para combatir el discurso del odio. Aborda retos como la integración de conocimiento externo, la evaluación y la generación automática de textos en lenguas con pocos recursos. Los métodos de evaluación incluyen BERTScore, ROUGE-L, BLEU y la clasificación JudgeLM. El estudio hace hincapié en la necesidad de definir contranarrativas de calidad (persuasividad, longitud, agresividad, corrección), mejorar los métodos de evaluación y los sistemas de generación en lenguas de escasos recursos. Con 100 sistemas evaluados, avanza en la generación de contranarrativas y promueve debates en línea más seguros.

---

sciencesconf.org:lt4all2025:617985

Hyperglot – a toolkit for discovering language support in digital fonts

David Březina [1]

1 : Rosetta Research

Hyperglot helps answer a seemingly simple question about language support in fonts that is deceptively complex: "When can I use a font to set texts in a particular language?" or, more generally, "What do I need to represent a language in writing in a digital environment?". Hyperglot is an open-source toolkit built around a database of language orthographies. The database currently includes 783 languages. The ultimate goal is to describe the needs that all languages of the world have from the digital ecosystem of keyboards, encodings, and word shaping systems (also called text stack) and any additional design requirements.

Hyperglot pomáhá odpovědět na zdánlivě jednoduchou otázku ohledně jazykové podpory v písmech: „Kdy mohu použít písmo pro sazbu textů v určitém jazyce?“ nebo obecněji: „Co potřebuji k zápisu konkrétního jazyka v digitálním prostředí?“. Hyperglot je open-source sada nástrojů postavená nad databází znakových sad. Databáze v současné době obsahuje 783 jazyků. Konečným cílem je popsat potřeby všech světových jazyků v rámci digitálního ekosystému klávesnic, kódování a renderovacích systémů (tzv. text stack) a doplňujících požadavků na design.

sciencesconf.org:lt4all2025:616935

Introducing Unicode CLDR Keyboards: an open keyboard specification and repository designed for Indigenous languages

Marc Durdin [1] , Andrew Glass [2] , Tex Texin [3] , Steven Loomis [4] , Jan Kučera [5]

1 : SIL Global (SIL)

7500 West Camp Wisdom Rd Dallas TX 75236 - États-Unis

2 : Microsoft Corporation [Redmond]

Redmond, WA 98052 - États-Unis

3 : Translation Commons

4 : Code Hive TX

5 : Unicode Consortium

The CLDR (Common Locale Data Repository) Keyboard Subcommittee has developed an XML definition for keyboards, defining text input requirements for the world's languages. This format allows the physical and virtual (on-screen or touch) keyboard layouts for a language to be defined in a single file. The goal of this project is that, where otherwise unsupported languages are concerned, CLDR becomes the common source for keyboard data, for use by platform/operating system developers and vendors. CLDR will also become the point of contact for keyboard authors and language communities to submit keyboard layouts.

Le CLDR (dépôt commun des données de locales) a développé une définition pour les claviers, définissant les spécifications en matière de saisie de texte pour des langues du monde. Ce format permet de définir la disposition des claviers physiques et virtuels (à l'écran ou en mode tactile). L'objectif est que CLDR devienne la source commune de données pour les claviers des langues qui ne seraient pas encore prises en charge, à destination des développeurs et des fournisseurs de plates-formes et de systèmes d'exploitation. CLDR deviendra également le point de contact unique pour ceux qui souhaitent soumettre des dispositions de claviers.

---

sciencesconf.org:lt4all2025:615965

La traduction automatique et le multilinguisme dans les sciences

Nicolas Bacaer [1]

1 : Institut de Recherche pour le Développement

Institut de Recherche pour le Développement

Avec la traduction automatique relue et corrigée, il est désormais assez facile de diffuser des livres scientifiques dans de nombreuses langues. Il est aussi possible de traduire dans sa propre langue de nombreux textes scientifiques en accès libre, comme certains livres et comme les textes de présentation de certains prix scientifiques internationaux. Mais le conformisme et la servilité des classes dominantes de nombreux pays vis-à-vis de la superpuissance dominante du moment sont des freins très puissants à un plus grand multilinguisme dans la littérature scientifique.

Mithilfe von korrekturgelesenen maschinellen Übersetzungen ist es heute recht einfach, wissenschaftliche Bücher in vielen Sprachen zu verbreiten. Es ist auch möglich, viele frei zugängliche wissenschaftliche Texte in Ihre eigene Sprache zu übersetzen, wie z. B. einige Bücher und wie die Ankündigungstexte für einige internationale Wissenschaftspreise. Doch der Konformismus und die Unterwürfigkeit der herrschenden Klassen in vielen Ländern gegenüber der aktuell dominierenden Supermacht sind sehr starke Bremsen für eine größere Vielsprachigkeit in der wissenschaftlichen Literatur.

sciencesconf.org:lt4all2025:617593

Language Technologies Helping against Free Will Manipulation in CEDMO2 Project

Ondřej Bojar [1]

1 : Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL)

Malostranske nam. 25, Praha 1, 11800 - République tchèque

The need to protect human free will from technologically determined manipulation is one of the conditions for the existence of democracy in the 21st century. The CEDMO2 project combines technological, social, scientific, media and institutional aspects and aims to create multiple tools and measures, expanding current empirical knowledge. The poster presents a selection of these tools: real-time speech translation aimed for fact checking and disinformation detection, tools for manual and automatic fact checking, and a foreseen experiment where large language models support teaching of media literacy.

Potřeba chránit lidskou svobodnou vůli před technologicky vedenou manipulací je jednou z podmínek existence demokracie v 21. století. Projekt CEDMO2 spojuje technologické, sociální, vědecké, mediální a institucionální aspekty s cílem vytvořit více nástrojů a opatření, která rozšíří současné empirické poznatky. Na posteru představujeme výběr z těchto nástrojů: simultánní překlad mluvené řeči zaměřený na kontrolu faktů a detekci dezinformací, nástroje pro manuální a automatickou kontrolu faktů a plánovaný experiment, kde velké jazykové modely podpoří výuku mediální gramotnosti.

sciencesconf.org:lt4all2025:618581

Linguistic data acquisition in participatory sciences

Jouitteau Mélanie [1]

1 : Centre de recherche sur la langue et les textes basques (IKER)

Université de Pau et des Pays de l'Adour, université Bordeaux Montaigne, Centre National de la Recherche Scientifique : UMR5478

Château-Neuf Place Paul Bert 64100 BAYONNE - France

The future of linguistic diversity in language technologies lies in our ability to collect processable data for the thousands of languages spoken today. The scale of the challenge calls for a digital approach, but above all for the development of elicitation tools on a global scale. Beyond corpus linguistics for which NLP tools exist, descriptive and formal linguistics have decades of experience in fieldwork elicitation practice and theory, but their data remains digitally under-exploited. I advocate equipping these disciplines and mobilizing their expertise.

L'avenir de la diversité linguistique dans les technologies du langage repose sur notre capacité à collecter des données traitables pour les milliers de langues parlées aujourd'hui. L'ampleur du défi impose une approche numérique, mais nécessite surtout de développer une élicitation outillée à l'échelle mondiale. Au-delà des linguistiques de corpus outillées par le TAL, les linguistiques descriptives et formelles ont des décennies d'expérience pratique et théorique en élicitation de terrain, mais leurs données restent sous-exploitées numériquement. Je plaide pour l'outillage de ces disciplines et la mobilisation de leur expertise.

sciencesconf.org:lt4all2025:619765

LYRA - Language verY Rare for All

Ibrahim Merad [1] , Amos Wolf [1] , Ziad Mazzawi [1] , Yannick Léo [1,*]

1 : Emerton Data

Emerton Data 16 avenue hoche - France

* : Corresponding author

In the quest to overcome language barriers, encoder-decoder models like NLLB have expanded machine translation to rare languages, with some models (e.g., NLLB 1.3B) even trainable on a single GPU. While general-purpose LLMs perform well in translation, open LLMs prove highly competitive when fine-tuned for specific tasks involving unknown corpora. We introduce LYRA (Language verY Rare for All), a novel approach that combines open LLM fine-tuning, retrieval-augmented generation (RAG), and transfer learning from related high-resource languages. This study is exclusively focused on single-GPU training to facilitate ease of adoption. Our study focuses on two-way translation between French and Monégasque, a rare language unsupported by existing translation tools due to limited corpus availability. Our results demonstrate LYRA's effectiveness, frequently surpassing and consistently matching state-of-the-art encoder-decoder models in rare language translation.

Session O3: LT Solutions for Revitalizing Endangered Languages

sciencesconf.org:lt4all2025:616790

The SWiP Project: Enhancing the Digital Presence of South African Languages on Wikipedia

Menno Van Zaanen [1], Nomsa Skosana [1]

1 : South African Centre for Digital Language Resources (SADiLaR)

Building A7 North-West University Potchefstroom Campus Potchefstroom South Africa - Afrique du Sud

---

Most South African languages have little coverage on Wikipedia, which limits their digital presence. The SWiP project is a collaboration between the South African Centre for Digital Language Resources (SADiLaR), Wikipedia, and the Pan South African Language Board (PanSALB). Its aim is to train and mobilize language communities to advance the contributions in South African languages on Wikipedia. During its initial phase, training events (throughout the country) and a writing competition were organized with a specific focus on isiNdebele. The next phase expands this initiative in size (e.g., through train-the-trainer events), and towards other languages and organizations, potentially also outside South Africa.

Amalimi amanengi weSewula Afrika awakajameleki ku-Wikipedia, okwenza bonyana aqalane nobudisi bokunzinza nokutholakala kwawo eenkundleni zedijithali. Ihlelo le-SWiP lisebenza ngokuhlanganyela ne-South African Centre for Digital Language Resources (SADiLaR), Wikipedia, ne-Pan South African Language Board (PanSALB). Umnqophalo ihlelweli kubandula kanye nokubuthelela imiphakathi yamalimi ukwenzela bonyana kuthuthukiswe igalelo emalimini weSewula Afrika ku-Wikipedia. Ngesikhathi sesigaba sayo sokuthoma, kube neemfundobandulo enarheni yoke nephaliswano lokutlola. Koke lokhu bekuhlelelwe khulu ukuthuthukisa isiNdebele ekundleni ye-Wikipedia. Isigaba esilandelako sizokukhulisa iphrojekthi le ngokwenza iimfundobandulo zokubandula ababanduli, nokubandakanya amanye amalimi neenhlangano ebezingakabandakanywa esigabeni sokuthoma. Sizokuqalelela nokusebenzisana nezinye iinarha ngaphandle kweSewula Afrika.

---

sciencesconf.org:lt4all2025:617438

Learning in Our Language Aprender en nuestro lingua

Cat Kutay [1, 2], Nicola Bidwell [3, 4]

1 : Charles Darwin University [Australia]

2 : Australian National University - Department of engineering

3 : Charles Darwin Univeristy (Australia)

Ellengowan Drive - Australie

4 : Rhodes University (South Africa)

In Australia we have over 350 languages, spoken by whole communities or only as occasional words in Aboriginal English or Kriol. We are collaborating with different First Nations to develop LLM tools that translate formal English documents into Aboriginal English or Kriol and then into local languages. First Nations students using the institution's learning system will be our first clients. First Nations workshop participants explore the models we develop, as we explain the technology and development processes using real-world examples. Concerns are raised around harm in the collection and use of language material, local agency in model creation and ownership, and impacts on employment.

En Australia tenemos más de 350 idiomas, hablado por comunidades enteras o sólo como palabras ocasionales en inglés aborigen o kriol. Estamos colaborando con diferentes Primeras Naciones para desarrollar herramientas LLM que traduzcan documentos del inglés formal al inglés aborigen o kriol y luego a los linguas locales. Los estudiantes de las Primeras Naciones que utilicen el sistema de aprendizaje de la institución serán nuestros primeros clientes. Los participantes del taller de las Primeras Naciones exploran los modelos que desarrollamos, mientras explicamos la tecnología y los procesos de desarrollo utilizando ejemplos del mundo real. Se plantean preocupaciones en torno al daño en la recopilación y el uso de material lingüístico, la agencia local en la creación y propiedad de modelos y los impactos en los empleo.

---

sciencesconf.org:lt4all2025:616362

Speech Technology for Preservation and Revitalization of the Ainu Language

Tatsuya Kawahara [1]

1 : Kyoto University

Sakyo-ku, Kyoto 606-8501 - Japon

This talk will introduce our recent projects on speech and language technology to preserve and revitalize the Ainu language, which is listed as a critically endangered language in Japan. A large stock of recordings of Ainu oral folklore has been collected since the 1970s, but its annotation was challenging and time-consuming. We have developed an automatic speech recognition (ASR) system to transcribe the speech and conduct alignment of speech and text. As a result, speeches of over 300 hours have been archived to be open to the public. We have also developed a text-to-speech (TTS) system, which can generate folklore speeches without audio recordings. It is also used to create a reference for speech practice. These technologies will hopefully assist various kinds of speaking training. The projects were conducted with the support of the Japanese government and in collaboration with the National Ainu Museum.

本講演では、きわめて深刻な消滅の危機にあるアイヌ語の保存と再生のための音声言語技術に関するプロジェクトを紹介する。1970年代から口頭伝承が数多く収録されてきたが、その書き起こしとアノテーションは非常に困難であった。私たちは、音声認識システムを開発し、音声とテキストの対応付けを自動化した。 その結果、300時間を超える音声アーカイブ化され、一般公開されるようになった。また、録音の存在しない説話の音声を生成できる音声合成システムも開発した。 これはスピーチ練習のための参照としても使われている。これらの技術は様々な発話訓練に役立つことが期待される。このプロジェクトは、日本政府の支援を受け、国立アイヌ民族博物館と協力して実施された。

Smugri: Language Technologies for Low-Resource Finno-Ugric Languages and Dialects

Mark Fishel [1]

1 : Institute of Computer Science [University of Tartu, Estonie]

Narva mnt 18, 50090, Tartu - Estonie

The languages in the Finno-Ugric family range from Estonian, Finnish and Hungarian, each spoken by millions of speakers, to dozens of resource-poor and endangered languages like Livonian, Votic and Pite Sami. The goal of the Smugri project is to create open and free language technologies for this whole language family, focusing primarily on its resource-poor members. More generally we develop methodology of extremely low-resource translation, language modeling, speech processing and computer-aided language learning. Our work relies on cross-lingual knowledge transfer, careful management of data and annotation quality and tight collaboration with the speakers' communities.

Soome-ugri keelerühma kuuluvad nii eesti, soome ja ungari keel, mida räägivad kokku miljonid inimest, kui ka kümned ressurssivaesed ja ohustatud keeled, nagu liivi, vadja ja pite saami. Smugri projekti eesmärk on luua avatud ja vabu keeletehnoloogiaid kogu sellele keelerühmale, keskendudes eeskätt väheste ressurssidega keeltele. Üldisemalt tegeleme ülivaeste ressurssidega masintõlke, keelemudelite, kõnetöötluse ja arvutipõhise keeleõppe metoodika arendamisega. Meie töö tugineb keeltevahelisele teadmussiirdele, andmete ja märgenduskvaliteedi hoolikale haldamisele ning tihedale koostööle nende keelte kogukondadega.

Digital Tools for Cherokee and the Role of a Morphological Analyzer in Language Empowerment

Jeannette Stewart [1, *]

1 : Translation Commons

* : Corresponding author

Cherokee is an endangered Iroquoian language, spoken in Oklahoma and North Carolina (Montgomery-Anderson 2008). The Cherokee community is actively working on revitalization efforts, including speaker training programs at the Cherokee Language Institute (https://language.cherokee.org/). Cherokee is a polysynthetic language, and has a wide range of prefixes and suffixes. These affixes undergo a complex set of vocalic changes based on phonetic context. This can be particularly challenging for language learners. This project presents Morphy, a morphological analyzer that can both parse and generate Cherokee verbs. Morphy is part of a larger effort to build tools to support Cherokee language students.

Le cherokee est une langue iroquoienne en voie de disparition, parlée en Oklahoma et en Caroline du Nord (Montgomery-Anderson, 2008). La communauté Cherokee est très active dans les efforts de revitalisation, y compris des programmes de formation de locuteurs au Cherokee Language Institute (https://language.cherokee.org/). Le cherokee est une langue polysynthétique avec un large éventail d'affixes. Ces affixes subissent un ensemble complexe de changements vocaliques. Cette complexité présente des obstacles pour apprendre la langue. Ce projet présente Morphy, un analyseur morphologique capable à la fois d'analyser et de générer des verbes cherokee. Morphy fait partie d'un effort plus large visant à créer des outils pour soutenir les étudiants en langue cherokee.

sciencesconf.org:lt4all2025:616318

The importance of digital language and music archives for language and cultural continuity, an example from the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)

Nick Thieberger [1, 2, *]

1 : University of Melbourne

2 : ARC Centre of Excellence for the Dynamics of Language

* : Corresponding author

Many records of the world's languages are inaccessible, on paper, or audio tape, and are in need of preservation. Digital language archives look after these important records, digitising analog recordings and considering access conditions that need to apply. They also allow material to be accessed remotely, so that contents of records can be understood and the appropriate language group can be identified. In this talk I outline our experience at the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) in its 22nd year and show how we return copies of material using local wifi transmitters (Raspberry Pi).

