



DISRUPTING DIGITAL MONOLINGUALISM

**A report on multilingualism
in digital theory and practice**

DISRUPTING DIGITAL MONOLINGUALISM

REPORT 2021

Written by Paul Spence with input from workshop collaborators and participants

Layout and design by Renata Brandão. Photo by [Ryutaro Uozumi](#) on [Unsplash](#)

Event website <https://languageacts.org/digital-mediations/event/disrupting-digital-monolingualism/>

This report forms part of a series of reports produced by the Digital Mediations strand of the Language Acts & Worldmaking project, in this case in collaboration with the translingual strand of the Cross-Language Dynamics project (based at the Institute of Modern Languages Research), both funded by the UK Arts and Humanities Research Council's Open World Research Initiative. Digital Mediations explores interactions and tensions between digital culture, multilingualism and language fields including the Modern Languages.

[DOI: 10.5281/zenodo.5743283](https://doi.org/10.5281/zenodo.5743283)

License used: This work is licensed under a Creative Commons Attribution 4.0 International license (CC BY 4.0): <https://creativecommons.org/licenses/by/4.0/>. The license applies to all content unless noted otherwise in the report. This license lets others copy and redistribute the material in any medium or format, and to adapt (remix, transform, and build upon the material) the content for any purpose, even commercially, as long as it is attributed correctly.



Please contact us with any questions about using the report: paul.spence@kcl.ac.uk

Table of Contents

EXECUTIVE SUMMARY	1
INTRODUCTION	2
ABOUT THE WORKSHOP	4
Part One (State of the Art)	5
Part Two (Future Directions)	5
General information	6
PART ONE – STATE OF THE ART	8
Contexts	8
Panel/invited talk	8
Lightning Talks, demos, posters and mini-workshop: key themes	12
PART TWO – FUTURE DIRECTIONS	20
Theme group 1: Requirement profile for multilingually enabled digital knowledge infrastructures with a special focus on non-Latin scripts	21
Theme group 2: Multilingual challenges in the use of multimodal corpora	23
Theme group 3: Transcultural and translingual approaches to digital study	24
Theme group 4 - Artificial intelligence, machine learning and NLP in language worlds	25
Concluding the theme groups	26
CONCLUSIONS	29
Since the workshop	29
Final thoughts	32
REFERENCES	34
APPENDICES	37
Appendix 1: About the organisation of the event	37
Appendix 2: Aims and themes	37
Appendix 3: List of presentations/contributions	40
Appendix 4: Speakers and contributors	42
Appendix 5: Key links	45

Executive Summary

The Disrupting Digital Monolingualism workshop aimed to map the current state of multilingualism in digital theory and practice through, and across, languages and cultures.

This virtual workshop brought together leading researchers, educators, digital practitioners, language-focused professionals, policy-makers and other interested parties to address the challenges of multilingualism in digital spaces and to collectively present new models and solutions.

In the first part of the workshop, a series of contributions in different formats aimed to capture the breadth of responses to digital monolingualism on topics such as:

- Linguistic diversity in digital knowledge production
- Scripts, and in particular non-Latin scripts
- Multilingual language resources
- Multilingualism in media, platform, and software studies
- Language technologies, including Natural Language Processing (NLP)
- Low-resourced languages
- Transcultural and translingual dynamics
- Language pedagogies
- Multilingualism in practice

In the second part of the workshop, four ‘Theme groups’ were formed, which examined the following themes:

1. Linguistic and geocultural diversity in digital knowledge infrastructures
2. Working with multilingual methods and data
3. Transcultural and translingual approaches to digital study
4. Artificial intelligence, machine learning and NLP in language worlds

Key conclusions from the workshop:

- The workshop demonstrated both the benefits and challenges in bringing together different academic and professional fields to attend to digital multilingual issues, while also proving the value of providing holistic responses which draw on the perspectives of researchers, industry and language communities alike.
- Discussion around digital multilingualism is still fragmented (and sometimes marginalised), but momentum is growing across different fields to disrupt monolingual/Anglo-centric or linguistically/geoculturally exclusionary approaches to language technology.
- Access to open and multidisciplinary data, services and tools needs to be made easier and expanded to cover a wider range of languages. Training is also needed to lower the barrier to entry for those wishing to work with new digital tools in less technically-oriented fields of study.
- Digital multilingualism is not just about linguistics, or language. It has a cultural and socio-political dimension which is crucial in studying increasingly transcultural and translingual dynamics and in helping us to understand what are ultimately human-designed and complex (digital) cultural artefacts.
- Languages are not clearly bounded objects, and tools and technologies risk being based on an overly narrow conceptualisation of languages if they are not connected to people’s actual multilingual practices.
- People interested in digital multilingualism need more opportunities to engage with each other in exploring these questions, and to promote alliances between academic, commercial, third sector and language community respondents.

Introduction

There are currently something like 7,000 human languages in the world, but many experts fear that a very large proportion may face extinction due to the consolidation of what Abram de Swaan calls ‘supercentral languages’, with English as the ‘hypercentral’ language (de Swaan, 2002). Far from disrupting this dynamic, digital culture and infrastructure often impede language diversity, reinforcing the mistaken assumption that monolingualism is the global norm, when, in fact, most of the world negotiates communication in multilingual spaces. A large proportion of digital tools are constructed with a monolingual mindset, and most often the language is English. The text-centric nature of many digital approaches is an additional barrier for the many languages with a stronger oral culture/tradition.