De nombreux documents sur les langues du monde sont inaccessibles, sur papier ou sur bande audio, et doivent être préservés. Les archives numériques de langues prennent soin de ces documents importants, numérisent les enregistrements analogiques et prennent en compte les conditions d'accès qui doivent s'appliquer. Elles permettent également d'accéder au matériel à distance, afin que le contenu des documents puisse être compris et que le groupe linguistique approprié puisse être identifié. Dans cet exposé, je décris notre expérience aux Archives régionales et du Pacifique pour les sources numériques des cultures en danger (PARADISEC) dans sa 22e année et montre comment nous restituons des copies de documents à l'aide d'émetteurs wifi locaux (Raspberry Pi).

THEME: CHALLENGE

Session Keynote 1

Session O4: Language Technologies for All: Empowering Communities or Reinforcing Dependence?

Lauleo: Bringing our voices together for the future of ʻōlelo Hawaiʻi

Keoni Mahelona [1]

1 : Te Reo Irirangi o Te Hiku o Te Ika (Te Hiku Media)

Eia ke komo mai nei ka AI ma nā kelepona a me nā lolo uila i mea e kikokiko ʻia ai ka ʻōlelo i hoʻopuka ʻia e ke kanaka (leo-a-kiko - speech to text), a i ʻole, e ʻōlelo mai i ke kanaka (kiko-a-leo - text to speech), akā ma ka ʻōlelo Pelekania wale nō i kēia manawa. ʻO kekahi pahu hopu o Lauleo ka hoʻomohala ʻana i ia ʻenehana like no ka ʻōlelo Hawaiʻi ma ka huliāmahi ʻana me ke kaiāulu ʻōlelo. E aho hoʻi ko kākou alu like ʻana, o hana auaneʻi ʻo "Big Tech" mā. He papahana ʻo Lauleo e kākoʻo pū ʻia nei e Te Hiku Media ma Aotearoa a me nā hui o Hawaiʻi, akā, e ʻilau nā leo e kō ai ka hana.

AI language tools are rapidly becoming a part of our world, allowing our phones and other devices to understand our voices and type what we say or read to us in our language. Lauleo seeks to work with the language community and its supporters to develop these tools. Lauleo is a partnership between Te Hiku Media and various Hawaiʻi organizations and institutions, but it will require many voices—I lau nā leo—to be a success.

Fostering Language Community Empowerment and Development: The Holistic Approaches – Some Cameroonian and Canadian Cases

Evelyn Chibaka Epse Fogwe [1,*]

1 : University of Buea (UB)

P.O.Box 69, Buea, SW Region - Cameroun

* : Corresponding author

One of the main objectives of 'language revitalization' is to use our indigenous knowledge and cultural heritage artifacts to foster the intergeneration transfer, empowerment and long-lasting development of our communities. So far, the traditional methods of focusing on isolated sectors and/or factors interventions when empowering our language communities have not yielded or taken us to our intended long-term development and sustainable changes in today's high speed technologically advancing world. In this paper, we are arguing and proposing the use of 'Holistic Approaches' for a comprehensive, collaborative, impactful, and multiple interconnected networking method in fostering strategic Community empowerment and innovative development.

L'un des principaux objectifs de la « revitalisation des langues » est d'utiliser nos connaissances autochtones et nos artefacts du patrimoine culturel pour favoriser le transfert intergénérationnel, l'autonomisation et le développement durable de nos communautés. Jusqu'à présent, les méthodes traditionnelles consistant à se concentrer sur des secteurs isolés et/ou des interventions sur des facteurs lors de l'autonomisation de nos communautés linguistiques n'ont pas cédé ou nous ont menés à notre développement à long terme et à nos changements durables prévus dans le monde technologiquement avancé d'aujourd'hui. Dans ce document, nous soutenons et proposons l'utilisation d'« approches holistiques » pour une méthode de réseautage complète, collaborative, percutante et interconnectée afin de favoriser l'autonomisation stratégique des communautés et le développement innovant.

Indigenous Peoples and Languages of Indonesia: Challenges and Realities in a Diverse Archipelago

Marolop Sahat Martua Manalu [1]

1 : Aliansi Masyarakat Adat Nusantara (The Indigenous Peoples Alliance of The Archipelago) (AMAN)

https://aman.or.id/contact-us - Indonésie

Indonesia, an archipelagic nation with 17,380 islands, is home to 1,340 ethnic groups and thousands of native languages. The Indigenous Peoples Alliance of the Archipelago (AMAN) estimates around 80 million Indigenous Peoples belong to thousands of communities, with 2,569 communities (21 million people) as AMAN members. While many Indigenous Peoples still use their native languages, others face extinction due to declining speakers and the loss of indigenous territories. Despite Indonesia's linguistic diversity, language technology mainly supports Bahasa Indonesia and major Indigenous languages like Javanese and Sundanese. Most other Indigenous languages remain uncovered, further accelerating their endangerment and digital exclusion.

Indonesia ima sada negara namanghamham 17.380 pulo, na gabe jabu tu 1.340 bangso dohot marribu hata. Aliansi Masyarakat Adat Nusantara (AMAN) mamparhirahon 80 juta halak adat namardomu tu marribu punguanpunguan di Indonesia, 2.569 punguan (21 juta halak) nungga gabe ruas ni AMAN. Saonari godang dope namamangke hatanabe, asing ni i saonari mangadopi hapunuan disiala naung moru parhatana jala mago tano adat nasida. Asing sian ragam ni hata di Indonesia, tehnologi hata nuaengon holan mangangkupi hata Indonesia dohot hata gomodangan dope songon hata Jawa dohot Sunda. Santurpuk godang nari dang haidaan dope nagabe pagirahon hapunuan jala siding secara digital nasida.

sciencesconf.org:lt4all2025:619485

LIV - Un écosystème IA adapté aux langues et cultures régionales. LIV - A Comprehensive, local , Community adapted AI Ecosystem

Romain Deceuninck [1]

1 : LIV.CORSICA

Calvi 20260 Corsica - France

À travers **LIV**, nous avons doté la Corse de son propre Ecosysteme IA, **composée de Livia** et **Liviu**.
L'un, **Livia**, est un Compagnion éducatif conçu pour les enfants, avec un langage adapté favorisant l'apprentissage naturel du corse. L'autre, **Liviu**, s'adresse aux adultes : véritable assistant culturel, il génère tout type de contenu, renseigne et accompagne les Corses dans la compréhension de leur langue, de leur histoire et de leurs spécificités culturelles.
LIV **tend** à s'affranchir de la dependance aux grands modèles linguistiques pour préserver et transmettre un patrimoine vivant, tout en étant conçu et développé exclusivement par les insulaires.

Through LIV, we have equipped Corsica with its own AI ecosystem, composed of Livia and Liviu.
Livia is an educational companion designed for children, using language adapted to encourage the natural learning of Corsican. **Liviu**, on the other hand, is aimed at adults: a true cultural assistant, it generates all types of content, informs, and helps Corsicans understand their language, history, and cultural specificities.
LIV strives to break free from dependence on major linguistic models to preserve and transmit a living heritage while being entirely **designed and developed by the island's inhabitants.**

sciencesconf.org:lt4all2025:616585

Recognition and Collaboration: An Approach to Linguistic Diversity

Eugenia Urrere [1, *]

1 : Indigenius (LSP)

* : Corresponding author

In an increasingly globalized world, notions of inclusion and culture are diffuse. Indigenous communities perceive themselves as bicultural. More than being empowered or dependent on others, it is crucial to facilitate, accompany; this is the real challenge. Technology can play a vital role in their revitalization, as long as it is accessible and respectful of their contexts, without falling into cultural relativisms. It is essential to learn how to adapt to the rhythms of communities and avoid biases that reproduce old structures. The proposal is to collaborate and facilitate the strengthening of their social status and cultural self-esteem.

En un mundo cada vez más globalizado, las nociones de inclusión y cultura son difusos. Las comunidades indígenas, se auto perciben como biculturales. Más que ser empoderadas o depender de otros, es crucial facilitar, acompañar; este es el verdadero desafío. La tecnología puede jugar un papel vital en su revitalización, siempre y cuando sea accesible y respetuosa de sus contextos, sin caer en relativismos culturales. Es fundamental aprender a adaptarnos a los ritmos de las comunidades y evitar sesgos que reproduzcan viejas estructuras. La propuesta es colaborar y facilitar el fortalecimiento de su estatus social y la autoestima cultural.

Poster Session P4: Language Technology for Language Revitalization

sciencesconf.org:lt4all2025:616925

Bridging Generations: A Pathway to Revitalise Endangered Indigenous Languages

Natasha Martin [1, 2, 3, *] , Jonathan Newchurch [4, 5, 6, *]

1 : Te Arawa

2 : Ngāti Ranginui

3 : Warra Wangkatitya

Kaurna Yarta - Australie

4 : Kaurna

5 : Narungga

6 : Warra Wangkatitya

Kaurna Yarta - Australie

* : Corresponding author

Ko Te Arawa me Ngāti Ranginui ōku iwi. Nō Tauranga Moana ahau. Ko Natasha Demi Parekauwau Martin tōku ingoa. Ko Bidois tōku ingoa whānau. E noho ana au i te whenua o Kaurna Yarta.

Warra Wangkatitya, meaning “to speak language” in Kaurna, is an Indigenous-led not-for-profit founded by Dr. Jonathan Newchurch (Kaurna and Narungga) and myself, dedicated to re-establishing the intergenerational transmission of Kaurna language.

The Purruna Kauwi (Healing Waters) Framework draws on Indigenous methodologies and the profound significance of water as a symbol of language. Like water, language is dynamic and interconnected—it does not follow a fixed, unidirectional path but moves through different spaces.

The framework consists of three key components:

- Warra Pudna (Language Spring)—inspired by kōhanga reo (Aotearoa NZ), representing the nurturing of language in children at its purest source.
- Warra Pari (Language River)—our adult immersion program, influenced by the Salish Fluency Transfer System, providing fluid, adaptable pathways to fluency.
- Warra Yarlu (Language Ocean)—the vast body of ancestral and cultural knowledge.

Just as water nourishes and sustains life, a thriving language sustains identity, intergenerational knowledge, and wellbeing. This framework informs language technology by shaping pedagogy, digital tools, and immersion strategies, creating opportunities to revitalise and sustain intergenerational language use and transmission.

Inclusion or Digital Colonialism : The Challenges of Indigenous Digital Sovereignty and AI Decolonialism

Cristian Ahumada Oliva [1] , Fatiha Sadat *

1 : Université du Québec à Montréal = University of Québec in Montréal (UQAM)

Université du Québec à Montréal CP 8888, succursale Centre-ville Montréal (Québec) H3C 3P8 - Canada

* : Corresponding author

Indigenous peoples have suffered forced migration, territorial dispossession, colonization and cultural genocide; the digital spaces can be tools for resistance and cultural reconstruction. Protecting culture does not mean hiding it but reproducing and strengthening it. Data and languages technologies for Indigenous languages should be developed and/or incorporated in a respectful, ethical and correct way (by, with and for the Indigenous communities), with respect to the sovereignty and the culture, ensuring that the Indigenous knowledge is not distorted—both within the community and in the world. Additionally, we need to ensure that AI is diverse and inclusive (no language is left behind) and considers Indigenous knowledge without resorting to the famous quote “divide and conquer”.

Les peuples autochtones ont subi des migrations forcées, des dépossessions territoriales, la colonisation et le génocide culturel ; les espaces numériques peuvent être des outils de résistance et de reconstruction culturelle. Protéger la culture ne signifie pas la cacher mais la reproduire et la renforcer. Les technologies langagières ainsi que les données autochtones doivent être développées et/ou intégrées de manière respectueuse, éthique et correcte (par, avec et pour les communautés autochtones), dans le respect de la souveraineté et de la culture, en veillant à ce que les connaissances autochtones ne soient pas déformées, tant au sein de la communauté que dans le monde. En outre, nous devons nous assurer que l'IA est diversifiée et inclusive (aucune langue n'est laissée de côté) et qu'elle prend en compte les connaissances autochtones sans recourir à la célèbre citation « diviser pour mieux régner ».

---

sciencesconf.org:lt4all2025:619792

Join us in protecting humanity's linguistic heritage for eternity!

Katrine Loen [1] , Rune Bjerkestrand [1,*]

1 : Arctic World Archive (AWA)

Grønland 56, 3045 Drammen - Norvège

* : Corresponding author

The **Arctic World Archive (AWA)** is a repository for world memory located on Svalbard in the Arctic Ocean.
AWA ensures the survival of endangered languages by storing audio recordings, transcripts, and linguistic data. Using **PiqlFilm**, a future-proof, migration-free, unalterable, searchable and resilient storage medium, spoken traditions and linguistic heritage can be safeguarded for centuries.
As languages disappear, their unique knowledge and identity risk being lost forever. AWA provides a secure, long-term solution to keep these voices alive for future generations.
arcticworldarchive.org

Rejoignez-nous pour protéger le patrimoine linguistique pérenne de l'humanité.
The Arctic World Archive (AWA) est un dépôt de la mémoire mondiale situé à Svalbard, dans l'océan Arctique.
AWA assure la préservation des langues menacées de disparition en stockant des enregistrements audios, transcriptions et données linguistiques. Grâce à **PiqlFilm**, un support de stockage pérenne, sans migration, inaltérable, consultable et extrêmement résistant, les traditions orales et le patrimoine linguistique peuvent être sauvegardés pour des siècles.
Avec la disparition des langues, un savoir unique et une identité précieuse risquent d'être perdus à jamais.
AWA offre une solution sécurisée et durable pour préserver ces voix et les transmettre aux générations futures.
Participez à cette mission et contribuez à la sauvegarde de la diversité linguistique mondiale!
arcticworldarchive.org

---

sciencesconf.org:lt4all2025:617101

Localization in Africa's highly multilingual ecosystems: Bridging research and practice

Pierpaolo Di Carlo [1, 2]

1 : State University of New York at Buffalo (SUNY)

Buffalo, NY 14260 - États-Unis

2 : Observatoire du Plurilinguisme en Afrique (OPA) (OPA)

Rue 21x20 pôle urbain de Diamniadio, Dakar - Sénégal

Messages are received differently depending on the language that is used to convey them. This is especially evident in African postcolonial contexts, where each of the several languages populating local ecosystems connects to a distinct image of ideal speaker. Such language-specific images evoke specific memories, values, and expectations in multilingual individuals. In different contexts, the same language may pattern differently with, e.g., expected reliability of a message. Does this apply to chatbot interaction, too? The poster presents a proposed study to be implemented in collaboration with an existing dense network of linguists and communities in Cameroon and Senegal.

[Wolof] Làkk wi ñuy wasaaree bataaxal yi mooy firnde li yooyule bataaxal di tekki. Loolu dafay gëna fës ci jamonoy ginaaw kolonisasioŋ ci Afrique, muy làkk bu nekk ci ecosystem yu dëkk bi, muy lëkkaloo ak nataal bu wuute ci aji wax bu gën bi. Nataal yu mel noonu, yu jëm ci làkk wi, dañuy yee ci nit ñi làkk yu bari ay fàttaliku, ay valeur ak ay xaar-xaar. Ci yeneen mbir yu wuute, benn làkk mën na wuute ci anam wu wuute, lu ci melni, koolute gu ñuy seentu ci bataaxal bi. Ndax noonu la deme itam ci jëfeek chatbot ? Afiis bi dafay wane ab njàngale buñ nara amal ci lëkkaloo ak ab reso bu mak buu yengu ci mbirum làkk yi ak askan wi ci Cameroun ak Senegal.

sciencesconf.org:lt4all2025:618915

LT for revitalization of Less-Resourced Languages

Zygmunt Vetulani [1, *, @]

1 : Adam Mickiewicz University, Poznań

* : Corresponding author

Revitalizing natural language means restoring communication functionality. Its impoverishment is not limited to the Less-Resourced Languages. This may result from a decreasing speakers population or language covering (Sanskrit, Latin). In some cases, revitalization is effective (Hebrew). For extinct languages (Ainu), and relict non-verbal languages (as Silbo Gomero), restitution remains. We observe two spontaneous development trends: 1) taking over the role of lingua franca for a specific area or languages, 2) revitalization of (evolving) endangered languages. This second trend is of particular interest to UNESCO. In this presentation we discuss the most important challenges for languages poorly equipped with language technologies.

Rewitalizacja języka naturalnego to przywrócenie funkcjonalności komunikacyjnej. Jej zubożenie nie ogranicza się do języków technologicznie niedoposażonych. Bywa on skutkiem zmniejszania się populacji użytkowników lub zakresu stosowania języka (Sanskryt, Łacina). W niektórych przypadkach rewitalizacja jest skuteczna (Hebrajski). Dla języków wymarłych (Ajnu), ale także reliktowych języków niewerbalnych (Silbo Gomero), pozostają próby restytucji. Obserwujemy dwie spontaniczne tendencje rozwojowe: 1) przejęcie roli *lingua franca* dla określonego obszaru, lub języków. 2) rewitalizacja języków zagrożonych, podlegających ewolucji. Ta druga tendencja stanowi szczególny obiekt zainteresowania UNESCO. W tej prezentacji omawiamy najważniejsze wyzwania stojące przed językami słabo wyposażonymi w technologie językowe.

NüshuRescue: Reviving the Endangered Nüshu Language with AI

Ivory Yang [1,*], Weicheng Ma [1], Soroush Vosoughi [1]

1 : Dartmouth College [Hanover]

Hanover, NH 03755 USA - États-Unis

* : Corresponding author

Nüshu is a rare script historically used by Yao women in China for self-expression within a patriarchal society. As an endangered language, its revitalization is challenging due to limited resources. To address this, we introduce NüshuRescue, an AI-driven framework that trains LLMs on endangered languages with minimal data. We developed NCGold, the first 500-sentence Nüshu-Chinese parallel corpus, achieving 48.69% translation accuracy with GPT-4-Turbo using just 35 examples. NüshuRescue also generated NCSilver, a set of 98 new translations, and supports research with FastText and Seq2Seq models. Code and dataset available at: https://github.com/ivoryayang/NushuRescue.

女书是中国瑶族女性在父权社会中用于自我表达的罕见文字。作为濒危语言，其复兴因资源有限而充满挑战。为此，我们提出NüshuRescue，一个利用少量数据训练LLM的AI框架。我们构建了NCGold，首个500句女书-汉语平行语料库，并用GPT-4-Turbo在仅35个示例下实现48.69%翻译准确率。此外，NüshuRescue生成了NCSilver（98句新译文），并支持FastText和Seq2Seq模型研究。代码与数据集公开：https://github.com/ivoryayang/NushuRescue。

Nüshu est un script rare utilisé historiquement par les femmes Yao en Chine pour s'exprimer dans une société patriarcale. En tant que langue en danger, sa revitalisation est difficile en raison du manque de ressources. Pour y remédier, nous introduisons NüshuRescue, un cadre IA entraînant des LLMs avec peu de données. Nous avons développé NCGold, le premier corpus parallèle Nüshu-Chinois de 500 phrases, atteignant 48,69 % de précision avec GPT-4-Turbo sur seulement 35 exemples. NüshuRescue a aussi généré NCSilver (98 nouvelles traductions) et prend en charge FastText et Seq2Seq. Code et données: https://github.com/ivoryayang/NushuRescue.

Preserving cultural and linguistic diversity in Kyrgyzstan

Jyldyz Bakashova [1, *]

1 : Director of the National History Museum of the Kyrgyz Republic, Corresponding Member of the National Academy of Sciences of the KR, Professor, National Coordinator of Horizon Europe in the KR, Member of the KR National Commission for UNESCO (NHM KR)

Bishkek, Ala Too Square 203 - Kirghizistan

* : Corresponding author

Kyrgyzstan is a place where the West and the East meet, a crossroads of languages and cultures, a country where representatives of 83 nationalities live. The indigenous population is the Kyrgyz, who make up 73% of the total population of the country. One of the main traditional values of Kyrgyzstan is respect for cultural and linguistic diversity and national identity. The country adheres to the principles of unity, equality, values of tolerance and multiculturalism.

Кыргызстан - это место, где встречаются Запад и Восток, перекресток языков и культур, страна, где проживают представители 83 национальностей. Коренное население - кыргызы, которые составляют 73% от общей численности населения страны. Одним из главных приоритетов экономики Кыргызстана является уважение к культурному и языковому разнообразию и национальной самобытности. Страна придерживается борьбы с единством, равенством, ценностной толерантностью и мультикультурализмом.

sciencesconf.org:lt4all2025:616208

Revitalization of minority languages of Nepal

Netra Mani Rai [1]

1 : Netra Mani Rai (SIL.org)

Gokarneshwor Municipality ward no. 9, Arubari, Kathmandu 09/022 - Népal

Abstract
This presentation focuses on the current situation of Nepal's minority languages and their preservation methods and processes. In this context, SIL Nepal focuses on preserving the language and culture of indigenous Nepalese communities. Primarily focuses on the development of writing systems, story writing, and linguistic analysis utilising the latest technologies. In addition to this, the development of dictionaries, grammar writing, and literature has been prioritized. It also provides training and consultancy for contemporary linguistic activities such as conducting multilingual education and literacy classes. Furthermore, it collaborates with governmental and non-governmental organizations in linguistic research and publishes linguistic research reports and books.

tam semusi nepalbim camnabtim brʌmupo tejom musi, hampo tusi lam ka musi lampo dubi halpom gota. tam dubi, sil nepal.a nepalbim kʰamam adzʰom minumupo brʌ kirsahu nu: bita. tejom prʌbidʰi lʌpsoa lamlu tsʌpsilampo uduwa, tsuri tsʌpsi, brʌpo andʌmbi tsoisim gota. mambikajo sʌbdʌkos bʌnemusi, bjakʌrʌn tsʌpsi ka sahitjʌpo bʌrsilai tso tsonpo gota. tama bʌhubʰasi sikshja ka saksherta tsendumu tʰuksa heŋam tejom brʌpo sulamduwamupo tsemlam ka banlam jo biksa muta. mampikajo, tama sʌrkari kajo gʌirʌ sʌrkari sʌnstʰakajo tuhe duwa musoka brʌpo dubi siŋso-lʌmsoa prʌtibedʌn, dzendʌmmu laksa muta.

sciencesconf.org:lt4all2025:620585

Rising Voices Catalyst Program

Marco Martínez [1, *]

1 : Global Voices (GV)

Kranenburgweg 135 A 2583 ER The Hague Netherlands - Pays-Bas

* : Corresponding author

Rising Voices' Catalyst Program is the result of more than 14 years of learning as a network. We have listened to hundreds of speakers of Indigenous, minoritized or low-resource languages who have expressed their interest in using the Internet and digital media to promote and revitalize their languages. In Mexico, Colombia, and Guatemala, this program gathers participants with projects related to the use, strengthening, revitalization and/or promotion of an Indigenous language through digital media and tools. The Program involves peer learning, community mentors and accompaniment, and providing much needed resources to make community-based digital projects a reality.

El Programa Catalizador de Rising Voices es el resultado de más de 14 años de aprendizaje como red. Hemos escuchado a cientos de hablantes de lenguas indígenas o lenguas minorizadas que han expresado su interés en utilizar Internet y los medios digitales para promover y revitalizar sus lenguas.En México, Colombia y Guatemala, este programa reúne a participantes con proyectos relacionados con el uso, fortalecimiento, revitalización y/o promoción de una lengua indígena a través de medios y herramientas digitales. El programa incluye aprendizaje entre pares, mentoría comunitaria y acompañamiento, además de brindar recursos necesarios para hacer realidad proyectos digitales comunitarios.

The conundrum of using Wikipedia as training set data / Čuolmâ Wikipedia kevttimist tiätuamnâstuv škovliimân

Kimberli Mäkäräinen [1, 2, 3] , Jack Rueter [4] , Trond Trosterud [5]

1 : Wikimedia Finland (WMFI)

2 : Wikimedia Language Diversity Hub

3 : Wikimedia Norway (WMNO)

4 : University of Helsinki

5 : Uit The Arctic University of Norway

“Leverage the power of Wikipedia to train your LLMs!” But not all Wikipedia text is of good quality, particularly for indigenous and minority languages. Orthographical issues, multiple languages in one Wikipedia, editors from language communities without a written tradition, and editors who use various tools to write in languages they do not know all contribute to this. These latter editors contaminate datasets, which is not always evident to the people harvesting Wikipedia to train LLMs since they do not know the language(s) either. A vicious circle arises when tools based on these LLMs are used to create new Wikipedia texts.

"Ávhástâl Wikipedia stuorrâ kielâmyensterij škovliimist!" Mut puoh Wikipedia teevstâi kvaliteet ij lah šiev, eromâšávt algâaalmug- teikkâ ucceeblovokielâg Wikipediain. Toos láá maaŋgah suujah: ortografisiih čuolmah, maaŋgah kielah oovtâ Wikipediast, kevtteeh kielâsiärváduvâin, moi kulttuurân ij kuulâ čäällimärbi teikkâ -kielâ, kevtteeh, kiäh kevttih ereslágán niävuid teevstâi čälimân veikkâ iä määti kielâ. Taah majemuuh kevtteeh nyeskideh tiätuamnâstuvváid, mut taat äšši ij lah čielgâs taid ulmuid, kiäh ráápuh teevstâid Wikipediast stuorrâ kielâmyensterij škovliimân veikkâ iä sijgin määti täid kielâ(id). Ko uđđâ niävuh ráhtojeh kielâmyensterij vuáđuld, já kevtteeh rähtih uđđâ Wikipedia artikkâlijd toiguin, te ponjos juátkoo.

---

sciencesconf.org:lt4all2025:617482

Using offline digital tools to share knowledge in Indigenous Languages

Phil Malone [1]

1 : Access Agriculture

40 Rue Washington 1050 Brussels - Belgique

Access Agriculture will be showcasing how quality, practical training videos to agroecological principles can inspire communities to improve livelihoods using indigenous and local languages. By having translations in the spoken word, the videos are easily accessible to women and youth.
Teams of young “Entrepreneurs for Rural Access” are using solar powered smart projectors to provide services showing videos which improve incomes in harmony with the environment. This means effective communication, even in remote areas where internet, mobile signal and electricity connections are unreliable.

Utiliser des outils numériques hors ligne pour partager des connaissances dans les langues autochtones
Access Agriculture montrera comment des vidéos de formation pratiques et de qualité sur les principes agroécologiques peuvent inspirer les communautés à améliorer leurs moyens de subsistance en utilisant les langues autochtones et locales. Grâce à des traductions orales, les vidéos sont facilement accessibles aux femmes et aux jeunes.
Des équipes de jeunes « entrepreneurs pour l'accès rural » utilisent des projecteurs intelligents alimentés à l'énergie solaire pour fournir des services montrant des vidéos qui améliorent les revenus en harmonie avec l'environnement. Cela signifie une communication efficace, même dans les zones reculées où les connexions Internet, le signal mobile et l'électricité ne sont pas fiables.

Session O5: The Impact of Language Technology: Enhancing or Undermining Human Roles?

We don't trust those mobiles!

Steven Bird [1,*]

1 : Charles Darwin University [Australia] (CDU)

* : Corresponding author

"You want to sit down and share your knowledge with your young one, and you tell them to come, but they say 'no, I want to play Call of Duty', or whatever. You can't stop them... The families talked about putting Kunwinjku in the phones for the children to learn, but we don't like that idea. We don't trust those mobiles!" This experience reveals how mobile devices are accelerating the very problem they are intended to solve. I argue that we must stop centering language technologies, and start centering minoritised communities and their aspirations for linguistic and cultural regeneration.
« Vous voulez vous asseoir et partager vos connaissances avec votre enfant, et vous lui dites de venir, mais il dit « non, je veux jouer à Call of Duty », ou autre. Vous ne pouvez pas l'en empêcher... Les familles ont parlé d'installer Kunwinjku dans les portables pour que les enfants apprennent, mais nous n'aimons pas cette idée. Nous ne faisons pas confiance à ces téléphones portables ! » Cette expérience montre comment les appareils mobiles accélèrent le problème même qu'ils sont censés résoudre. Je soutiens que nous devons cesser de centrer les technologies linguistiques et commencer à centrer les communautés minoritaires et leurs aspirations à une régénération linguistique et culturelle.

The Role of Human Moderators in Low-Resource Languages: Kiswahili as a Case Study

Mona Elswah [1]

1 : Center for Democracy & Technology (CDT)

Content moderators play a crucial role in safeguarding the information environment and ensuring a fair online experience. However, they often work under poor conditions with inadequate training, leading to moderation errors. This is particularly evident in the Majority World, especially in low-resource languages like Kiswahili, spoken by over 100 million people in East Africa. This presentation explores the challenges faced by human moderators in Kiswahili content moderation across platforms in Kenya and Tanzania. Findings highlight the limitations of language technology and underscore the essential need for human expertise to ensure accurate and culturally relevant content moderation.

Arabic translation
غير عمل لظروف يتعرضون المحتوى مشرفي أن إلا .الإنترنت مستخدمي لكل عادلة تجربة وضمان المعلوماتية البيئة حماية في هامًا دورًا المحتوى مشرفو يلعب اللغات في وخاصة ،العالمي الجنوب دول في خاص بشكل هذا يتضح .المحتوى على الإشراف في أخطاء إلى يؤدي مما ،العمل لطبيعة كافٍ غير تدريب مع جيدة مشرفو واجهها التي التحديات نستعرض ،الحديث هذا في .إفريقيا شرق في شخص مليون ١٠٠ من أكثر بها يتحدث التي السواحيلية اللغة مثل المحدودة الموارد ذات إلى الحاجة وتوضح اللغة تقنيات محدودية إلى النتائج تشير .وتنزانيا كينيا في الاجتماعي للتواصل منصات عدة عبر السواحيلية اللغة مع يتعاملون الذين المحتوى الغربية غير اللغات في للمحتوى ثقافيًا وملائم دقيق إشراف لضمان بشريين مشرفين

The Hidden Human Labour Powering Artificial Intelligence

James Muldoon [1]

1 : University of Essex

Big Tech has sold us the illusion that artificial intelligence is a frictionless technology that will bring wealth and prosperity to humanity. But hidden beneath this smooth surface lies the grim reality of a precarious global workforce of millions that labour under often appalling conditions to make AI possible. We present the results of fieldwork at specialised data annotation facilities in East Africa that brings to light the working conditions of those often deliberately concealed from view and the systems of power that determine their future. It shows how AI is an extraction machine that churns through ever-larger datasets and feeds off humanity's labour and collective intelligence to power its algorithms.

Les grandes entreprises technologiques nous ont vendu l'illusion que l'intelligence artificielle est une technologie sans friction qui apportera richesse et prospérité à l'humanité. Mais sous cette surface lisse se cache la triste réalité d'une main-d'œuvre mondiale précaire de millions de personnes qui travaillent dans des conditions souvent épouvantables pour rendre l'IA possible. Nous présentons les résultats d'un travail de terrain effectué dans des installations spécialisées dans l'annotation de données en Afrique de l'Est, qui met en lumière les conditions de travail de ceux qui sont souvent délibérément dissimulés et les systèmes de pouvoir qui déterminent leur avenir. Il montre comment l'IA est une machine d'extraction qui traite des ensembles de données de plus en plus vastes et se nourrit du travail et de l'intelligence collective de l'humanité pour alimenter ses algorithmes.

Modern Slaves, Precariat, Proletariat: AI and Work in a Developing Country

Emmanuel Lallana [1]

1 : University of the Philippines (UP)

Quezon Hall, UP Diliman campus, Quezon City - Philippines

In this presentation I discuss how AI is affecting two types of digital workers in the Philippines: 1) online freelance workers who do ‘microwork' (the precariat) and 2) those engaged in Business Process Outspourcing - the contracting of the operations and responsibilities of a specific business process in developed country to a service provider in the country (the proletariat).

Sa presentasyong ito, tatalakayin ko kung paano naaapektuhan ng AI ang dalawang uri ng digital workers sa Pilipinas: 1) mga “online freelance workers” na gumagawa ng microwork (‘precariat') at 2) mga nasa business process outsourcing - pagkontrata ng isang proseso ng negosyo mula sa maulad na bansa sa tagapagbigay ng serbisyo na nasa Pilipinas (proletariat).

I should be the one asking the questions! Taking turn and defining roles in AI-interpreted conversations

Alice Delorme Benites [1]

1 : Université des Sciences Appliquées de Zurich

Speech technologies and the foundation models on which they are based are advancing at lightning pace in terms of accuracy and speed. The research presented here can be seen as a kind of clinical trial: what is the impact of using automatic speech translation on communication? Exploratory experiments show that AI used as an interpreter forces speakers to redefine their role in the interaction - giving greater autonomy of speech to people from linguistic minorities.

Les technologies langagières et les modèles de fondation sur lesquelles elles s'appuient enregistrent des progrès fulgurants en termes de précision et de rapidité. La recherche présentée ici se comprend comme une forme d'essai clinique : quelles conséquences entraîne l'usage de traducteurs automatiques vocaux sur la communication ? Des expériences exploratoires montrent que l'IA utilisée comme interprète force les interlocuteurs à redéfinir leur rôle dans l'interaction – conférant une meilleure autonomie de parole aux personnes issues des minorités linguistiques.

Bridging the Digital Linguistic Divide in the AI Era

Tahar Bouhafs [1, *]

1 : CSA Research M&A Practice (csa-research)

3 Kennedy Drive #901 North Chelmsford, MA , 01863 - États-Unis

* : Corresponding author

Tahar Bouhafs, CEO of CSA Research, will explore the critical issue of digital linguistic inequality in the AI era. Large language models (LLMs) rely on biased training data, favoring North American English and widening the digital divide. With 63% of the world controlling 93% of online wealth, underrepresented languages face further marginalization. This session examines whether LLMs will bridge or exacerbate disparities. Bouhafs will also discuss potential solutions to close this gap, ensuring more inclusive and equitable technological advancements across languages and communities through diverse language data, digital infrastructure investment, and local AI initiatives supported by governments, companies, and NGOs.