There has been increasing attention in the past few years to the challenges of multilingualism in digital practice and it has been widely accepted that digital ecosystems have a ‘language and geocultural diversity’ problem. At present they have a strong bias towards firstly English, and then a small group of (largely European) languages. Over the years, numerous initiatives have attempted to address this imbalance in a variety of ways, whether driven by *practice* or *theory*. There are vibrant and well-established research communities exploring various aspects of digital multilingualism from multiple disciplinary perspectives and, in preparing for the workshop reported on here, we were inspired in particular by:

- **Empirical and theoretical work on multilingualism in digital spaces** (Danet and Herring, 2007; Lee, 2016; Maffi, 2005; Maaya Network, 2012).
- **Projects to measure and foster digital language diversity** (The Digital Language Diversity Project, 2019; Vrana et al., 2020; Soria et al., 2016; Ceberio Berger et al., 2018).
- Initiatives such as [CLARIN](#), [Language Grid](#) and [Nexus Linguarum](#) supporting **new language technology landscapes for research and commerce** at European level.
- **Digital research communities which include an explicit linguistic focus** (Humanistica, 2014; Network for Digital Humanities in Africa, 2020).
- **Minority, endangered or heritage languages archives/platforms** ([Endangered Languages Archive](#); [Endangered Languages Documentation programme](#); [PARADISEC](#); [First Peoples’ Map](#)).
- **Modern Language and Area Studies** where researchers increasingly explore digital transformations within language learning, cultural studies or transnational social sciences. (Taylor et al., 2017; Pitman and Taylor, 2017; Patti, 2018; Schneider, 2015; Vierthaler, 2020).
- **Multilingual initiatives in digital research fields, including the digital humanities** (workshops on [non-Latin scripts](#) and [African Languages and digital humanities](#) at the DH2019 conference in Utrecht; the [Multilingual DH](#) initiative; the [MARKUS](#) tool; and the Right to Left/RTL workshops at [DHSI summer schools](#)).
- Work being done to explore the **implications for linguistic diversity in increasingly machine and data-driven processes** (*META-NET White Paper Series*, 2012; Nekoto et al., 2020).

Digital culture has transformed many of the ways in which we engage with languages – whether through language learning apps like Duolingo or machine translation services like Google Translate – and yet, broadly speaking, the study of languages and associated cultures continues to suffer from a precarious existence both institutionally and in the popular imagination.

If machines can instantly translate and interpret for us, the popular logic increasingly goes, why bother understanding the role of (or learning) languages, or why do we need to trouble ourselves with intercultural competence? This same logic pervades much of the current debate on language technologies, which frequently lacks a 'critical-cultural' perspective.

Various commentators have challenged the dynamics of ‘language indifference’ or ‘language insensitivity’ present in “the unmarked monolingual assumptions of numerous academic disciplines and non-academic sectors” (Forsdick, 2017). A report on *Transnationalizing Modern Languages: Reframing language education for a global future* in 2018 argued that “Language learning constitutes both a social and an economic resource, while the ability to move between linguistic and cultural systems has become a feature of everyday life” and so we should “make the work of language and of translation visible” (Burdett et al., 2018). At present, ‘language work’ is frequently invisible, or fragmented across disciplines, and the task of understanding the huge implications of digital transformations within modern languages and cultures is split between research areas with little connection or mutual dialogue. The [Digital Modern Languages seminar series](#) was launched in 2019 to address some of these discontinuities and to bring together and raise the visibility of Modern Languages research which engages with digital culture, media and technologies. In general terms, languages (in the plural) and multilingualism lack agency in debates, initiatives and research focusing on digital transformation.

Finally, when exploring interactions between languages and digital culture, the overriding trend is to explore how the 'digital' disrupts languages. It is much less common to examine the relationship in the other direction, to ask how 'languages' disrupt digital studies and practice. Based on these circumstances, in 2019 a group of researchers set out to organise an event which would examine how languages, translation, and translingual/transcultural dynamics disrupt the predominantly monolingual, and particularly anglophone, model on which digital culture and technology are based.

This report describes the [Disrupting Digital Monolingualism](#) workshop, which aimed to map the current state of multilingualism in digital theory and practice through, and across, languages and cultures. It explored multilingual and geocultural challenges facing digital studies, addressing criticisms that digital research and practice are often ‘language indifferent’ and generally have weak engagement with the many ways in which languages play an important part in how we ‘make worlds’, in digitally mediated research and teaching practices.

The workshop had the following objectives:

- To identify areas of linguistic (especially anglophone) bias and 'language indifference' in digital methodologies and infrastructure.
- To discuss the value and role of languages in digital theory and practice and their implications for language study and professions.
- To bring together experts in languages-driven digital study and practice to discuss priorities for future action and potential collaboration.
- To explore emerging models for linguistic diversity and languages-aware digital practice in academia, education and private/third sectors, and to document best practice.

The workshop was held online on June 16th and 17th 2020 and hosted by the [Language Acts & Worldmaking](#) project with the support of the [Cross-Language Dynamics: Reshaping Community project](#), both projects funded by the UK Arts and Humanities Research Council as part of its [Open World Research Initiative](#).

About the workshop

The event was conceived as a two-part workshop:

- The first part aimed to showcase **existing research and practice**, including some of the most exciting and innovative responses to digital monolingualism from across digital practice and theory.
- The second part aimed to bring together leading **researchers, educators, digital practitioners, language-focused professionals and policy makers** to address the challenges of **multilingualism in digital spaces** and to propose **key areas of strategic development** in multilingual digital theory and practice.

To some extent, the event involved traditional conference-like formats (like thematic panels or lightning talks), but the fundamental aim was to set up practical and agile engagements between a wide variety of stakeholders (across different fields in and beyond academia, including language-based professionals) in order to evaluate existing multilingual and languages-focused tools and methods, and seek new ways for people to examine existing collections through different linguistic and geocultural frames.

Originally intended to take place as a primarily face-to-face event, due to the COVID-19 pandemic, it was converted into a virtual workshop, with a series of synchronous/live and asynchronous/pre-recorded interventions, using a combination of audio-visual and text-based tools. Many of these interventions were recorded and published, and we will summarise the debates and findings in them later, but in this section, the report will first describe the event structure in more detail, and then outline some of the practical challenges involved in moving an event like this online at short notice, before summarising the scale and nature of interventions at the workshop.

Figure 1: DDM workshop registration



Source: Image by Alexandra Koch from Unsplashd

Part One (State of the Art)

Panel

Our first invited session at the workshop brought together four panelists: Anasuya Sengupta (Whose Knowledge? campaign); Cosima Wagner (Freie Universität Berlin); Kalika Bali (Microsoft Research India); and Eduard Arriaga (University of Indianapolis). The panel was asked to present what they saw as the main challenges in relation to digital multilingualism right now, to propose some potential models for addressing these challenges and to outline the roles, communities and organisations that needed to be involved in this discussion going forward.

Keynote

Our other invited speaker was Mandana Seyfeddinipur, Director of the Endangered Languages Documentation Programme and Head of the Endangered Languages Archive at SOAS University of London, whose main research focus is on documentation of endangered languages, language use and video recording. Her presentation, titled *If it's not written it did not happen: The written bias in the digital world*, examined how we can improve access and content production in languages other than the small number of majority languages which currently dominate digital ecosystems.