Tahar Bouhafs, PDG de CSA Research, abordera l'inégalité linguistique numérique à l'ère de l'IA. Les grands modèles de langage (LLMs) utilisent des données biaisées, favorisant l'anglais nord-américain et aggravant la fracture numérique. Avec 63 % de la population mondiale contrôlant 93 % de la richesse en ligne, les langues sous-représentées risquent l'exclusion. Cette session examinera si les LLMs réduisent ou accentuent ces écarts. Bouhafs discutera également des solutions pour combler ce fossé, garantissant des avancées technologiques plus inclusives et équitables via des données linguistiques diversifiées, des infrastructures numériques et des initiatives locales en IA soutenues par gouvernements, entreprises et ONG.

Humanity in the 21st Century

Brett Frischmann [1, *]

1 : Villanova University

* : Corresponding author

Humans have been shaped by technology since the dawn of time, and of course, humans have shaped other humans through technology for a very long time as well. Yet techno-social engineering of humans exists on an unprecedented scale and scope, and it is only growing more pervasive as we embed networked sensors in our public and private spaces, our devices, our clothing, and ourselves. The fine-grained, hyper-personalized, ubiquitous, continuous and environmental aspects of the techno-social engineering make the scale, scope, and consequences for humanity unprecedented. In my brief talk, I will touch on the following themes, developed more extensively in Re-Engineering Humanity (2018): * When does technology diminish our humanity? * When and how do humans become programmable? * Can we detect when this happens? How will we evaluate? * What makes us human? What about being human matters?

Poster Session P5: Language-specific Tools

sciencesconf.org:lt4all2025:617620

Advancing Language Technology for Ethiopia's Diverse Linguistic Landscape through the EthioNLP Collaborative Effort

Seid Muhie Yimam [1]

1 : Fachbereich Informatik - Informatics Department [Hamburg] (UHH)

Universität Hamburg, MIN-Fakultät, Fachbereich Informatik, Vogt-Kölln-Straße 30, 22527 Hamburg - Allemagne

---

The EthioNLP community unites global researchers dedicated to advancing language technology for Ethiopia's multilingual society, encompassing over 83 distinct languages. Previously scattered, resources are now consolidated through EthioNLP's collaborative efforts. Our focus areas include NLP corpus and dataset creation, language model building, research and collaboration, enhancing education quality, and fostering academy-to-industry linkages. We have developed models and datasets for tasks such as POS tagging, NER, sentiment analysis, hate speech analysis, machine translation, and news generation. We are currently focusing on improving and fine-tuning various large language models for the Ethiopian context, with some already published, such as EthioLLM and Walia-LLaMA, paving the way for continued innovation and integration.

ኢትዮ የተፈጥሮ ቋንቋ የሥራ ሂደት (EthioNLP) በኢትዮጵያ ውስጥ በሚገኙ ከ83 በላይ የተለያዩ ቋንቋዎች ላይ የቋንቋ ቴክኖሎጂን ለማሻሻልና ለማበልጸግ የተመሠረተ የኮምፒዩተር ሳይንስ የባለሙያዎች ማኅበር ነው። ከአሁን በፊት በዓለም ዙሪያ በተበታተነ መልኩ ይሠሩ የነበሩ ጥናቶችን እና ምርምሮችን አንድ ላይ በማሰባሰብ በአሁኑ ጊዜ በEthioNLP አማካኝነት ወጥ ሥራን ለመሥራት ችሏል። የእኛ ዋና ዋና ግቦች የቋንቋ መረጃዎችን ማሰባሰብ፣ የቋንቋ ሞዴሎችን ማበልጸግ፤ የጥናት እና ምርምር ትብብርን መፍጠር፣ የትምህርት ጥራትን ማሻሻል፣ እንዲሁም የትምህርቱን እና የኢንዱስትሪውን ዘርፍ ማስተሳሰር ይገኙበታል። እስካሁን ድረስ የቋንቋ ሞዴሎችን፣ የማሽን ትርጉም፣ ጥያቄና መልስ፣ ሐሳብን መረዳትና መተንተን፣ የነገሮችን ስም መለየት፣ የጥላቻ ንግግሮችን መለየት እናም ዜናዎችን መመርመር፣ መተንተን እና ማዘጋጀት የመሳሰሉትን ጥናቶች አካሂደናል። በአሁኑ ጊዜ በተለይ EthioLLM እና Walia-LLaMA የመሳሰሉትን የልህቀት ቋንቋዎች ማዘመን በመሥራት ላይ እንገኛለን። ይህም ሰው ሠራሽ አስተውሎትን እና የቴክኖሎጅ ፈጠራን ለማሳለጥ ጥሩ መንገድን ይከፍታል ብለን እናምናለን።

---

sciencesconf.org:lt4all2025:617171

Current State of Language Technologies in Sorbian Languages

Daniel Sobe [1] , Ivan Kraljevski [2]

1 : Foundation for the Sorbian people, Bautzen, Germany

2 : Fraunhofer Institute for Ceramic Technologies and Systems IKTS, Dresden, Germany (Fraunhofer IKTS, Dresden, Germany)

This poster gives an overview of all publicly available Upper and Lower Sorbian language technology resources, shown from the perspective of an end user.
In order to preserve areas where Sorbian can be spoken to a mixed German and Sorbian audience, creating a simultaneous translation from Sorbian to German is one of our core tasks.
To create such a system, we need to work on the individual building blocks first. The poster shows achievements at each of these three blocks speech recognition, machine translation and speech synthesis. Furthermore, results from relevant fundamental research are included.
Most of the technology and knowledge is publicly availabe. We value every contribution to these systems.

Na tutym posteru namakaće přehlad wšitkich technologiskich resursow za hornjo- a delnjoserbšćinu, kotrež su wužiwarjam zjawnje přistupne.
Zo bychu so rěčne rumy zdźerželi, w kotrychž móže so w přitomnosći Němcow dale serbować, je simultane přełožowanje ze serbšćiny do němčiny jedyn z našich jadrowych nadawkow.
Za wutworjenje tajkeho systema dyrbimy najprjedy jednotliwe technologije wuwić. Poster pokazuje wuslědki w kóždym z třoch wobłukow - spóznawanje rěče, mašinelny přełožk a synteza rěče. Nimo toho su wuslědki z přisłušneho zakładneho slědźenja zapřijate.
Najwjetši dźěl technologije a wědy steji zjawnje k dispoziciji. Ważimy sej kóždy přinošk k tutym systemam.

---

sciencesconf.org:lt4all2025:618053

Dialectal and Low-Resource Machine Translation for Aromanian

Alexandru-Iulius Jerpelea [1,*] , Sergiu Nisioi [2] , Alina Radoi [3]

1 : Tudor Vianu National College of Computer Science

Arhitect Ion Mincu 10, Bucharest, Romania - Roumanie

2 : University of Bucharest (UniBuc)

90, Panduri Street, Sector 5, 050663, Bucharest - Roumanie

3 : Universitatea de Vest din Timișoara [România] = West University of Timișoara [Romania] = Université Ouest de Timișoara [Roumanie] (UVT)

Blvd Varsile Parvan 4, Timisoara 300223, Timis - Roumanie

* : Corresponding author

This work presents the development of a neural machine translation system for English, Romanian, and Aromanian—an endangered Eastern Romance language. Our key contributions include (1) the largest Aromanian-Romanian parallel corpus to date (79,000 sentence pairs) and (2) a comparative analysis of optimized Aromanian translation models. We introduce auxiliary tools, such as a language-agnostic sentence embedding model for text mining and automated evaluation, alongside a diacritics conversion system. This research brings contributions to both computational linguistics and language preservation efforts by establishing essential resources for a historically under-resourced language. All datasets, trained models, and associated tools are public: https://arotranslate.com

Această lucrare prezintă dezvoltarea unui sistem de neural machine translation pentru engleză, română și aromână – o limbă romanică orientală pe cale de dispariție. Contribuțiile noastre principale includ (1) cel mai mare corpus paralel Aromanian-Romanian realizat până în prezent (79.000 de perechi de propoziții) și (2) o analiză comparativă a modelelor optimizate pentru traducerea în Aromanian. Introducem instrumente auxiliare, precum un model de sentence embedding independent de limbă pentru text mining și evaluare automată, alături de un sistem de conversie a diacriticelor. Această cercetare aduce contribuții atât în computational linguistics, cât și în eforturile de conservare a limbilor, oferind resurse esențiale pentru o limbă istoric sub-reprezentată. Toate seturile de date, modelele antrenate și instrumentele asociate sunt publice: https://arotranslate.com.

Dialogue Summarization and Hoax Classification for Buginese, Balinese, Minangkabau (Local Languages in Indonesia)

Ayu Purwarianti [1,*] , Mokhammad Wildan Marzuqon [2] , Reza Nawawi [2] , Agung Baptiso Sorlawan [2] , Alham Fikri Aji [3] , Dea Adhista [2] , Samuel Cahyawijaya [4] , Yusrina Sabila [2] , Aulia Adila [5] , Miftahul Mahfuzh [2]

1 : Institut Teknologi Bandung (ITB)

Jl. Ganesha no. 10, Bandung - Indonésie

2 : Prosa Solusi Cerdas (prosa)

Jl. Otten no. 10, Bandung - Indonésie

3 : Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

4 : The Hong Kong University of Science and Technology (HKUST)

5 : Japan Advanced Institute of Science and Technology (JAIST)

1-1 Asahidai, Nomi, Ishikawa 923-1292 Japan - Japon

* : Corresponding author

We presented two NLP tasks for three low resources local languages in Indonesia, namely Buginese, Balinese, and Minangkabau. We developed several datasets for Dialogue Summarization and Hoax Classification tasks. For the dialogue summarization task, we asked native speakers to write dialogues between two people based on several defined topics, along with the summary. For the hoax classification task, we collected hoax confirmation articles on climate change and asked native speakers to write various queries related with the articles. Each query is completed with its stance detection class and topic label. We conducted several experiments using various pretrain language models for several tasks: dialogue summarization, stance detection, topic and sub topic classification.

Kami melakukan penelitian terhadap dua NLP task untuk tiga bahasa daerah dengan sumber daya rendah di Indonesia, yaitu Bugis, Bali, dan Minangkabau. Kami mengembangkan beberapa set data untuk tugas Peringkasan Dialog dan Klasifikasi Hoaks. Untuk peringkasan dialog, kami meminta penutur asli untuk menulis dialog antara dua orang berdasarkan beberapa topik yang ditentukan, beserta ringkasannya. Untuk tugas klasifikasi hoaks, kami mengumpulkan artikel konfirmasi hoaks tentang perubahan iklim dan meminta penutur asli untuk menulis berbagai pertanyaan yang terkait dengan artikel tersebut. Setiap pertanyaan dilengkapi dengan kelas deteksi sikap dan label topiknya. Kami melakukan beberapa eksperimen menggunakan berbagai model bahasa pra-latih untuk beberapa tugas: peringkasan dialog, deteksi sikap, klasifikasi topik dan subtopik.

---

sciencesconf.org:lt4all2025:618003

Generative AI for communal Sámi storytelling / Generatiiva AI veahkkin muitaleame servođaga muitalusaid

Lars Ailo Bongo [1, 2] , Ernie Roby-Tomić [2] , Ánde Somby [3] , Samuel Valkeapää [2]

1 : Uit The Arctic University of Norway

2 : Sami University of Applied Sciences

3 : Uit The Arctic University of Norway

The democratization of AI offers new opportunities for indigenous people to share their stories and preserve their cultural heritage, but current generative AI models do not represent indigenous people accurately. However, by clever design it is possible to leverage AI for indigenous language applications. We have developed a Tabletop Roleplaying Game (TTRPG) that uses generative AI for communal storytelling for the indigenous Sámi people. It provides a platform that can be used to create roleplaying sessions for both traditional stories and discussion of current Sámi issues. We believe this approach can be adapted for storytelling in other indigenous communities.

AI boahtin álbmogiid olamuddui addá álgoálbmogiidda ođđa vejolašvuođaid juogadit iežas muitalusaid ja seailluhit iežas kultuvrra. Muhto dálá generatiiva AI modeallat eai doarjjo álgoálbmogiid albmaládje. AI:a čehpet hábmemin lea vejolaš veahkehit álgoálbmotgielaid geavaheami. Mii leat ráhkadan beavderollaspealu (Tabletop Roleplaying Game, TTRPG) mii geavaha generatiiva AI veahkkin muitalusaid muitaleamis. Dainna maid sáhttá ráhkadit árbevirolaš muitalusaid ja áigeguovdilis sámi áššiid vuođul rollaspealuid. Mii jáhkkit ahte dákkár lahkoneami sáhttá geavahit maid eará eamiálbmot servošiid muitalanárbevieruid oktavuođas.

---

sciencesconf.org:lt4all2025:617429

Language Resources for Guarani: How to Work with a Low Resource Language

Luis Chiruzzo [1, *]

1 : Universidad de la República [Montevideo] (UDELAR)

* : Corresponding author

We work with the Guarani language, a South American indigenous language that has not been completely explored in the NLP community. We started by building a small Guarani-Spanish parallel corpus and a few other resources like a Guarani-Spanish code switching corpus and a small Guarani wordnet. Lately, we created Guarani and Spanish feature grammars that let us build synthetic parallel text for both languages, and we show that using this text it is possible to highly improve the performance of a neural translation system, even using little real data, a method that could be extended to other low resource languages as well.

Trabajamos con el idioma Guaraní, un idioma de pueblos originarios de América del Sur, que no está muy explorado en la comunidad de PLN. Comenzamos construyendo un pequeño corpus paralelo Guaraní-Español, y otros recursos como un corpus de code-switching y un pequeño wordnet en guaraní. En nuestro último trabajo, desarrollamos gramáticas de rasgos para el guaraní y el español, que nos permitieron crear texto sintético para las dos lenguas, y mostramos que es posible usar este texto para mejorar los resultados de un modelo de traducción neuronal, incluso teniendo pocos datos reales, este método podría utilizarse también en muchos escenarios de lenguas con pocos recursos.

sciencesconf.org:lt4all2025:617222

Machine Translation Literacy Project: Building Blocks for Empowering Communities

Lynne Bowker [1]

1 : Université Laval (UL)

2325, rue de l'Université Québec (Québec) G1V0A6 - Canada

Machine translation tools hold enormous potential for breaking down language barriers, but they are not foolproof. Communities must therefore develop machine translation literacy skills to become responsible and savvy tool users. The Machine Translation Literacy Project provides free resources to empower people to perform risk assessment, to understand the relevance of data and the importance of transparency, to safeguard their privacy, and to interact effectively with tools to get better results. These resources are available at:
https://sites.google.com/view/machinetranslationliteracy/

Les outils de traduction automatique ont un potentiel énorme pour briser les barrières linguistiques, mais ils ne sont pas infaillibles. Les communautés doivent donc développer une approche raisonnée de la traduction automatique pour se transformer en utilisateurs responsables et avertis. Le projet « Machine Translation Literacy Project » fournit des ressources gratuites pour permettre aux gens d'effectuer une évaluation des risques, de comprendre la pertinence des données et l'importance de la transparence, de protéger leurs données sensibles et de travailler de manière efficace avec les outils afin d'obtenir de meilleurs résultats. Ces ressources sont disponibles à l'adresse suivante :
https://sites.google.com/view/machinetranslationliteracy/

sciencesconf.org:lt4all2025:620384

NLP for Ethnic Languages of Myanmar

Win Pa Pa [1, *]

1 : Naypyitaw State Polytechnic University

* : Corresponding author

Myanmar is home to 135 officially recognized ethnic groups, each with its own language. Many of these languages remain largely underrepresented in technological advancements. Our research and development efforts in Machine Translation for the Wa, Chin, Lisu, Shan, Mon, and Myanmar languages are presented.

In "**Mon**" Language
ကောန်ချူင့်အရင့်မတန်တံဒုံင့် ပ္လဲကွိုန်ချူင့်ဗ္ဘာဏာအ် နကိုဘာသာညးကိုညး ဟိုဂးစကာရပ်စပ်မံင့် အရေဝ်ဘာသာကောန်ချူင့်အရင့်တအ် မန္ဒံ (၁၃၅)ဂကူရ ။ သုင့်စော်စကာ အရေဝ်ဘာသာဂတာပ်ခေတ် ကျော်ဒီဂျေတ်တေလ်ဗွဲမဂိုင်္ဂတအ်ဂ့် လဇုလဇုံင့်အိုတ်ရ။ ပ္လဲလိက်စံင့်ဝွံဏာအ် ကောန်ချူင့်အရင့် မန္ဒံပ္လဲကွိုန်ချူင့်ဏာအ် မပ္တံကို ဂကူ "o" ၊ ချေင့် ၊ လီသူ ၊ သေမ် ကျော် မန် တအ်ဂ့် ပါလုပ်ဒုံင့် ပ္လဲလွေင့်ကမ္ဘောန် NLP တအ်ရ။

sciencesconf.org:lt4all2025:617052

Text-to-speech synthesis in Walloon, an endogenous language of Belgium

Philippe Boula De Mareüil [1, *] , Felipe Espinosa [1], Marc Evrard [1]

1 : Laboratoire Interdisciplinaire des Sciences du Numérique

Université Paris-Saclay, Centre National de la Recherche Scientifique

* : Corresponding author

This study describes a text-to-speech synthesis system for Walloon, a Gallo-Romance language spoken in parts of Belgium and France. The system uses recordings of a translation of *The Little Prince*, read entirely (156 minutes) and partly (18 minutes) by two speakers. The corpus was transcribed into phonemes by a rule-based grapheme-to-phoneme converter. The synthesis system is based on adversarial learning (VITS), and several models were trained: with or without grapheme-to-phoneme conversion, using or not a fine-tuned model pre-trained on French data. Perceptual evaluation was conducted with Walloon speakers. Results suggest that the models using French data are only preferred in the 18-minute corpus training condition.