Other contributions

In early 2020 we launched a Call for Proposals to present at the workshop in a number of formats including lightning talks, posters, demos, roundtables and mini-workshops, and we also welcomed proposals for experimental formats. We received a wide range of submissions on topics from 'digital responses to language endangerment and script displacement' to 'digital homelands for diasporic languages' and from 'open online platforms for multimedia language learning' to 'Multilingualism as an instrument for fostering diversity in South Africa'.

Accepted contributions consisted of:

- Fifteen **lightning talks**, which were pre-recorded and then played live at the event, with discussion at the end of each session.
- Five **demos and posters**, which were presented live in virtual breakout rooms.
- A half-day **workshop** on *Multilingual data in ELTeC: enacting European literary traditions*, which introduced participants to the Distant Reading for European Literary History project and allowed them to explore linguistical and computational challenges arising from the creation of the ELTeC (European Literary Text Collection) multilingual corpus.

The [Full Programme](#) is listed on the workshop website.

Part Two (Future Directions)

The second part of the *Disrupting Digital Monolingualism* workshop consisted of one week of focused/mixed asynchronous-synchronous discussion around pre-defined topics which aimed to build on debates from the 'live' first part of the workshop and to propose **key areas of strategic development** in multilingual digital theory and practice. Originally, the event was planned to happen predominantly in person at King's College London, but as a result of the pandemic we resolved to establish four virtual 'theme groups' instead, which were formed before the workshop and started their activities directly after the synchronous event on June 16th & 17th 2020. For each group, we invited two facilitators per theme to elicit responses to a series of starting questions for each theme, over a one-week period.

Within this framework, each group decided on its own makeup and mode of operation. Some groups arranged a one-off meeting, one sent daily prompts which members responded to in a writing sprint, and another organised a public survey. We simply asked that each group report back on what they had achieved (which they did a week later) and publish any outcomes on the workshop website.

Each group loosely covered one of four areas:

1. Linguistic and geocultural diversity in digital knowledge infrastructures.
2. Working with multilingual methods and data.
3. Transcultural and translanguaging approaches to digital study.
4. Artificial intelligence, machine learning and Natural Language Processing (NLP) in language worlds.

General Information

Technical plan/Switch to virtual

From a technical point of view, the rapid shift from a face-to-face workshop to an online event presented a number of challenges, particularly since the transition happened mid-way through the planning process.

In practical terms, the event was delivered using the following platforms and tools:

- Zoom for livestreamed presentations and interactions.
- YouTube for live broadcast (and later as a partial archive).
- Twitter (using the hashtags #DDM20 and #DigitalMultilingualism) and Slack for back-channel community discussion.

The technical specification for the event was produced with guidance from the organisers of the [Global Digital Humanities symposium](#), to whom we offer thanks.

We aimed to facilitate dynamic and informal discussion around the presentations, although this presented challenges due to time considerations and the relatively early stage in the process of wider acclimatisation to new pedagogies of interaction provoked by the pandemic. Additionally, presenters and audience members were working across numerous different time zones. We encouraged presenters to translate their presentations and to consult resources such as the [GO::DH Translation toolkit](#) (<https://go-dh.github.io/translation-toolkit/conferences/>) for ideas on how to make their presentations more multilingual.

Speakers and Contributions

There were 25 speakers with institutional affiliation in 12 different countries over two days for the synchronous event (Part One), and the theme groups included 41 participants based in 17 countries for Part Two. Registrants came from 49 countries spanning six continents. The workshop brought together people from academia, language professions, secondary education, local and global digital media companies, the cultural heritage sector, the Galleries Libraries Archives Museums (GLAM) sector, funding agencies, international policy organisations and the creative arts sector.

Registrants were asked why they joined the workshop, and a wide variety of responses was provided, which, broadly summarising, included the following topics:

- General understanding of linguistic landscape online
- Greater multilingual perspectives on digital study and practice

- Multilingual approaches to Artificial Intelligence (AI), Natural Language Processing (NLP) and text mining
- Language technologies
- Low-resourced languages
- Translation and intercultural exchange
- Intersections and borders
- Language education, Modern Languages and Area Studies
- Language activism and justice

We provide more detailed information about participants in Appendix 4.

Our positioning

Geography and language are never neutral. We would like to acknowledge the largely anglophone/European positioning of the organising team. While we have aimed to be as globally inclusive as possible, we recognise that our perspective inevitably excludes other linguistic and cultural contexts which we may not be aware of.

Outcomes

The Covid-19 pandemic not only changed our original plans for the workshop design; it also significantly delayed both this final report and other reports by some of the theme groups. Nevertheless, an initial report about the event was posted shortly afterwards, along with YouTube videos and slides (for those presentations for which we had permission) and [summaries of the makeup and activity of each 'Theme Group'](#). You can see videos of most of the interventions from the public/first part of the event on our [YouTube channel](#), and they are listed on the workshop website for convenience:

<https://languageacts.org/digital-mediations/event/disrupting-digital-monolingualism/ddm-workshop-videos/>

The workshop confirmed considerable appetite for continuing to disrupt digital monolingualism in its various forms, and the rest of this report describes the main themes from the workshop and suggests some possible future lines of enquiry and areas for potential collaboration.

Part one - state of the art

The first part of the workshop aimed to capture the breadth in digitally-mediated multilingual theory and practice through, and across, languages and cultures, and to draw connections between numerous overlapping digital and languages-driven conversations and initiatives. In this section we summarise the key topics covered in the lightning talks, posters, demos, mini-workshops, panel and keynote presentation.

Contexts

As noted earlier, numerous academic and professional fields of expertise were represented at the workshop covering both ‘linguistic’ and ‘cultural’ perspectives embedded within the concept of ‘multilingualism’. The workshop was inaugurated by Catherine Boyle, the Principal Investigator on the [Language Acts and Worldmaking](#) project, who welcomed participants as part of an international community of people “convinced that language research is central to our understanding of the world”.

“Our words make worlds”

Language Acts and Worldmaking

Panel/invited talk

A panel of experts in digital multilingualism explored digitally mediated ‘worldmaking’ from four different perspectives. Firstly, Anasuya Sengupta from the [Whose Knowledge? Campaign](#) presented the [Decolonizing the Internet’s Languages](#) initiative, which aims to raise the profile of

Figure 2: “The internet does not look or sound like most of us in the world”



Source: Decolonizing the Internet's Languages initiative (CC BY-SA 4.0)

marginalised language communities and examine the relationship between language diversity, digital access and epistemic injustice (Vrana et al., 2020).

Sengupta calculated that only 7% of the world’s languages are represented in published material, and that the proportion of those which have been digitised is even smaller. She estimated that users from only 10 languages represent over 75% of those on the Internet, with English and Mandarin Chinese dominating (based on figures from <https://www.internetworldstats.com/stats7.htm>). This in spite of 59% of the world being online, 75% of whom are from the Global South. Drawing on collaborative work between the [Center for Internet & Society, Oxford Internet Institute](#) and Whose Knowledge?, Sengupta demonstrated the severe lack of support for languages in interfaces to some of the most widely used digital platforms and the power relations of languages in Wikipedia, where the largest number of articles about a particular country is often in a foreign language (predominantly English).

Contending that languages represent a key ‘proxy’ indicator for global knowledge dynamics and systems, the campaign aims to contribute to “a truly multilingual Internet embodying multiple forms of knowledge” not only captured in text but also in audio-visual/multimedia forms. Arguing that epistemic justice is dependent both on testimonial and hermeneutical factors, Sengupta asked “Can you only be on the internet in your nearest colonial language?” [Sengupta]

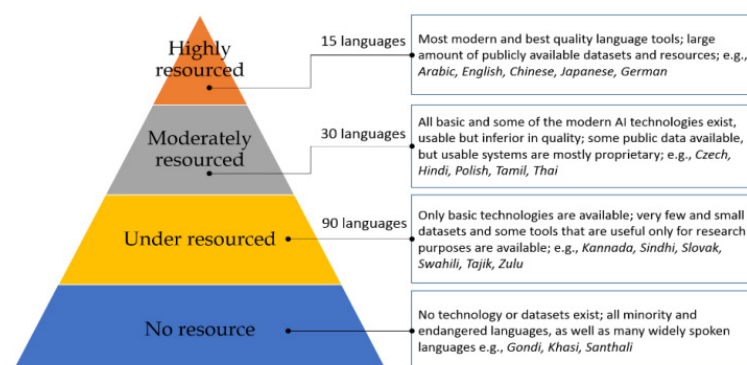
“Digital technology has never been global”

Eduard Arriaga

Eduard Arriaga, co-editor of a recent volume titled ‘Afro-Latinx Digital Connections’ (Arriaga and Villar, 2021), extended the discussion of embodied digital multilingualism through a decolonial approach which reminded us that digital technology, far from being equally distributed globally, follows historic and emerging colonial power dynamics. He identified two kinds of challenge for understanding languages in digital spaces: one *epistemological*, symbolized by the enduring strength of utilitarian models in shaping discussions about technology in mainstream discourse, and the other *representational*, namely that lower-resourced languages struggle to break out of their marginality in digital spaces due to the “Silicon Valley” mindset which still pervades much of digital practice.

Arriaga’s approach explicitly looked at these issues from an Afro-Latin American and Afrolatinx viewpoint, highlighting the lack of digitised materials developed by, or for, Afro-descended peoples in the Americas and the fact that cultural production in this context is not always necessarily defined by a textual ‘canon’. His research has put the spotlight on Afro-Latinx communities who appropriate and redeploy digital methods and technologies to suit “a more complex set of relations” embodied in their own self-representational norms and customs. Arriaga’s work encourages us to comprehend an “expanded conception of infrastructure” which goes beyond

Figure 3: Classification of languages by digital support



Source: Bali et al (2019), “ELLORA: Enabling Low Resource Languages with Technology” in LT4All Collection of Research Papers, pages 160–163, Paris, UNESCO Headquarters, 4-6 December, 2019. © 2019 European Language Resources Association (ELRA), licenced under CC-BY-NC

technopositivist views on digital culture to connect humans, technology, networks, culture, and society. In responding to the question we set panellists of “who should be involved?”, he proposed engagements between a wide range of stakeholders, from technology experts integrated in language communities, to librarians, archivists, translators, cultural practitioners, educators, policy makers and the language communities themselves. [Arriaga]

Language communities were at the centre of the panel presentation by Kalika Bali, principal researcher at Microsoft Research in India, a profoundly multilingual country with 16 languages enjoying primary official language status at state level and 29 languages with more than a million native speakers. Highlighting the high degree of concentration of AI, NLP and speech technology systems on a very small number of languages, Bali’s presentation centred on how to build technology for low-resourced languages. In global terms, no indigenous Indian language qualifies in her category of the 15 ‘highly-resourced’ global languages in terms of digital support, and most sit in the ‘under resourced’ or ‘no source’ categories. “Can we truly impact a low resource language community?” asked Bali, before presenting the work of Microsoft Research India on projects such as [ELLORA](#) (Enabling Languages with Low Resources), which aims to support speech and language systems in low-resourced settings.

This report will explore the implications for language technologies (including AI and NLP) in more detail later, but at the heart of Bali’s argument was the need to engage with language communities, to understand how they need/wish to engage with digital technology, and to build systems which allow them to access and interface with technology using linguistically and culturally sensitive human-computer interaction models. [Bali]