Ci studia cial discrît ene sinteze do pårlaedje pol walon, on lingaedje galiroman djåzé e l' Beldjike ey e l' France. Li sistinme est rashiou so des eredjistrumints d' on ratournaedje do *Ptit Prince*, léjhou e-n etir (156 munutes) u djusse les prumîs tchaptrês (18 mn). Li coirpusse fourit riscrît a foninmes pa on cvierseu grafinme-foninme erîlé. Li sinteze est rashiowe so l' aprindaedje afrontiveus (VITS) emey sacwants modeles : avou u sins cvierseu grafinme-foninme, eyet e-n eployant u nén on modele neuronike dedja etrinné pol francès. Al difén, on dmanda a des waloncåzants di noter les rzultats. I shonne ki l' sistinme pre-etrinné n' est meyeu **k' avou** l' waire d' eredjistrumints (18 munutes).

Uyghur in Smart Communication Tools

Abduweli Ayup [1, *]

1 : Uyghur Hjelp

* : Corresponding author

Smart communication tools have significantly influenced Uyghur language usage and attitudes. Early mobile tools, like pagers, lacked Uyghur support, forcing users to communicate in Chinese. This reinforced the perception that Uyghur was unsuitable for modern communication. Later, in the digital era, Uyghur speakers faced exclusion from essential services and increased scrutiny for using their language. These challenges made Uyghur seem impractical and risky in digital spaces. This presentation examines the impact of smartphones on Uyghur from 2001 to 2021, highlighting how technology, policies, and social pressures have reshaped linguistic behavior and identity.

ئەقىلفونلاردا ئۇيغۇر تىلى

ئابدۇۋەلى ئايۇپ

ئۇيغۇر تىلىنىڭ ئىشلىتىلىش دائىرەسى، پۇرسىتى ۋە تەرەققىياتى سىياسىي، ئىقتىسادى ۋە ئىجتىمائىي ئامىللاردىن باشقا تېخنىكا ئامىلىنىڭمۇ تەسىرىگە ئۇچراپ كەلدى. بۇ سۈنئىمدا ئۇيغۇر تىلىنىڭ چاقىرغۇ دەۋرى، كونۇپكا تېلىفون دەۋرى ۋە ئەقىلفون دەۋرى يورۇتۇلىدۇ. بولۇپمۇ ئەقىلفون دەۋرىدە ئۇيغۇر تىلىنىڭ قانداق خىرىسلارغا دۇچ كەلگەنلىكى، تېلىفون تەكشۈرۈش، نازارەت يۈمتاللىرى قاتارلىقلارنىڭ كىشىلەرنى تىل ۋە يېزىقتىن ۋاز كېچىشكە سەۋەپ بولۇۋاتقانلىقى ئوتتۇرىغا قويۇپ تەدبىرلەر بايان قىلىنىدۇ.

Session O6: Multilingual Communication: AI-based Translation vs Lingua Franca?

sciencesconf.org:lt4all2025:617743

AI-Based Translation of All of the World's Languages: Realistic Goal or Far-Fetched Dream?

Katharina Von Der Wense [1, 2]

1 : Johannes Gutenberg University Mainz (JGU Mainz)

Staudingerweg 9 55128 Mainz - Allemagne

2 : University of Colorado Boulder (CU Boulder)

In this talk, we will discuss how close we are to having widely available, high-performing machine translation systems for Indigenous languages of the Americas. First, we will go over the findings of the AmericasNLP 2021-2024 shared tasks on machine translation systems for Indigenous languages. Second, we will highlight concerns members of Indigenous communities have with regards to the development of AI-based translation systems for their languages.

En esta charla, discutiremos qué tan cerca estamos de contar con sistemas de traducción automática de alto rendimiento y ampliamente disponibles para las lenguas indígenas de las Américas. En primer lugar, analizaremos los resultados de las tareas compartidas de AmericasNLP 2021-2024 sobre sistemas de traducción automática para lenguas indígenas. En segundo lugar, destacaremos las inquietudes que tienen los miembros de las comunidades indígenas con respecto al desarrollo de sistemas de traducción basados en IA para sus lenguas.

Reform of research assessment, lingua franca, and AI-based translation

Janne Pölönen [1]

1 : Federation of Finnish Learned Societies (TSV)

Kirkkokatu 6, 00170 Helsinki - Finlande

The Coalition for Advancing Research Assessment (CoARA) supports quality-focused evaluation practices that recognize diverse research contributions irrespective of language and move away from discriminating metrics. Using a lingua franca enhances global communication yet risks compromising fairness, inclusivity and outreach of science. Although translation quality for less-resourced languages poses difficulties, AI-based technologies can help reduce linguistic obstacles for researchers and the public alike. However, neither a common language nor AI-based translations can substitute for the creation and dissemination of knowledge in various languages. Multilingualism is crucial for the advancement of both science and society.

Tutkimuksen arviointia edistävä yhteenliittymä (CoARA) tukee laatuun keskittyviä arviointikäytänteitä, jotka huomioivat erilaiset tutkimustuotokset kielestä riippumatta ja välttävät syrjivien mittareiden käyttöä. Lingua francan käyttö edesauttaa globaalia viestintää, mutta vaarantaa myös oikeudenmukaisuuden, osallisuuden ja tavoittavuuden tieteessä. Vaikka käännösten laatu vähemmän resursoiduissa kielissä on haaste, tekoälypohjaiset teknologiat voivat auttaa alentamaan kielimuureja tutkijoiden ja yleisön keskuudessa. Kuitenkaan lingua franca tai tekoälyyn perustuvat käännökset eivät voi korvata tutkimustiedon tuottamista ja viestimistä eri kielillä. Monikielisyys on välttämätöntä sekä tieteen että yhteiskunnan edistymiselle.

Inventory and comparisons of all methods for the measure of languages online: implications for the Internet's lingua franca

Daniel Pimienta [1]

1 : Observatoire de la diversité linguistique et culturelle dans l'Internet (OBDILCI)

Seven existing approaches for data about languages online are summarized and parameters exposed (institution, number of languages, method, updates frequency, peer-review). Bias's analysis tells that English presence is within 20%-27% and multilingualism has become the characteristic of the Internet. Presence online of some 750 languages allows more than 95% of world population to interact with L1 or L2. Yet, this is only 10% of the wealth of languages; the challenge remains for respecting the right of all languages to be digital. The rate of multilingualism of the www, a key unknown parameter, may have crossed its equivalent for Human beings (1.44). The lingua franca of the Internet is now translation boosted by IA.

Sept approches existantes pour la proportion des langues en ligne sont résumées avec paramètres (institution, nombre de langues, méthode, mises à jour, évaluation par les pairs). L'analyse des biais conclut en une présence de l'anglais entre 20% et 27 %, le multilinguisme caractérise l'Internet aujourd'hui. La présence en ligne d'environ 750 langues permet à plus de 95 % de la population mondiale d'interagir. Cependant, il ne s'agit que de 10% de des langues ; le défi reste dans le respecter du droit de toutes les langues au numérique. Le taux de multilinguisme du www, un paramètre clé inconnu pourrait avoir dépassé son équivalent humain (1,44). La lingua franca d'Internet est désormais la traduction boostée par l'IA.

sciencesconf.org:lt4all2025:617162

Reliance on Lingua Franca: kulaktan kulağa, Stille Post, tichá pošta, rikkinäinen puhelin, χαλασμένο τηλέφωνο, téléphone sans fil in humanitarian crises

Aimee Ansari [1]

1 : CLEAR Global

In humanitarian crises, effective communication requires a combination of AI and human expertise to ensure language access for all. This panel explores how AI-driven translation and speech recognition can be enhanced through localized, human-led assessments tailored to specific language needs. By analyzing language gaps per location and addressing marginalized languages, we can develop more inclusive communication strategies. Participants will learn how CLEAR Global integrates AI technology with human validation to bridge linguistic divides. Through real-world examples, we will showcase innovative solutions that adapt to diverse linguistic landscapes, ensuring communities receive vital, life-saving information.

Dans les crises humanitaires, une communication efficace nécessite un combinaison d'IA et d'expertise humaine pour assurer un accès aux langues pour tous. Ce panel explore comment la traduction et la reconnaissance vocale basées sur l'IA peuvent être améliorées grâce à des évaluations localisées et dirigées par des humains, et adaptées aux besoins linguistiques spécifiques. En analysant les lacunes linguistiques par localisation et en nous intéressant aux langues marginalisées, nous pouvons développer des stratégies de communication plus inclusives. Les participants découvriront comment CLEAR Global intègre la technologie de l'IA à la validation humaine pour combler les fossés linguistiques. À travers des exemples concrets, nous présenterons des solutions innovantes qui s'adaptent à des paysages linguistiques diversifiés, garantissant ainsi que les communautés reçoivent des informations essentielles et vitales.

Machines against Lingua Franca

Ondřej Bojar [1]

1 : Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL)

Malostranske nam. 25, Praha 1, 11800 - République tchèque

In the talk, I will present the achievements of thr EU project ELITR which delivered highly multilingual speech translation system and a more recent experiment with remote calls over the language barrier. Our subjects were speaking mutually unintelligible languages and had to rely on automatic speech translation. Assuming shared interest in running the communication, the technical solution avoiding the need for a lingua franca works, save for a few necessary features.

Poster Session P6: Language Technologies and the People

21st Century Tools for Indigenous Languages -- A Canadian Experience in Co-creating Language Technology

Antti Arppe [1, 2]

1 : University of Alberta

2 : 21st Century Tools for Indigenous Languages (21C)

In response to the well-documented threats to Indigenous cultures and languages, our Partnership project "21st Century Tools for Indigenous Languages" has worked since 2014 to support the revitalization and sustained daily use of Indigenous languages spoken in Canada and elsewhere in North America by developing modern digital tools and resources for these languages. Crucially, this work has been pursued in collaboration with these language communities, speaking Algonquian (Plains Cree, Odawa, Northern East Cree, Blackfoot, Arapaho), Dene (Tsuut'ina, Upper Tanana) and isolate languages (Haida), so that the development of language technological applications has been driven by their needs.

---

sciencesconf.org:lt4all2025:618029

Blocked by BigTech - How big tech makes using indigenous languages impossible

Sjur Moshagen [1, *] , Brendan Molloy

1 : University of Tromsø (UiT)

The Arctic University of Norway, PO Box 6050, Langnes 9037 Tromsø - Norvège

* : Corresponding author

The Divvun group at UiT the Arctic University of Norway, Tromsø, has developed language technology for the Sámi languages for more than 20 years. From the beginning, integrating tools with the software and systems that the speakers use has been a challenge, and the challenge has not diminished over the years. The poster gives illustrative examples of the challenges we have met, and also describes the reality today. The poster serves as a technical background for a presentation in Session 8.

NYNORSK (nn):
Blokkert av BigTech - Korleis bigtech gjer det uråd å bruka urfolksspråk
Divvun-gruppa ved UiT Noregs arktiske universitet, i Tromsø, har utvikla språkteknologi for dei samiske språka i meir enn 20 år. Heilt frå starten var det ei utfordring å integrera verktøya med dataprogram og system som språksamfunna bruker, og utfordringa har ikkje vorte mindre med åra. Plakaten gjev illustrative døme på utfordringane vi har møtt, og skildrar òg røynda i dag. Plakaten fungerer som teknisk bakgrunn for ein presentasjon i økt 8.

---

sciencesconf.org:lt4all2025:617455

Case Studies of Philippine Government-Industry-Academia-Language Communities Collaboration Scheme

Nathaniel Oco [1]

1 : De La Salle University [Manila] (DLSU)

2401 Taft Avenue, Manila 1004, Philippines - Philippines

I present several case studies of funded projects in the Philippines, focusing on collaborative efforts between government, industry, academia, and language communities. The Department of Science and Technology (DOST), a government agency responsible for coordinating science and technology-related projects in the country, employs the 6Ps framework to quantitatively assess the outputs of programs and initiatives. The 6Ps framework evaluates projects based on six key dimensions: Publication, Patent, Product, People Service, Places and Partnership, and Policies. Using this framework, I analyze and present case studies of DOST-supported projects to demonstrate their impact and the roles of various stakeholders (government, industry, academia, and language communities).

Ilalahad ko ang ilang case study ng mga pinondohang proyekto sa Pilipinas na nakatuon sa kolaborasyon ng pamahalaan, industriya, akademya, at mga komunidad. Ang Department of Science and Technology (DOST), isang ahensiya ng pamahalaan na nangangasiwa sa mga proyektong may kaugnayan sa agham at teknolohiya sa bansa, ay gumagamit ng 6Ps framework upang masukat nang kwantitatibo ang mga output ng mga programa at inisyatiba. Sinusuri ng 6Ps framework ang mga proyekto batay sa anim na pangunahing salik: Publication, Patent, Product, People Service, Places and Partnership, at Policies. Gamit ang framework na ito, aking ilalahad ang case study ng mga proyektong sinuportahan ng DOST upang ipakita ang kanilang ambag at ginampanang papel ng iba't ibang stakeholder (pamahalaan, industriya, akademya, at mga komunidad ng wika).

---

sciencesconf.org:lt4all2025:617542

Collaborative Language Technologies for Empowering Indigenous Communities in a Digital Age

Umarani Pappuswamy [1, @]

1 : CENTRAL INSTITUTE OF INDIAN LANGUAGES (CIIL)

CENTRAL INSTITUTE OF INDIAN LANGUAGES, Ministry of Education - Inde

India, a hub of linguistic and cultural diversity, faces the dual challenges such as language endangerment and digital invisibility. Languages like Ruga spoken in Meghalaya, with only three speakers, remain under-documented. In the digital era, cyberspace visibility is key to survival and empowerment. Yet most indigenous languages lack digital resources. Collaborative Language Technologies (CLT) offer a transformative approach to preserve these languages. CLT can enhance digital storytelling through community-led films and TV programs. Key initiatives include talking dictionaries, gamified revitalisation apps, and AI-driven oral history preservation. This proposal addresses ethical, infrastructural, and policy challenges to integrating CLT into India's language technology ecosystem.

மொழியியல் மற்றும் கலாச்சார பன்முகத்தன்மையின் மையமாக உள்ள இந்தியா, பழங்குடி மொழிகளின் வேகமான அழிவு மற்றும் அம்மொழிகளின் மின்னணு வெளிப்பாட்டின்மை போன்ற இரட்டை சவால்களை சந்திக்கிறது. மேகாலயாவில் மூன்றே நபர்களால் பேசப்படும் ருகா போன்ற மொழிகள் போதியளவு பதிவு செய்யப்படவில்லை. இன்றைய கணிணியுகத்தில், 'இணையத்தில் காட்சி பெறுதல்' என்பது மொழி உயிர்வாழ்தல் மற்றும் மேம்படுத்தலுக்கும் இன்றியமையாததாக உள்ளது. இருந்தபோதிலும், இந்தியாவில் உள்ள பெரும்பாலான பழங்குடி மொழிகளில் மின்னணு வளங்கள் இல்லை. இந்த மொழிகளை பாதுகாப்பதற்கான மாற்று தீர்வாக கூட்டு மொழி தொழில்நுட்பங்கள் (CLT) அமைகின்றன. இவை பழங்குடி மொழிகளை உள்ளடக்கிய திரைப்படங்கள் மற்றும் தொலைக்காட்சி நிகழ்ச்சிகள் போன்ற சமூகம்-தலைமையிலான முயற்சிகள் மூலம் மின்னணு கதைசொல்லலை மேம்படுத்த முடியும். பேசும் அகராதிகள், கேமிஃபைட் மொழி மறுமலர்ச்சி பயன்பாடுகள், வாய்வழி வரலாறுகளைப் பாதுகாக்க செயற்கை நுண்ணறிவு-இயங்கும் தளங்கள் ஆகியவை புதிய திட்டங்களில் அடங்கும். கூட்டு மொழி தொழில்நுட்பங்களை மொழித் தொழில்நுட்பச்சூழல் அமைப்பில் ஒருங்கிணைப்பதற்கான கொள்கைப்பரிந்துரைகளும் வழங்குப்பட்டுள்ளன.

Digital Initiatives for Indigenous Languages

Genner Llanes-Ortiz [1]

1 : Bishop's University [Sherbrooke, Canada]

2600 College St., Sherbrooke, QC J1M 1Z7 - Canada

Le kaartel je'ela' ku ts'áaik jump'éel tuukul yóok'ol u nu'ukulil “Diijital meyajo'ob tia'al u muuk'ankunsik le máasewal t'aano'obo'” ts'íibta'an tumen Genner Llanes-Ortiz yéetel u yáantajil Global Voices yéetel UNESCO. Le pikil ju'una' meenta'abi' yéetel u tsolxikinil Eddie Avila yéetel le máaxo'ob ku t'aniko'obo' máasewal t'aano'ob ku táakpajal tu múuch' tsikbal ichil Rising Voices. Le pikil ju'uno' ku ye'esik oocho u p'éel noj bejo'ob ti'al u meyajtik yéetel diijital nu'ukulo'ob ku bin yóok'ol tumen le máasewáal wíiniko'ob te yóok'ol kaabe'. Le kaartel ku ye'esik tu'ux táan u meyajo'ob yéetel le pikil ju'uno' ti'al u ka'ansik máak, yéetel ba'ax maas u binetik k-xak'altik tia'al u jóo'sik béentaja ti' nu'ukul diijital.

This poster offers an overview of the Global Voices/UNESCO toolkit titled “Digital Initiatives for Indigenous Languages,” authored by Genner Llanes-Ortiz. The toolkit was developed with the guidance and support of Eddie Avila and the Rising Voices networks of Indigenous language advocates. The toolkit discusses eight key approaches that Indigenous language users and signers worldwide have adopted to leverage digital technologies. The poster will also showcase the practical applications of the toolkit and will also identify potential areas for further development.

sciencesconf.org:lt4all2025:617257

For All? Including Minorities within the Minority: the Irish ABAIR Experience / Do Gach Éinne? ABAIR ag Freastal ar Mhionlaigh sa Mhionlach

Ailbhe Ní Chasaide [1]

1 : Phonetics and Speech Lab, Trinity College Dublin

Trinity College Dublin, Dublin 2, Ireland - Irlande

Challenges and approaches to developing speech technologies, adequate to the language communities and end-users, are discussed in the light of ABAIR's work on Irish (Gaelic). As often in minority languages, Irish has no spoken standard, but rather 3 distinct dialects: TTS/ASR systems must provide for them. Linguistic resources were developed, to capture the linguistic structure (very different from English) and cross-dialect differences. Requirements for the most critically-needed applications were prioritised: educational applications, central to language transmission; applications for those disabilities, a minority within the minority. Community partnership and close collaboration with end-user groups guide all stages of development.

Pléitear saothar ABAIR (don Ghaeilge) i gcomhthéacs na ndúshlán i bhforbairt teicneolaíochtaí atá oiriúnach do phobail na teanga agus d'úsáideoirí. Mar is minic do mhionteanga, níl canúint ‘chaighdeánach' amháin ann, ach trí mhórchanúint: caithfidh na córais sintéise/aitheanta cainte freastail orthu. Forbraíodh acmhainní teangeolaíochta, chun struchtúr na teanga (atá an-difriúil ón Bhéarla) agus na difríochtaí idir canúintí a léiriú sa teicneolaíocht. Dírítear tosaíochtaí na forbartha ar na riachtannais is práinnigh: áiseanna oideachasúla, atá tábhachtach do chaomhnú na teanga; áiseanna dóibh siúd atá faoi mhíchumas – mionlach sa mhionlach. Tá comhpháirtíocht pobail agus comhoibriú le saineolaithe/úsáideoirí na n-áiseanna lárnach i ngach gné forbartha.

sciencesconf.org:lt4all2025:616543

How globalization, superdiversity, diaspora, and increasing multilingual patterns need to impact the future of Language Technology

Mark E. Karan [1]

1 : SIL Global

SIL Global 7500 W. Camp Wisdom Road Dallas, TX 75236-5629 USA Telephone: +1 972 708-7300 - États-Unis

---

Current globalization, superdiversity, diaspora, and increasing multilingual patterns need to be considered in the creation of the Language Technologies. Policies, theories, solutions, and best-practices used in the past are no longer adequate.
Practices that need to be reconsidered include:

- Language development policies and practices
- Expectations concerning the loss of smaller languages
- Patterns of choosing a one-language context for any app or solution
- As languages are social constructs, any language technology related activity or initiative must incorporate the social change planning intrinsic to its acceptance and application

The Future of Language Technology needs to be built in consideration of today's and tomorrow's sociolinguistic realities and ecologies.
La mondialisation actuelle, la superdiversité, la diaspora et les modèles multilingues croissants doivent être pris en compte dans la création des technologies linguistiques. Les politiques, théories, solutions et meilleures pratiques utilisées dans le passé ne sont plus adéquates.

Les pratiques qui doivent être reconsidérées comprennent :

- Politiques et pratiques de développement linguistique
- Attentes concernant la perte de langues plus petites
- Modèles de choix d'un contexte dans une seule langue pour toute application ou solution
- Les langues étant des constructions sociales, toute activité ou initiative liée à la technologie linguistique doit intégrer la planification du changement social intrinsèque à son acceptation et à son application.

L'avenir des technologies linguistiques doit être construit en tenant compte des réalités et des écologies sociolinguistiques d'aujourd'hui et de demain.

---

sciencesconf.org:lt4all2025:618924

Human expertise powered by AI

Ildikó Horváth [1, *]

1 : Translation Centre For the Bodies of the EU (EU_CdT)

12E, rue Guillaume Kroll L-1882 Luxembourg - Luxembourg

* : Corresponding author

The Translation Centre is strongly committed to implementing state-of-the-art technology and processes. Fully conscious of the impact of AI-based language technologies in today's world, we endeavour to incorporate this technology into our business operations as far as possible, and continue providing our partners with enhanced and modern services.
It is essential to underline that in our view the power of machines cannot be overstated. However, the human element remains irreplaceable. Although the Centre is using and developing state-of-the-art language technology, it is taking a hybrid approach to machine translation, keeping humans in the loop.

A Fordítóközpont elkötelezett a legkorszerűbb technológia és folyamatok megvalósítása mellett. A MI-alapú nyelvi technologiak mai világra gyakorolt hatásának teljes tudatában törekszünk arra, hogy ezt a technológiát a lehető legnagyobb mértékben beépítsük üzleti tevékenységeinkbe, és folyamatosan korszerű szolgáltatásokat nyújtsunk partnereinknek.
Fontos hangsúlyozni, hogy véleményünk szerint a gépek erejét nem lehet túlbecsülni. Az emberi elem azonban pótolhatatlan marad a fordítás folyamata során. Jóllehet a Központ a legkorszerűbb nyelvi technológiát használja és fejleszti, hibrid megközelítést alkalmaz a gépi fordítás területén, melynek során a több terület szakembereinek szerepe nélkülözhetetlen.

---

sciencesconf.org:lt4all2025:617185

Small Data, Big Dreams: Lessons Learned from the Faroese Centre for Language Technology

Barbara Scalvini [1,*], Iben Debess [1,*], Dávid í Lág [1]

1 : University of the Faroe Islands

14 J.C. Svabos gøta, Tórshavn 100, Îles Féroé - Îles Féroé

* : Corresponding author

Developing language technology for low-resource languages presents numerous challenges, including a shortage of high-quality data, inadequate computational infrastructure, and limited expertise—from skilled annotators to engineers. These challenges often force low-resource communities to rely on international partners and major tech companies, thereby restricting their influence over how their languages are represented in the digital realm. The Language Technology Centre at the University of the Faroe Islands addresses these issues holistically by focusing on sustainable, open-source technological development, fostering continuous community dialogue, and assessing innovative data annotation strategies. This poster outlines the Centre's key initiatives for advancing Faroese language technology.

At menna máltøkni til smá mál hevur við sær nógvar avbjóðingar. Millum annað vantar dátugrundarlag av góðari góðsku, tøknilig undirstøðukervi eru ófullfíggjað, og staðbundin serfrøði er avmarkað í øllum frá mál- og tekstviðgerð til verkfrøði. Hesar avbjóðingar noyða ofta smá málsamfeløg at dúva upp á altjóða samstarvsfelagar og tøknirisar. Hetta avmarkar møguleikarnar at ávirka, hvussu egið mál verður umboðað í talgilda rúminum. Máltøknidepilin á Fróðskaparsetri Føroya bøtir um hesar umstøður við at arbeiða fyri burðardyggari og opnari menning av tøkni, fremja framhaldandi samfelagssamskifti og meta um nýhugsandi dátuviðgerðarhættir. Henda plakat lýsir høvuðsátøkini hjá Máltøknideplinum at menna føroyska máltøkni.

---

sciencesconf.org:lt4all2025:616955

Testing Māori Data Sovereignty Principles with AI models

Te Taka Keegan [1]

1 : University of Waikato [Hamilton]

Hillcrest, Hamilton 3216 - Nouvelle-Zélande

Generative AI technologies like Claude, ChatGPT, and Gemini demonstrate concerning proficiency in te reo Māori, the indigenous language of Aotearoa New Zealand. This capability raises alarms as Large Language Models appear to have been developed without consent or involvement from Māori people, the languages' guardians, representing a breach of Māori Data Sovereignty and a form of data colonization. Through testing various open-source generative LLM models, we explored whether Māori Data Sovereignty principles could be maintained in the Generative AI era. Our findings indicate that Māori AI tools can be developed in a Māori Data sovereignty environment.

Kua eke panuku ngā hangarau AI whēnei i a Claude, ChatGPT, me Gemini ki te kōrerorero ki te reo Māori, te reo taketake o Aotearoa. Ka ara ake te āwangawanga nui i te mea ko ēnei taputapu reo Māori kua hangaia i te kore whakaaetanga me te ārahi-kore o tētehi iwi Māori, o ngā kaitiaki o te reo. He takahitanga tēnei i te Mana Raraunga Māori, he momo colonization raraunga anō hoki. Mā te whakamātau i ngā tauira LLM matawhānui, i tūhura mātou kia kite mehemea e taea ana te hanga AI e mau tonu ana i ngā mātāpono o te Mana Māori motuhake, o te mana Māori Raraunga. E whakapono ana mātou he huarahi e taea ana e mātou te takahi.

sciencesconf.org:lt4all2025:619984

The Welsh Government's work on Welsh language technology, and how this could help your language. Gwaith Llywodraeth Cymru ar dechnoleg Gymraeg, a sut y gallai hyn helpu dy iaith di.

Gareth Morlais [1] , Dr Jeremy Evas, Dr Indeg Marshall

1 : The Welsh Government

---

1. To increase daily use of my language
2. Translation; Voice; AI
3. Suitably permissive licensing –training data for all.
4. Reduce friction - we switched Microsoft 365 to Welsh in 379 Welsh schools. Then 78,086 children woke up to Welsh by default.
5. Reflect the way we speak - audio transcribers and synthetic voices Bangor University have developed can just switch language mid-sentence without having to load a new module or voice.
6. Collaborate - handwriting recognition research by OpenAI with the National Library of Wales so languages with a handwritten history can be understood and digitised.
7. Think big - as part of our partnership with Microsoft, we have collaborated to create a simultaneous interpretation facility in Microsoft Teams meetings. .

https://www.gov.wales/welsh-language-technology

1. I gynyddu defnydd dyddiol fy iaith
2. Cyfieithu; Llais; AI
3. Trwyddedu caniataol addas –data hyfforddi i bawb.
4. Lleihau ffrithiant - fe wnaethon ni droi Microsoft 365 i'r Gymraeg mewn 379 o ysgolion Cymraeg. Yna fe ddeffrodd 78,086 o blant i'r Gymraeg yn ddiofyn.
5. Adlewyrchu'r ffordd rydym yn siarad – mae trawsgrifiwr sain a lleisiau synthetig Prifysgol Bangor yn gallu newid iaith ganol brawddeg heb orfod llwytho modiwl neu lais newydd.
6. Cydweithio - ymchwil cydnabyddiaeth llawysgrifen gan OpenAI gyda Llyfrgell Genedlaethol Cymru fel y gellir deall a digideiddio ieithoedd sydd â hanes llawysgrifen.
7. Amdani! -fel rhan o'n partneriaeth ni â Microsoft, rydyn ni wedi cydweithio i greu cyfleuster cyfieithu ar y pryd mewn cyfarfodydd Microsoft Teams

https://www.llyw.cymru/technoleg-cymraeg

Session O7: How to Properly Evaluate Language Technologies: Addressing Challenges in Quality, Reliability, Usability, and Ethical Frameworks

Platform of digital language resources for the indigenous languages of Latin America: Its creation and the challenges for indigenous languages in the development of language tools

Daniel Héctor Prado [1]

1 : Daniel Héctor Prado

English:

Platform of digital language resources for the indigenous languages of Latin America: Its creation and the challenges for indigenous languages in the development of language tools

After a thorough inventory of the digital language resources available for the indigenous languages of Latin America, the author proposes the creation of a platform for the dissemination of these resources, as well as the cooperation and provision of tools for the development of applications for the indigenous languages of the region.

This paper addresses the challenges in terms of quality, reliability, usability and ethical frameworks involved in both the development of such applications and the use of these tools.

Español:

Plataforma de recursos lingüísticos digitales para las lenguas indígenas de América Latina: Su creación y los desafíos para las lenguas indígenas en la elaboración de herramientas lingüísticas

Tras realizar un minucioso inventario de los recursos lingüísticos digitales disponibles para las lenguas indígenas de América Latina, el autor plantea la creación de una plataforma destinada a la difusión de estos recursos, así como a la cooperación y provisión de herramientas para el desarrollo de aplicaciones dirigidas a las lenguas indígenas de la región.

Esta comunicación aborda los desafíos en términos de calidad, fiabilidad, usabilidad y marcos éticos que conlleva tanto el desarrollo de dichas aplicaciones como la utilización de estas herramientas.

---

sciencesconf.org:lt4all2025:619126

Towards Reliable and Effective Multilingual Evaluation for all Languages of the World

Sunayana Sitaram [1]

1 : Microsoft Research India [Bangalore]

"Vigyan", #9, Lavelle Road, Bangalore 560 001, India - Inde

Evaluating Artificial Intelligence systems rigorously and reliably is essential to understand their capabilities, limitations, and the progress of the field. Multilingual evaluation with extensive coverage of different languages and cultures across the world is important to ensure that advancements in language technologies do not exacerbate the digital divide. This talk will outline the current state of multilingual evaluation, highlighting significant challenges such as language and task coverage, English-centric datasets, and evaluation methodology. Additionally, it will discuss recent developments in alternative evaluation methods, including community-centered participatory design and the use of AI to enhance multilingual evaluation efforts.

ಶೀರ್ಷಿಕೆ: ಪ್ರಪಂಚದ ಎಲ್ಲಾ ಭಾಷೆಗಳಿಗೆ ವಿಶ್ವಾಸಾರ್ಹ ಮತ್ತು ಪರಿಣಾಮಕಾರಿ ಬಹುಭಾಷಾ ಮೌಲ್ಯಮಾಪನದ ಕಡೆಗೆ

ಕೃತಕ ಬುದ್ಧಿಮತ್ತೆ ವ್ಯವಸ್ಥೆಗಳನ್ನು ಕಟ್ಟುನಿಟ್ಟಾಗಿ ಮತ್ತು ವಿಶ್ವಾಸಾರ್ಹವಾಗಿ ಮೌಲ್ಯಮಾಪನ ಮಾಡುವುದು ಅವುಗಳ ಸಾಮರ್ಥ್ಯಗಳು, ಮಿತಿಗಳು ಮತ್ತು ಕ್ಷೇತ್ರದ ಪ್ರಗತಿಯನ್ನು ಅರ್ಥಮಾಡಿಕೊಳ್ಳುವುದು ಅತ್ಯಗತ್ಯ. ಭಾಷಾ ತಂತ್ರಜ್ಞಾನಗಳಲ್ಲಿನ ಪ್ರಗತಿಗಳು ಡಿಜಿಟಲ್ ಅಂತರವನ್ನು ಉಲ್ಬಣಗೊಳಿಸದಂತೆ ಖಚಿತಪಡಿಸಿಕೊಳ್ಳಲು ಪ್ರಪಂಚದಾದ್ಯಂತ ವಿವಿಧ ಭಾಷೆಗಳು ಮತ್ತು ಸಂಸ್ಕೃತಿಗಳ ವ್ಯಾಪಕ ವ್ಯಾಪ್ತಿಯೊಂದಿಗೆ ಬಹುಭಾಷಾ ಮೌಲ್ಯಮಾಪನವು ಮುಖ್ಯವಾಗಿದೆ. ಈ ಭಾಷಣವು ಬಹುಭಾಷಾ ಮೌಲ್ಯಮಾಪನದ ಪ್ರಸ್ತುತ ಸ್ಥಿತಿಯನ್ನು ವಿವರಿಸುತ್ತದೆ, ಭಾಷೆ ಮತ್ತು ಕಾರ್ಯ ವ್ಯಾಪ್ತಿ, ಇಂಗ್ಲಿಷ್-ಕೇಂದ್ರಿತ ಡೇಟಾಸೆಟ್‌ಗಳು ಮತ್ತು ಮೌಲ್ಯಮಾಪನ ವಿಧಾನದಂತಹ ಗಮನಾರ್ಹ ಸವಾಲುಗಳನ್ನು ಎತ್ತಿ ತೋರಿಸುತ್ತದೆ. ಹೆಚ್ಚುವರಿಯಾಗಿ, ಸಮುದಾಯ-ಕೇಂದ್ರಿತ ಭಾಗವಹಿಸುವಿಕೆಯ ವಿನ್ಯಾಸ ಮತ್ತು ಬಹುಭಾಷಾ ಮೌಲ್ಯಮಾಪನ ಪ್ರಯತ್ನಗಳನ್ನು ಹೆಚ್ಚಿಸಲು AI ಬಳಕೆ ಸೇರಿದಂತೆ ಪರ್ಯಾಯ ಮೌಲ್ಯಮಾಪನ ವಿಧಾನಗಳಲ್ಲಿನ ಇತ್ತೀಚಿನ ಬೆಳವಣಿಗೆಗಳನ್ನು ಇದು ಚರ್ಚಿಸುತ್ತದೆ.