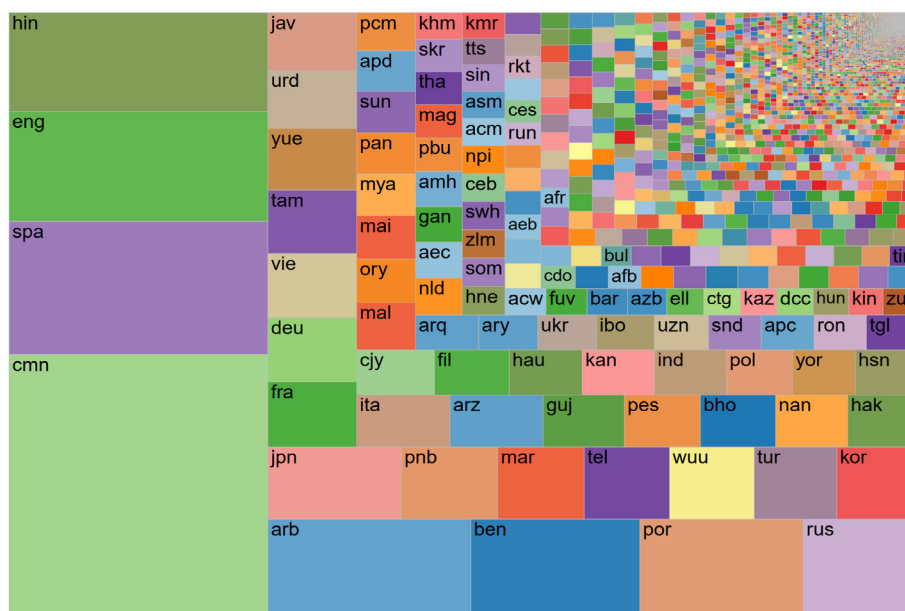
The final presentation in the panel was by Cosima Wagner, research librarian for East Asian Studies and expert in the challenges for non-Latin scripts in digital spaces. Wagner’s presentation ‘Towards multilingual enabled digital knowledge infrastructures’ looked at the main challenges for digital multilingualism in today’s academic research infrastructures. Situated at the nexus between Japanese studies, Library & Information Sciences and Science & Technology Studies, work she has carried out in collaboration with her colleague Martin Lee explores the way in which technological design shapes knowledge infrastructures and the concrete impact this has on research on/from the East Asian region when mediated by current platforms, tools usage and frameworks for multilingual discovery. We look at multilingual infrastructure in a special section on the topic later in this report, but here we should briefly mention the important work Lee and Wagner have done on exploring challenges for non-Latin scripts (NLS) in digital scholarship, which Wagner presented at the Disrupting Digital Monolingualism (DDM) event. The workshops Lee and Wagner organised in 2019 aimed to identify the particular challenges for NLS-based study in digital ecosystems and to consolidate a network of NLS practitioners and researchers to share and build dedicated digital methods and tools (Lee and Wagner, 2019). This initiative has also become part of a wider community of practice which aims to improve multilingual support and awareness in the digital humanities (and other fields of digital study) through initiatives such as Multilingual DH (‘Multilingual DH’, 2020) initiated by Quinn Dombrowski. [Wagner]

“If it’s not written it did not happen: The written bias in the digital world”

Mandana Seyfeddinipur

‘If it’s not written it did not happen: The written bias in the digital world’ was the title of the keynote presentation by Mandana Seyfeddinipur, director of the [Endangered Languages Archive](#). The DDM workshop received few proposals exploring the implications of orality for disrupting digital monolingualism, so her talk was important in exploring the “reality ... that you need to be literate in English or a few other majority languages to be able to participate in the digital world.” Given the fact that the vast majority of the world’s languages do not have a written form, this obliges speakers of those languages to seek digital access via majority languages. She described how this situation calls for action because of our political responsibility (to ensure linguistic justice),

Figure 4: Graph visualisation of Language spoken in world




Source: Created by Kilu von Prince based on data from Simon and Fennig (2017), used here with her permission

our ethical responsibility (to provide diversity support), a historical/humanitarian responsibility and an academic responsibility (to guarantee information integrity and reproducibility).

Her research suggests that this has been particularly apparent during the communication crisis around COVID-19 where, theoretically, a digital approach can amplify and accelerate access to information, but in practice it is “hampered by its monolingual and written bias”. Seyfeddinipur presented the [VirALLanguages](#) project, an initiative to “reach marginalized communities and share reliable and memorable information so people know what to do to stop the spread of coronavirus” by allowing them to share video-based health advice in their own languages “presented by their own trusted sources through the medium communities are used to”.

There is a systemic bias in the digital world, Seyfeddinipur argued, because information available is generally monolingual and written, so we have an urgent need to understand how multilingualism and orality function better digitally. The speaker concluded by advocating a more creative and language community-driven process in supporting diversity through technology. [Seyfeddinipur]

Figure 5: Non-latin scripts initiative workshop



Non-Latin scripts initiative: two community building workshops

Workshop1 (7/2018)

Title: „Non-Latin Scripts in Multilingual Environments: research data and digital humanities in area studies“

Participants: 27



Detailed report: <https://blogs.fu-berlin.de/bibliothek/2019/01/18/workshop-nls2018/#more-24479>

Workshop 2 (7/2019)

Title: "Towards Multilingualism In Digital Humanities: Achievements, Failures And Good Practices In DH Projects With Non-latin Scripts" (ADHO DH2019 Konferenz in Utrecht/Niederlande, WS program: <https://hackmd.io/s/ry0yFF1oE>)

Participants: 29

Presentations: Zenodo Community Multilingual DH <https://zenodo.org/communities/multilingual-dh/>

Source: © Cosima Wagner

Lightning Talks, demos, posters and mini-workshop: key themes

The majority of the first part of the workshop programme came from an open call for lightning talks, posters, demos and mini-workshops, which saw numerous submissions from all over the world. Here we offer a thematic overview of the presentations, highlighting the key themes which emerged.

Wikipedia/Linguistic diversity in digital knowledge production

A number of presentations took Wikipedia as a case study in global/multilingual open knowledge production which is, however, subject to wider inequalities in linguistic dynamics. Bunty Avieson analysed Wikipedia's role as "pharmakon: poison and cure for minority languages", maintaining that while it offers the potential to expand knowledge in any geography and any language, in practice the increasing cultural power of English Wikipedia potentially "poses a risk to the linguistic and cultural diversity of the world", since its volunteers are principally male, white and from the global North. As she recognised, the Wikimedia foundation is aware of the challenge, introducing the concept of 'knowledge equity' in 2017 as part of its strategic direction, and her research explored how Wikipedia might become a "cultural bulwark", which instead of "over-writing local knowledge" with a Western lens might fully realise its potential as a multimedia platform and a "dynamic site for [multilingual] renewal". Engaging with the Dzongkha linguistic community in Bhutan, she explored how Wikipedia could be used to present local histories, and how its talk pages could foster cultural dynamism. [Avieson]