LT 4 whom?

Mandana Seyfeddinipur [1, 2, 3, *]

1 : Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)

2 : Endangered Languages Archive

3 : Endangered Languages Documentation Programme

* : Corresponding author

LT4 whom?
Developing language technology today requires a substantial corpus of written text and these just do not exist for more than probably 80% of the world's languages. Hence, Language Technology exists for around 2% of the languages of the world. At the same time, language activists are fighting all over the world for support to develop language technology protecting the languages of becoming endangered and providing services to these speech (and signing) communities. Connecting communities and industry for collaborative approaches to create technology; developing approaches working with other modalities than writing will allow to close this gap and also pushes creativity in tech.
LT4 für wen?
Die Entwicklung von Sprachtechnologie heutzutage erfordert einen umfangreichen Korpus geschriebener Texte, und die es für ueber 80 % der Sprachen der Welt nicht gibt Daher existiert Sprachtechnologie für gerade mal 2 % der Sprachen der Welt. Gleichzeitig kämpfen Sprachaktivisten auf der ganzen Welt um Unterstützung für die Entwicklung von Sprachtechnologie, um die vom Aussterben bedrohten Sprachen zu schützen und diesen Sprach- (und Gebärden-) gemeinschaften Dienste anzubieten. Die Verbindung von Gemeinschaften und Industrie für gemeinsame Ansätze bei der Entwicklung von Technologien und die Entwicklung von Ansätzen, die mit anderen Modalitäten als der Schrift arbeiten, wird es ermöglichen, diese Lücke zu schließen und auch die Kreativität in der Technik zu fördern.

---

sciencesconf.org:lt4all2025:617083

Language technologies across the world's languages: fairness, incentives, and policy

Damian Blasi [1]

1 : Catalan Institute for Advanced Study (ICREA)

In this presentation I will summarize a number of empirical observations on contemporary LT with the aim of answering the following questions: what percentage of all R&D in language technologies goes into English? Can macro-economic indicators of the populations of language users explain LT inequalities across languages? Does academia bolster the expansion of LT over under-resourced languages?

---

sciencesconf.org:lt4all2025:616751

Data Farming and the QRUE Frameworks: The NaijaVoices Experience

Gloria Monica Tobechukwu Emezue [1]

1 : Lanfrica / Alex Ekwueme Federal University Ndufu Alike, Ebonyi State, Nigeria

This presentation argues that the entire spectrum of frameworks for Quality, Reliability, Usability, and Ethics (hereafter referred to as QRUE Frameworks) for curating speech datasets in African languages would be more meaningful and relevant if the guiding principle and methodology for data curation in African languages shifted from the concept of data mining to data farming. The 1,800-hour NaijaVoices Speech Dataset, which involved the participation of no fewer than 5,000 people, working within four months, will be used to illustrate how Data Farming enabled the successful application of QRUE frameworks to speech data curation in African languages.

sciencesconf.org:lt4all2025:616960

Māori Algorithmic Sovereignty: Idea, Principles, and Use

Te Taka Keegan [1]

1 : University of Waikato [Hamilton]

Hillcrest, Hamilton 3216 - Nouvelle-Zélande

As data technologies and algorithms emerge in Aotearoa New Zealand for decision-making, frameworks are needed to maximize opportunities and minimize risks. For algorithms using Māori data, special consideration is required due to systemic biases affecting Māori in data and algorithm development. Algorithms represent a specific data use, allowing existing data frameworks to be extended. Māori data sovereignty principles, already used by researchers and government agencies, can be adapted for algorithms. This extension creates Māori algorithmic sovereignty principles, addressing responsible algorithm development from a Māori perspective.

I te aranga ake o ngā hangarau raraunga me ngā hātepe ki Aotearoa mō te whakatau whakaritenga, me whai anga hei whakanui i ngā āheinga me te whakaiti i ngā mōrea. Mō ngā hātepe e whakamahi ana i ngā raraunga Māori, me āta whakaaro ki ngā whakatau tōkeke e pā ana ki te iwi Māori i roto i te whanaketanga raraunga me te hātepe. Ko ngā hātepe he momo whakamahinga raraunga, ka taea te whakawhiti i ngā anga raraunga o nāianei. Ko ngā mātāpono mana raraunga Māori, e whakamahia kētia e ngā kairangahau me ngā tari kāwanatanga, ka taea te whakahāngai tonu ki ngā hātepe. Ko tēnei whakawhānuitanga ka hanga i ngā mātāpono mana hātepe Māori, e aro ana ki te whanaketanga hātepe tika mai i te tirohanga me te whakaro Māori.

Ai!: Building Inclusive Pathways to Indigenous Language Learning and Knowledge Access

Natasha Ita Macdonald [1, 2, 3] , Ali Mehdi [3, 4, *]

1 : Concordia University [Montreal]

1455 Boulevard de Maisonneuve O, Montréal, QC H3G 1M8 - Canada

2 : McGill's Faculty of Education

3 : Heritage Lab

4 : TÉLUQ University

* : Corresponding author

Heritage Lab's "Ai!" initiative demonstrates how Indigenous-led technology development can create inclusive pathways to language learning and cultural knowledge. By combining fine-tuned LLMs based on Inuit knowledge to deliver accessible learning solutions, the platform addresses digital barriers through dialect-aware translation, personalized learning paths, and searchable cultural archives. Through community-driven development engaging youth, Elders, and regional language committees, Ai! exemplifies how technology can enhance Indigenous language sovereignty while ensuring knowledge remains under community control. The platform also improves access to essential services by making critical information more accessible and understandable across sectors.

ᑖᒃᑯᐊ ᑌᔭᐅᔪᑦ Heritage Lab "ᖃᓇᑕᐅᔭᖅ ᖃᐅᔨᒪᔪᒻᒪᓇᒃ!" ᐱᕆᐊᖕᖓᑕᐅᒍᑎᖓ ᓴᕿᑎᑦᓯᕗᖅ ᖃᓄᖅ ᓄᓇᖃᖅᑲᓯᒪᔪᐃᑦ ᖃᐅᔨᓴᕐᕕᖓ ᖃᓇᑕᐅᔭᑎᒍᑦ ᐱᕚᓪᓕᐊᑎᑕᐅᓯᒪᔫᑦᓯᓂ ᐋᖅᑮᒍᓐᓇᑐᖅ ᐊᖅᑯᑎᑦᓴᐅᒍᓐᓇᑐᓂ ᐃᓕᓐᓂᐊᑕᐅᒍᓐᓇᑐᓂ ᐅᖃᐅᓯᓕᕆᓂᕐᒥᒃ ᐃᓗᕐᕈᓯᒥᓪᓗ ᖃᐅᔨᒻᒪᐸᓪᓕᐊᓂᕐᒥᒃ. ᐅᐊᕈᑦᔫᑐᕐᓯᕆᑦ ᖃᓇᑕᐅᓯᑦ ᖃᐅᔨᒪᓂᒻᒪᓇᖓᒍᑦ ᑐᖕᖓᕕᕆᑦᓯᕆᑦ ᐃᓄᐃᑦ ᖃᐅᔨᒪᔭᑐᖃᖏᑦ ᓄᐃᑦᓯᒍᓐᓇᑐᖅ ᐊᑐᕐᑕᐅᒍᓐᓇᕚᔭᕐᑐᓂᒃ ᐃᓕᓐᓂᐊᑎᑦᓯᒍᓯᕐᓂᒃ, ᑖᓐᓇ ᑲᑕᒃ ᓇᓗᓀᕐᓯᔪᖅ ᖃᓇᑕᐅᔭᒨᕐᑐᐃᑦ ᐊᐳᕐᓇᓲᖕᒍᕆᐊᖕᒋᑦ ᐅᖃᐅᓯᖃᖃᑎᖏᖕᖕᒋᑐᕐᑎᒍᑦ ᓇᓗᓀᕐᓯᒍᓐᓇᓯᓂ, ᓇᒻᒥᓂᕐᒎᓕᖕᖓᔪᒥ ᐃᓕᓐᓂᐊᑎᑦᓯᒍᓯᕐᒥᒃ ᐊᖅᑯᓯᐅᑎᒍᓐᓇᓂᖓ, ᐊᒻᒪᓗ ᕿᒥᕐᕈᕕᐅᒍᓐᓇᓯᓂ ᐃᓗᕐᕈᓯᓅᓕᖕᖓᔪᓂᒃ ᓴᓂᕐᕙᑕᐅᖃᑦᑕᓯᒪᔪᓂᒃ. ᓄᓇᓕᑦᑎᒍᑦ ᐊᔪᐃᓐᓈᓇᔭᐅᑦᓯᓂ ᐱᕚᓪᓕᐊᑎᑕᐅᓯᒪᔪᖅ ᐃᓚᒋᔭᐅᑦᓯᑎᒃ ᐅᕕᒃᑫᑦ, ᐃᓄᒻᒪᕇᑦ, ᐊᒻᒪᓗ ᓄᓇᓕᓖᒪᓂ ᐅᖃᐅᓯᓕᕆᓂᕐᒥᒃ ᑲᑎᒪᔩᑦ, "ᖃᓇᑕᐅᔭᖅ ᖃᐅᔨᒪᔪᒻᒪᓇᒃ!" ᓄᐃᑦᓯᓇᒃᑯ ᖃᓄᖅ ᖃᓇᑕᐅᔭᕈᑏᑦ ᓄᐃᑦᓯᕆᐊᓪᓚᕈᓐᓇᓂᖕᒋᓐᓄᑦ ᓄᓇᖃᖅᑲᓯᒪᔪᐃᑦ ᐅᖃᐅᓯᓪᓚᓇᖕᒋᓐᓂᒃ ᑖᒃᑯᐊᓗ ᐃᓗᓕᑦᑕᖕᒋᑦ ᑲᒥᔭᐅᑦᓯᐊᕈᓐᓇᓯᑎᒃ ᓄᓇᓕᑦᑎᒍᑦ. ᑖᓐᓇ ᑲᑕᒃ ᐃᓯᐊᕐᓂᓴᐅᑎᑦᓯᒥᔪᖅ ᐊᑐᐃᓐᓇᐅᑎᑦᓯᒍᓐᓇᓯᓂ ᐱᒍᑦᔫᑎᓂᒃ ᐊᑐᕐᑐᒪᓇᓐᓂᒃ ᑐᓴᐅᒪᔭᐅᓇᐊᓕᓐᓂᓗ ᐊᓐᓂᓇᕐᑐᓂᒃ ᐱᕕᑦᓴᖃᕐᑎᓯᑦᓯᐊᓯᓂ ᐊᒻᒪᓗ ᑐᑭᓯᓇᕐᑎᓯᑦᓯᐊᓯᓂ ᓇᐅᓕᒫᓂᒃ.

THEME: SOLUTIONS

Session Keynote 2

sciencesconf.org:lt4all2025:617452

A Shared Future in Language Technology

Ludmila Golovine [1, *]

1 : MasterWord (MasterWord)

303 Stafford Street Houston, Texas 77079 - États-Unis

* : Corresponding author

Most of the 7000 languages spoken today do not have a digital presence. Ludmila Golovine addresses this challenge, urging us to transform adversity into a movement of inclusive innovation and economic opportunity. By re-envisioning technology as a bridge—rather than a barrier—communities can shape their digital futures, from ethically governed language repositories to inclusive AI. This vision turns marginalized languages into engines of global creativity and economic growth, creating a roadmap for transformational change. Her keynote is a bold call to place the untapped potential of Indigenous and marginalized languages at the center of economic progress and innovation.

La plupart des 7 000 langues parlées à ce jour n'ont pas de présence numérique. Ludmila Golovine aborde ce défi et nous exhorte à transformer l'adversité en un mouvement d'innovation inclusive et d'opportunité économique. En repensant la technologie comme un pont – plutôt qu'une barrière – les communautés peuvent façonner leur avenir numérique, des référentiels linguistiques régis de manière éthique à l'IA inclusive. Cette vision transforme les langues marginalisées en moteurs de créativité mondiale et de croissance économique, créant ainsi une feuille de route pour un changement transformationnel. Son discours est un appel audacieux à placer le potentiel inexploité des langues autochtones et marginalisées au centre du progrès économique et de l'innovation.

Session O8: Co-Creating the Future of Language Technology: Government-Industry-Academia-Language Communities Collaboration Scheme

sciencesconf.org:lt4all2025:620217

AKILI UNDE - An English-Kiswahili AI Dictionary for a Multilingual Cyberspace

Misako Ito [1, 2, *]

1 : UNESCO Regional Office for Eastern Africa

2 : Bureau régional de l'UNESCO pour l'Afrique de l'Est

* : Corresponding author

While there are an estimated 7,000 languages and dialects in the world—of which African languages account for one-third—only 14 dominate the internet ecosystem. Less than 2% of Africa's estimated 2,000 languages have a real online presence, including in localized software and websites, translation and text-to-speech services, and content moderation algorithms. Indigenous, minority and low-resourced languages continue to be excluded from the benefits and opportunities of the digital world. Through finding equivalent of AI and digital terminologies in Kiswahili, this project demonstrates the urgent need to scale-up efforts to realize a truly multilingual and inclusive digital world through collaboration between industry, academia and language communities.

Alors que l'on estime à 7 000 le nombre de langues et de dialectes dans le monde - dont les langues africaines représentent un tiers -, seules 14 d'entre elles dominent l'écosystème de l'internet. Moins de 2 % des quelque 2 000 langues africaines sont réellement présentes sur Internet, notamment dans les logiciels et les sites web locaux, les services de traduction et de synthèse vocale et les algorithmes de modération de contenu. Les langues autochtones, minoritaires et à faibles ressources continuent d'être exclues des avantages et des opportunités du monde numérique. En trouvant un équivalent des terminologies de l'IA et du numérique en kiswahili, ce projet démontre qu'il est urgent d'intensifier les efforts pour créer un monde numérique véritablement multilingue et inclusif grâce à la collaboration entre l'industrie, le monde universitaire et les communautés linguistiques.

Mayan Language Preservation and Digitization: A Community Approach for Our Future. Ruyakik, rusamajixïk ri maya' ch'ab'äl chuqa rutz'ib'axik pa taq yakb'äl ch'ich': Jun samajixïk pa kamön richin ri qak'aslem chwa'q kab'ij apo..

Winston Scott [1]

1 : The Mayan Languages Preservation and Digitization Project (The Mayan Languages Preservation and Digitization Project)

No set address. All contributors are remote. - Guatemala

Equitable inclusion in digital spaces is among the most pressing concerns for marginalized languages today. Technological advances now offer unprecedented opportunities for linguistic preservation, revitalization, and the creation of digital resources that will ensure equity and full access in the digital world. By utilizing these tools we will open the door to economic, educational, and social benefits that are currently out of reach for millions of Indigenous language speakers. This presentation provides a practical roadmap for low-resource languages to start their own journeys from “zero to digital” and is a call to action for private and public industry to embrace these possibilities by working together with Indigenous communities to preserve and promote language access for all.

K'atzinel chupan taqramaj ch'ab'äl e jech'un, jajun ye'ok chupan taq k'ojlib'äl richin k'ak'a ch'ich'asamajib'äl. Runimilen na'ojinem nikiya' q'ij utzil majun tz'etajinäq richin ruyakik, ruchuq'a'il taqch'ab'äl, richin rub'anixik jujun taq k'ak'asamajib'äl chikan qitzij nuya' q'ij rujunamixik tz'aqät rukusaxik ronojel na'oj pa pitz'ib'äl. Rik'in rukusaxik samajib'äl re', niqajaq jun utzil, richin ruch'akik pwa'q, etamab'äl, winaqilal chi wakamïn, sib'ilaj achamaq'i' kich'ab'äl man k'o ta pa kiq'a'. Jun ruk'utik samaj re' nuya' jun rub'eyal k'atzinel, taq ch'ab'äl; rik'in jub'a' ok qax pe' yetikïr nikib'än pa kiyonil ri kib'ey “wa'ixk'a pa pitz'ib'äl” junpeyonïk samaj, chi mama' taq moloj richin champomal i k'o jujun kajaw, tikichapa' jun utzil, keto'on chike taq tinamït achamaq'i', richin nikiköl nikitaluj rukusaxik taqch'ab'äl kichin konojel.