In other presentations, Jorgensen studied the relationship between news topics (in this case the COVID-19 virus) and the production of pages in particular linguistic editions of Wikipedia, asking whether Wikipedia pages are indeed 'global' and assessing to what extent content is situated within specific languages and communities. [Jorgensen] Finally, in their poster Stuart Prior and Lucie-Aimée Kaffee from Wikimedia UK demonstrated how Wikipedia's sister project [Wikidata](#) relates concepts in different languages in machine-readable form in order to connect multilingual datasets, and the ethical and authorship/ownership challenges this presents. [Prior and Kaffee]

Multilingual infrastructure/Non-Latin scripts

Cosima Wagner followed her panel presentation, described earlier, with a lightning talk addressing the current state of research infrastructures from a multilingual digital humanities perspective. In her panel presentation, Wagner had outlined the following as key challenges for digital multilingualism:

- Multilingual services and solutions not developed sustainably.
- An overall knowledge deficit in how multilingual practices are conceived and practiced, with knowledge/experience often 'siloe'd' and disconnected.
- Insufficient involvement of stakeholders in research infrastructure design and over-dependence on "black box" commercial solutions.
- A fragmented research community (in geographic and disciplinary terms), making co-ordinated lobbying more difficult.

In her lightning talk, Wagner outlined an STS (Science and Technology Studies) approach to creating multilingually-aware infrastructures, based on two workshops co-convened with Martin Lee (Lee and Wagner, 2019), as part of their research into non-Latin scripts in digital spaces at the Freie Universität Berlin, which later led to broader collaborations with the multilingual digital humanities community. The particular challenges of applying digital methods to non-Latin scripts have been addressed elsewhere in some depth already from a digital East Asian studies perspective, but Wagner's presentation outlined current challenges, including:

- Limitations in the reproduction of non-Latin scripts including issues connected to directionality and image-text connection.

- Lesser availability of Optical Character Recognition (OCR)-based knowledge and content to build research on.
- The absence of agreed-upon and practiced standards for transcribed/Romanised textual content and metadata.
- Limits in knowledge about how algorithmically-based discovery works for these scripts.
- The cost of producing alternatives to the broad corporate solutions on which much of academic research infrastructures rely.

In particular, she highlighted the challenges in sustaining professional experience and knowledge about how to create and maintain multilingual digital infrastructure for non-Latin scripts (NLS) :

- Expertise is often grounded in short-term projects.
- NLS community is often peripheral to design decisions in infrastructure due to ‘small discipline’ status.
- Job instability for researchers and librarians with expertise in multilingual infrastructure, in particular related to NLS; lack of clear opportunities for career progression in this area.
- Shortage in opportunities for training and experimentation.
- Lack of a common repository of knowledge and resources.
- Challenges in providing library support for a variety of languages and scripts.

She concluded her talk by presenting a survey for multilingual-enabled digital knowledge infrastructures, which we explore in more detail under the work of Theme Group 1 in the section on the second part of the workshop. [Wagner]

Multilingual infrastructure was implicitly present in a number of other presentations, but the other lightning talk which explored this in depth from a practical perspective was a presentation by Pascal Belouin and Sean Wang on [RISE and SHINE](#), “An API-based Infrastructure for Multilingual Textual Resources”. Aiming to replicate the kind of support IIIF (the International Image Interoperability Framework) offers for images to textual resources, according to the presenters, the RISE and SHINE framework provides a secure data exchange format for multilingual texts in many different languages and formats across diverse tools and platforms. Currently supporting 130,000 resources in eight languages, the platform aims to break down data siloes and increase discovery of licensed and Open Access multilingual texts. [Belouin and Wang]

Multilingual language resources

The creation, maintenance and documentation of multilingual language resources was another major theme at the DDM workshop. Andiswa Bukula presented the work of the South African Centre for Digital Language Resources ([SADILAR](#)), whose research contributes to a wider strategy for fostering multilingual diversity in South Africa, which has 11 officially recognised languages. Bukula delineated the main features of multilingualism in the country, which, since apartheid, has benefited from policies to protect and foster indigenous languages at all levels of education. However, she also highlighted issues facing this multilingual strategy at higher education level. Developed as part of the [South African research infrastructure roadmap](#), SADILAR aims to develop digital language resources and software, including annotation tools, part-of-speech (POS) taggers and translation tools to foster multilingual digital scholarship in the humanities and social sciences based on South African language community requirements. Bukula’s lightning talk demonstrated the potential value of machine translation and other digital systems in driving these transformations, in a context where resourcing human translation is a challenge, and the wider implications this has for multilingual information/knowledge dissemination. [Bukula]

Elizabeth Marie Thaut’s presentation examined the hurdles facing “language documentation and description with(in) a digital diaspora” based on her experience with the [Sylheti project](#) in London, which brings together Sylheti speakers of different generations to increase the visibility of Sylheti as a language and to help sketch out its dialectal variation. Sylheti is spoken by more

than ten million speakers in north-eastern Bangladesh and India. Using both face-to-face and social media-led approaches across London and digital diaspora communities, the project has contributed to the “revitalisation of the endangered Sylheti/Siloti/Syloti Nagri script” with the creation of a significant community archive, Thaut explained. She emphasised the importance of integrating trained linguists and amateur language enthusiasts from language communities in an interdisciplinary and co-participatory approach which can connect language documentation to social media practices. [Thaut]

The ‘Multilingual data in ELTeC: enacting European literary traditions’ mini-workshop delivered by Ioana Alexandra Lionte and Roxana Patraş described the design and creation of a multilingual corpus focusing on European literary history, as part of the [Distant Reading for European Literary History](#) project. The mini-workshop detailed the sampling and annotation choices, the underlying linguistic and literary principles underpinning them and the implications for comparing features across a heterogenous and multilingual corpus. [Lionte and Patraş]

Multilingualism in media, platform and software studies

Peter Chonka’s lightning talk examined the effects of algorithmic power on African indigenous languages. Analysing the effects of search engine interactions for Af Soomaliga, the Somali language, Chonka drew on Safiya Umoja Noble’s work on Algorithms of Oppression (Noble, 2018) and suggested that “more research is needed into the ways in which different languages interact with the algorithmic functions of digital platforms used worldwide”. Using a case study based on the concept of ‘clan’ in autocomplete predictions/suggestions for Google Search, Chonka explored the politics of information retrieval on the Somali conflict and highlighted the implications of potentially problematic autocomplete results for how we understand identity construction, networked knowledge, and the accountability of platforms to global communities of users. He concluded that there is an urgent need to study relations and interactions between African indigenous languages and algorithmic models for Search and Social Media. This research is currently being expanded with a comparative focus on other languages in the region under the auspices of the [Datafication and Digital Rights in East Africa](#) network. [Chonka]

Scripts

“Let’s talk about scripts”, proposed Isabelle Zaugg in her lightning talk. Zaugg started by depicting the current state of scripts, and their relationship to languages. She estimated that there are approximately 300 historic or currently used scripts, and made the point that the relationship between languages and scripts is neither one-to-one nor fixed. Examples given (for one script-to-many languages) were Latin, Cyrillic or Devanagari, which are each used in various languages, and (for one language-to-many scripts) Azerbaijani, which has been represented in numerous scripts including Perso-Arabic, Latin and Cyrillic.