Indigenous-led tech and self-determination

Eddie Avila [1]

1 : Global Voices (GV)

Kranenburgweg 135 A 2583 ER The Hague Netherlands - Pays-Bas

Digital technologies for language revitalization continue to evolve, which brings about both opportunities and increasing ethical concerns surrounding linguistic sovereignty. Indigenous communities must play meaningful roles in deciding whether and how these tools are applied to their own languages ensuring alignment with their values, needs, and aspirations. Building capacity among Indigenous communities and deepening their understanding of these tools and their implications are critical for advocating for their own interests. By becoming trusted intermediaries, community members can be in a better position to apply a critical lens to both the risks and benefits of these technologies, while also helping their own communities navigate this complex field and make informed choices. This presentation will explore efforts to strengthen community expertise and the potential of stakeholder partnerships to support inclusive engagement before critical decisions are made.

Las tecnologías digitales para la revitalización de lenguas siguen evolucionando, creando oportunidades y preocupaciones éticas sobre la soberanía lingüística. Las comunidades indígenas deben tener un papel clave en decidir si y cómo se aplican estas herramientas a sus lenguas, y en cómo se alinean con sus valores y necesidades. Fortalecer sus capacidades y comprensión de estas tecnologías es esencial para defender sus intereses. Como intermediarios de confianza, estas comunidades indígenas pueden evaluar los riesgos y beneficios de estas tecnologías, ayudando a sus comunidades a tomar decisiones informadas. Esta presentación explorará esfuerzos para fortalecer la experiencia comunitaria.

Open Language — how to make technology speak all languages

Sjur Moshagen [1]

1 : The Arctic University of Norway [Tromsø, Norway] (UiT)

Hansine Hansens veg 18, 9019 Tromsø, Norvège - Norvège

The Divvun group at *UiT the Arctic University of Norway*, has developed language technology for the Sámi languages for more than 20 years. From the beginning, integrating tools with the software and systems that the speakers use has been challenging. The presentation assesses the impact these barriers have on language communities, and presents a model for how the technology industry can make the technology field open to all languages: open all apps to third party localisations, package them for easy installation, and open all natural language APIs to registered developers, unconditionally.

Norwegian nynorsk (ISO 639: nn) translation:
Open Language — slik kan vi få digital teknologi til å prata alle språk

Divvun-gruppa ved *UiT Noregs arktiske universitet*, i Tromsø, har utvikla språkteknologi for dei samiske språka i meir enn 20 år. Heilt frå starten var det ei utfordring å integrera verktøya med dataprogram og system som språksamfunna bruker. Presentasjonen går stutt gjennom kva for fylgjer desse barrierane har for språksamfunna, og legg deretter fram ein modell for korleis teknologiindustrien kan gjera heile teknologisektoren open for alle språk: opne alle appar for tredjepartsomsetjing, pakk dei inn slik at dei er lette å installera, og opne alle API-ar for naturlege språk for registrerte utviklarar, utan vilkår.

---

sciencesconf.org:lt4all2025:617779

Digital Future of Indigenous Languages – a ground to prosper or new battlefield?

Valts Ernštreits [1,*]

1 : University of Latvia Livonian Institute

Kronvalda bulv. 4 - Lettonie

* : Corresponding author

In March 2025, the First Global Survey of Indigenous Languages will be launched by the Global Task Force for making the International Decade of Indigenous Languages (2022–2032). This survey marks the first global effort to assess the state and needs of indigenous languages in digital domains.
It is already clear that the key challenge for the digital presence of indigenous languages lies in speakers' ability to generate sufficient digital language data, essential for technology development, and their freedom to use their language across all digital domains. Academia, technology developers, and governments must take responsible action to eliminate all barriers and support the inclusion and use of indigenous languages in all digital spaces.
Alliztrov kīeld digitāli tulbizāiga – kūož ēdrikšimiz pierāst agā ūž suodānuŗm?
2025. āigast märtsõs Alliztrovd kīeld āigastkim (2022–2032) Globāli jūodimiz kub tieutõb Ežmiz alliztrovd kīeld tieut kuoŗŗimiz. Se tieut kuoŗŗimi um ežmi globāli kōļimi arū sõdõ, kus digitālizõs mõīlmas alliztrovd kīeld ātõ paldīņ, ja mis näntõn um vajāg.
Jõbā paldīņ um sieldõ, ku amād alliztrovd kīeld kõlbatimiz pierāst digitālizõs mõīlmas amā tādzi ažā um nänt rõkāndijizt võimi lūodõ dattidi, mis ātõ tārpalizt tehnologijd kazāntimiz pierāst, ja nänt võimi kõlbatõ eņtš kīeldõ ämši digitāližis aŗši. Tuņšlijiztõn, tehnologijd kazāntijiztõn ja võļikštõkstõn um võtāmõst vastātõks iļ sīe, laz amād tõgõd kīeld kõlbatimiz pierāst sõgõd kõtõd, ja näntõn um tigtõmõst alliztrovd kīeld kõlbatimiz ämši digitāližis aŗši.

Language Resources Protection in the Digital Intelligence Era

Lining Wang [1]

1 : Center for the Protection and Research of Language Resources of China, Beijing Language and Culture University, China

There are 56 ethnic groups in China, a multi-ethnic, multi-lingual, and multi-dialect country with multiple-written characters. China has undertaken large-scale language resource surveys to protect language resources. Utilizing modern technological means, it collects and records actual language resource of Chinese dialects, minority languages, and oral language cultures. A large-scale, sustainable multimedia language resource repository has been established. Via the internet and digital technology, online query, learning, and research services are provided, allowing the public to easily access and utilize these resources. Currently, we are dedicated to building a national language resource museum through digital technology, ensuring that the language culture spanning thousands of years will be passed down from one generation to the next.
中国是一个多民族、多语言、多方言的国家。为了科学保护语言资源，中国开展了大规模语言资源调查，利用现代化技术手段，收集记录汉语方言、少数民族语言和口头语言文化的实态语料，通过科学整理和加工，建成大规模、可持续增长的多媒体语言资源库，通过互联网和数字技术，提供在线查询、学习和研究服务，让公众能方便地访问和使用语言资源。目前正致力于通过数字技术打造国家级语言资源博物馆，以实体馆为基础，合理运用数智化时代先进的科学技术手段，提供沉浸式、交互式的观展体验，让跨越千年的语言文化永存后世，代代相传。

Session O9: “Shaping the Future: Policies for Human-AI Coexistence”

De l'extinction à l'inclusion : redonner leur langue à des millions de personnes

Heiura Itae-Tetaa [1,*]

1 : E-Reo (E-Reo)

E-Reo

5 avenue du Régent Paraita, Tahiti, Polynésie française - France

* : Corresponding author

Dans un monde où plus de la moitié des langues sont absentes du numérique, il est crucial de doter les langues autochtones d'outils adaptés pour les digitaliser et bâtir une mémoire numérique vivante. E-Reo est une plateforme SaaS qui permet aux communautés de créer leurs propres applications linguistiques, intégrant les langues dans les usages du quotidien. Cet outil ne se limite pas à archiver les langues : il les rend interactives et accessibles. Avec une architecture scalable, E-Reo vise à intégrer 400 langues d'ici 2027, en collaboration avec des partenaires pour structurer et valoriser les données linguistiques existantes.

I roto i te hō'ē ao, e mea tītau-roahia ‘ia fāna'o te mau reo ihotupu i te rāve'a rorouira tano ‘ia nehenehe e fa'aineine i te tahi putura'a parau ora. E tahua “SaaS” ‘o E-Reo e ha'afāna'o nei i te ta'ata ‘ia hāmani i tā rātou iho rāve'a e te reo nō te fa'a'ohipa i te mau mahana ato'a. E'ita teie rāve'a e putu noa i te parau, e fa'ariro atu ‘oia i te reo ‘ei moiha'a ‘ana'anatae ‘e te ‘ōhie ‘ia fa'a'ohipa. Te fā i teie matahiti 2027, e fa'aō hau atu i te 400 reo e te mau ‘āmui nō te fa'anaho ‘e nō te ha'afaufa'a i te mau parau e vai ra i ni'a i te mau reo.

---

sciencesconf.org:lt4all2025:616887

Leveraging Language Technologies to Promote Resilience: A Community-Based Participatory Approach to Reduce School Dropout Among Namibian San Youth

Naftali Indongo [1,*], Heike Winschiers-Theophilus [2,*], Rosetha Kays [3,*], Shorty Kandjengo [4,*]

1 : Namibia University of Science and Technology (NUST)

13 Jackson Kaujeua Street Windhoek - Namibie

2 : Namibia University of Science & Technology (NUST)

13 Jackson Kaujeua Street Windhoek - Namibie

3 : Namibia University of Science & Technology (NUST)

13 Jackson Kaujeua Street Windhoek - Namibie

4 : Inclusive and Collaboratve Innovaton Tech Hub

Donkerbos Community, Omaheke Region - Namibie

* : Corresponding author

The project aims to address educational challenges and promote resilience and perseverance among Namibian San learners by integrating participatory design with advanced language technologies. To address such challenges, we developed an AI-powered Smart Mirror prototype that uses facial emotion detection to deliver personalised motivational stories using an LLM and a Bilingual Language Translation model to facilitate seamless conversion between Ju/'hoansi and English, enabling learners to engage with the system in their native language. The project aims to promote linguistic diversity, cultural preservation, and educational equity in line with current developments of the Namibian Action Plan for the International Decade of Indigenous Languages (IDIL).

Omapekapeko onkambadhala yokukandulapo omashongo osho wo okugandja omukumo kaanasikola yomuhoko gwaaSan moNamibia pamukalo gwokutula mumwe uungomba welaka osho wo elongo moka yena okudhana onkandangala. Okukandulupa omashongo goludhi nduka otwanduluka po esiipili lyopautekinika wopashinanena tau ithanwa oArtificial Intelligence ta lilongo pamukalo gokutaalela omaiyuvo gokoshipala ta li longinga oLLM osho wo omodela yokutooloka omalaka gayooloka ndjoka tayi toolokele okuza moshiingilisa okuukitha melaka lyoshiJu/'hoansi shoka tashikwathele aalongwa yiilonge uungomba melaka lyawo.

Linguistic Diversity through the Lenses of the UNESCO Recommendation on the Ethics of AI

Doaa Abu Elyounes [1]

1 : The United Nations Educational, Scientific and Cultural Organization (UNESCO)

UNESCO

The UNESCO Recommendation on the Ethics of AI, adopted by acclamation by all Member States in November 2021 highlights the need to ensure diversity and inclusiveness as a key value that should be respected throughout the AI life cycle. The Recommendation calls on all AI actors to ensure that the benefits of AI technologies are available and accessible to all, taking into consideration, among other thing the specific needs of different language groups.
This talk will unpack the different provisions of the Recommendation related to linguistic diversity, and explore UNESCO's role in enhancing linguistic diversity in AI.

اعتمدتها التي ،الاصطناعي الذكاء أخلاقيات بشأن اليونسكو توصية تسلط الاصطناعي الذكاء أخلاقيات بشأن اليونسكو توصية عدسات خلال من اللغوي التنوع الذكاء حياة دورة طوال احترامها يجب أساسية كقيمة والشمولية التنوع ضمان إلى الحاجة على الضوء ،2021 نوفمبر في بالإجماع الأعضاء الدول جميع مراعاة مع ،للجميع وإتاحتها الاصطناعي الذكاء تقنيات فوائد توفر ضمان إلى الاصطناعي الذكاء مجال في الفاعلة الجهات جميع التوصية وتدعو .الاصطناعي دور وتستكشف ،اللغوي بالتنوع المتعلقة للتوصية المختلفة الأحكام المحاضرة هذه ستتناول .أخرى أمور بين من مختلفة لغوية لمجموعات المحددة الاحتياجات .الاصطناعي الذكاء في اللغوي التنوع تعزيز في اليونسكو

Indigenous AI: Abundant Intelligences and IndigiGenius

Michelle Brown [1, 2]

1 : IndigiGenius

2 : Abundant Intelligences

This talk highlights two Indigenous-led technology initiatives: Abundant Intelligences and IndigiGenius. Both apply artificial intelligence (AI) and computer science (CS) to support and strengthen Indigenous languages and Indigenous Data Sovereignty. Abundant Intelligences is an Indigenous-led research program that conceptualizes, designs, develops, and deploys based on Indigenous Knowledge systems.
IndigiGenius is an Indigenous-led tech nonprofit brining AI and CS education to Indigenous communities through programs like the Lakota AI Code Camp, Lakota-designed AI and CS education for Lakota high school students, and First Languages AI Reality (FLAIR), automated speech recognition (ASR) models supporting endangered Indigenous languages.

AI ʻŌiwi: No ka Abundant Intelligences a me IndigiGenius

Pili nō kēia haʻi ʻōlelo i ʻelua mau papahana ʻenehana i alakaʻi ʻia e ka poʻe ʻōiwi: ʻo ka Abundant Intelligences lāua ʻo ka IndigiGenius. Hoʻohana nā mea ʻelua i ka AI a me ka ʻepekema lolo uila i mea e kākoʻo a hoʻoikaika ai i nā ʻōlelo ʻōiwi a me nā ea ʻikepili. No ka Abundant Intelligences, he papahana noiʻi e wānana, hakulau, hoʻomōhala, a hoʻōlapa ana hoʻi ma ke kahua o ia mea ka ʻōnaehana ʻike kupuna. No Indigigenius, he hui ʻenehana kaiaulu kumuloaʻa ʻole e aʻo ana i ka AI a me ka ʻepekema lolo uila i nā kaiaulu ʻōiwi ma o nā polokolamu e laʻa me ka Lakota AI Code Camp, ka hoʻonaʻauao ma ia mau kumumanaʻo i haku ʻia e ka poʻe Lakota no nā haumāna kula kiʻekiʻe Lakota, ka First Languages AI Reality (FLAIR), a me nā kumu hoʻohālike no ka automated speech recognition (ASR) i mea e kākoʻo ai i nā ʻōlelo ʻōiwi ʻane halapohe.

Building Language Revitalization Robotics Rooted in Ethical AI

Danielle Boyer [1]

1 : The STEAM Connection

330 E Maple Road Troy, MI USA 48083 - États-Unis

The Anishinaabeg, an Indigenous people of the Great Lakes region of North America, face critical language loss, endangering the future of Anishinaabemowin. This presentation explores the development of the SkoBot by Danielle Boyer, an Anishinaabe youth robotics inventor. SkoBot, an AI-powered personal language revitalization robot, supports the resurgence of Indigenous languages by bridging generational gaps in language learning. Through case studies in interactive and community-based learning, as well as digital archiving, we demonstrate how ethical AI applied through robotics—rooted in data sovereignty and tribal governance—can empower youth to lead language preservation efforts while ensuring long-term cultural continuity.

Les Anishinaabeg, un peuple autochtone de la région des Grands Lacs en Amérique du Nord, font face à une perte critique de leur langue, mettant en danger l'avenir de l'anishinaabemowin. Cette présentation explore le développement du SkoBot par Danielle Boyer, une jeune inventrice Anishinaabe en robotique. SkoBot, un robot de revitalisation linguistique alimenté par l'IA, soutient la renaissance des langues autochtones en comblant les écarts générationnels dans l'apprentissage linguistique. À travers des études de cas sur l'apprentissage interactif et communautaire ainsi que l'archivage numérique, nous démontrons comment une IA éthique appliquée à la robotique—ancrée dans la souveraineté des données et la gouvernance tribale—peut permettre aux jeunes de diriger les efforts de préservation linguistique tout en assurant la continuité culturelle à long terme.

The Icelandic Approach to Language Technology

Sofiya Zahova [1,*], Óttar Kolbeinsson Proppé [2]

1 : International Centre for Multilingualism and Intercultural Understanding

2 : Icelandic Center for Language Technology

* : Corresponding author

This talk elaborates on the paths to multicultural and multilingual digital environments on the example of Iceland's experiences in LT. It elaborates on the lessons learned from the Icelandic work with OpenAI and the keys to success that apply to languages with fewer speakers. We call for global action with a proposal for an open, international project to establish best practices and develop benchmarks in multicultural and multilingual AI. This future initiative will strive to become a trusted forum and a resource for long-tail language and cultural communities seeking to contribute local data and knowledge for use openly and fairly.

Í erindinu verður fjallað um leiðir að fjölmenningarlegu og fjöltyngdu stafrænu umhverfi með hliðsjón af reynslu Íslands á sviði máltækni. Megináhersla erindisins verður á þann lærdóm sem draga má af samstarfi Íslands við OpenAI um stuðning við íslensku í GPT-mállíkönunum og hvernig styðja megi við tungumál fárra málhafa í nýrri tækni. Við leggjum til að helstu hagsmunaaðilar og þátttakendur á sviði gervigreindar taki höndum saman í opnu alþjóðlegu átaki sem miðar að því að koma á viðmiðunarreglum og gagnreyndum aðferðum, þróa safn mæliprófa, safna fyrirmæla- og úrlausnargögnum og styðja við rannsóknir á sviði fjölbreyttrar tungumálakunnáttu og menningarþekkingar í gervigreind.