Zaugg argued that, while language extinction is a relatively well understood phenomenon, script displacement is also an important (and frequently overlooked) aspect of declining geocultural diversity. She responded to her own question “Why should we care?” by arguing that language and script loss lead to “the loss of intergenerational knowledge, cohesion, and identity”. Recognizing that scripts are subject to change through displacement, she picked out the spread of Latin script due to “dynamics of colonization, evangelization, hierarchical beliefs about its superiority, and gaps in digital supports” and identified the risks of transliteration for script integrity. The “market logic” behind language technology has led to disparities in digital support for many languages, particularly those represented in non-Latin scripts. This in turn affects both the language and script’s “digital vitality”, leading to losses in function, prestige and competence within many digitally disadvantaged language communities. Even for well-supported languages like Arabic, Zaugg explained, the tendency to transliterate into Latin affects “script linguistic integrity”.

She provided the case study of an Ethiopic character which typically represents three levels of meaning – as (1) a syllable, (2) a religious concept associated with the Ethiopian Orthodox

Tewahedo church and (3) a numerical significance – and so is impossible to represent in transliterated form. Zaugg closed her talk by arguing for:

- better Unicode support for currently under-resourced scripts through initiatives such as the [Script Encoding Initiative](#) (SEI).
- greater keyboard and input support – “users will not always have the most updated devices or know how to use them”.
- more language-diverse (and language-specific) NLP tools, including spellcheckers, autocomplete, autocorrect, OCR, voice dictation and machine translation functionality.
- a reversal of the current model, whereby language and script communities are forced to adapt to digital technology – digital technology should adapt to the needs of language communities.
- lobbying ‘Big Tech’ for better language and script support – “this should be at the core of their corporate social responsibility to serve global audiences”. [Zaugg]

Language technologies, including NLP

In her panel presentation (summarised earlier), Kalika Bali had asked the following questions in relation to language technologies:

1. “How many resources are available across the world’s languages, and do they correlate with the number of speakers?”
2. Which typological features have NLP [Natural Language Processing] been exposed to and which features have been underrepresented?
3. How inclusive has ACL [[Association for Computational Linguistics](#)] been in conducting and publishing research for different languages?
4. Does resource availability influence the research questions and publication venue?
5. What role does an individual researcher or community have in bridging the resource divide?”

Some of these questions have been explored in a recent paper on linguistic diversity at NLP conferences (Joshi et al., 2021).

As Bali and others at the workshop demonstrated, there are a number of key challenges for low- resourced languages and communities when engaging with language technologies. Bali presented the ‘Enabling languages with low resources’ ([ELLORA](#)) project, which aims to develop speech and language systems in areas of the world with limited access to resources. She outlined ELLORA’s work in three principal areas:

- Data – investigating how to produce the type and volume of data required to make language technology work tractable, including community-led efforts to generate data.
- Bootstrapping – adapting models and techniques applicable to higher resource languages to lower resource languages.
- Community-focused applications – analysing what communities actually need and what techniques are most applicable to these needs: ‘verticals with social impact’.

“Can we truly impact a low resource language community with technology?”

Kalika Bali

The underlying question, Bali explained, is the following: “Can we truly impact a low resource language community with technology?” Among the case studies Bali presented were the [Karya project](#), which aims to use paid crowdsourcing approaches to generate local language data at

scale, while fostering digital literacy and inclusion. According to this research, “most of the population is very capable of working on these platforms”. Another case study discussed a workshop on language technology for Gondi, a Dravidian language, spoken by about three million Gondi people in India. The language has no technology support but a highly engaged community, and the workshop brought together various stakeholders (including the community itself) to brainstorm potential applications of language technology for the target community. This exercise has since led to the development of Gondi linguistic resources, including the creation of children’s stories in the language and a radio app called Adivasi, which allows the community to consume news in their language using text-to-speech technology – for further information see Mehta et al., 2021. Bali emphasised that language technologies should be co-designed with language communities rather than being designed “for” them and warned of the dangers of assuming that “universalist” technical solutions will work for all languages equally; sometimes digital systems need to be adapted or designed for specific languages/communities. [Bali]

In recent years much attention has been given to differing levels of support for Natural Language Processing in various languages and the concentration of resources/research in a small group of languages (in particular English). AI and NLP multilingual challenges were the key focus of one of the theme groups to be discussed in the section on Part Two of the workshop, but here we will briefly review two presentations which explored NLP in Part One. The first of these, a demo by Andrew Janco, was built on the premise that creating NLP tools for new languages is an important feature of linguistic diversity in digital scholarship, as well as a means to help researchers generate linguistic data. Janco’s team have developed a web application called [Cadet](#) which can be used to create training data for a new language model for [spaCy](#) (an NLP library in Python) by employing automated suggestions from users and responsive learning methods. This tool makes it easier to annotate text in order to train a new language model, and in so doing helps to extend linguistic coverage in NLP research. [Janco]

In his lightning talk on ‘Multilingual Transformers: Linguistic Relativity for the 21st Century’, Michael Castelle explored a particular form of digital multilingualism present in AI-based architectures. Grounding his presentation in the waves of modelling architectures driven by deep learning, which have profoundly shifted how NLP is performed in the last five years or so, he critiqued the essential theories and “ideology of architecture” represented by such models, and made comparisons between the “geometric transformation in high-dimensional space” which they embody, and anthropological theories of linguistic relativity proposed a century ago by thinkers such as Edward Sapir, who conceived of translations as “psychologically parallel to passing from one geometrical system of reference to another” (Sapir, 1924). [Castelle]

Low-resourced languages

Several presentations addressed the particular challenges facing low-resourced, endangered or minority languages. Leonore Lukschy’s lightning talk appraised the extent to which endangered languages are constrained by archival practices, signalling a contradiction between the immense quantity of language data relating to “spoken, signed, whistled and drummed” languages and the access barriers which the affected communities face when attempting to access their own language resources. The principal reason for this is that interfaces or metadata relating to collections are often only in ‘majority’ languages (in particular English), and Lukschy argued that, on one level, large endangered languages archives “contribute to the linguistic exclusion of the speakers represented in their collections”.

Why is this so, and how could this be changed? The presenter gave us a user perspective on what it is like to navigate and deposit language data in an unfamiliar language and script, and she identified the following barriers to archiving endangered language content:

- The archive homepage. How easy is it to open an account, to understand what the archive contains or to contribute content?
- The catalogue interface. If the user does not understand the language(s) supported by the resource, how easy will it be for them to work out how to discover content?

- Metadata and language labels. Even if the user is able to navigate the interface, will they be able to effectively search for content based on descriptions in an unfamiliar language?

Figure 6: Viral Languages



Source: Virallanguages project

Lukschy acknowledged that while in “an ideal world an archive’s interface would be available in every language represented in its collections”, this is expensive and impractical. She outlined a series of strategies to create and maintain multilingual archival data in a sustainable way, including:

- Template-based guidelines for deposit using screenshots to overcome linguistic barriers.
- Multilingual metadata.
- Measures to address the textual bias, such as audio-visually based collection guides.

[Lukschy]

In her lightning talk Sarah McMonagle explored challenges in researching minority languages based on her work with bilingual adolescents in the Sorbian-speaking region of Germany. Recognising the importance of involving diverse stakeholders (representing language research, policymaking, digital media companies and speech communities), she asked how digital media influence language choice, and hence our understanding of multilingualism. In the sample her study researched, young people were confident in using Sorbian in their everyday lives, but often preferred German when using social media, and while they agreed with the statement that the internet is multilingual, this was sometimes interpreted in terms of major languages rather than minority languages. McMonagle discussed differing perceptions of multilingualism between speech communities and language researchers, and examined how our perceptions affect minority languages research involving digital (and social) media. [McMonagle]

Jessica Green’s poster was titled ‘Around the British Library in 40 Languages: Engaging with a Different Community Each Week #AToUnknown’. It catalogued an initiative by the Heritage Made Digital Team at the British Library to “promote awareness and engagements of BL digitised collections through the lens of world languages” on Twitter, which consisted of a weekly focus on different languages present in the British Library’s collections. The project included languages represented in the BL’s [Endangered Archives Programme](#) and those using non-Latin scripts, aiming to connect communities and to provide them with opportunities to engage with digitised versions of content in their languages and external related resources through a playful, interactive approach involving games such as a weekly sudoku in the chosen language. [Green]

Carlos Yebra López’s lightning talk explored the case of Ladino (or Judeo-Spanish), a diasporic minoritised language mostly spoken by the Jewish people expelled from what are now Spain and Portugal at the end of the 15th century. His presentation detailed how the growth of the World Wide Web helped the Ladino language community to overcome geographic isolation,

to promote inter-generational linguistic exchange/transmission and to foster the preservation of Ladino by facilitating diglossic isolation. Using Held's concept of 'digital home-lands' (Held, 2010), Yebra Lopez presented three case studies of Ladino-focused online communities: Ladino Forever (on Facebook), [Ladino 21](#) and the recent [uTalk course in Ladino](#). [Yebra López]

Figure 7: Postcard produced by participants to represent the many languages of Prato and advertise the La Nostra Prato exhibition



Source: © Youth in the city project

Transcultural and translingual dynamics

Transcultural and translingual dynamics were present throughout the workshop and were one of the four themes we had chosen for the 'theme groups' in the second part of the workshop. In his lighting talk on a [project to represent multilingual youth voices in urban spaces](#), Matteo Dutto assessed the potential of collaborative digital storytelling in mapping transcultural dynamics. Using the Italian city of Prato as a case study, the project aimed to challenge dominant narratives about the city and capture its cultural diversity through digital mapping, data visualization, open participatory storytelling techniques and free online educational materials. Through workshops, exhibitions, curated digital interactions and geolocalized stories, the project demonstrated the potential impact of multicultural urban youth on a city's self-perception. [Dutto]

Language pedagogies and platforms

There were two demos on language pedagogies at the workshop. Ethem Mandić showcased a free online language learning platform for the standard Montenegrin language, which drew on the collaboration and resources of the Montenegrin diaspora in a bid to present the language within its historical and modern cultural context for speakers of English, Spanish, Turkish and Albanian (the four most common languages used in the diaspora of this language community). [Mandić] Meanwhile, Caoimhín Ó Dónaill presented [CLILSTORE](#), an open online educational platform for language learning, based on Content and Language Integrated Learning (CLIL) principles, which allows various types of multimedia language resources (including text, images, audio, video, and web apps) to be created, shared and consumed in open access. Employing a rich set of metadata to facilitate discovery (including [Common European Framework Reference level indicators for languages](#), descriptive and technical summaries), the project also provides automatic interlingual referencing via a network of online dictionaries so that a learner can instantly check unfamiliar

vocabulary directly from their reading context using a split-screen interface. [Ó Dónaill]

Multilingualism in practice

In ‘what language(s) should work on multilingualism take place?’ the lightning talk by Ernesto Priani asked ‘¿Estás en un proyecto multilingüe? Considera esto! (Are you in a multilingual project? Consider this!)’. Taking an international research project studying multilingual 19th Century newspaper content using digital humanities as a case study, Priani’s presentation was one of the few speakers at the event to not present in English. We explore the reasons for this general tendency elsewhere in this report, but Priani exhorted the audience to build multilingualism into our everyday practice in research communication. Observing that, even in multilingual projects, the question of what language to communicate in is often not even raised for discussion, and that English is used as an automatic lingua franca in many contexts, he urged us to “avoid the easy path” by (1) being language conscious, (2) identifying and encouraging communication opportunities in linguistic subgroups, (3) translating project materials and (4) publishing “in as many languages as possible”. [Priani Saisó]

Pedro Nilsson-Fernàndez made similar points in his lightning talk titled ‘Digital Peripheries: A Postcolonial Digital Humanities Approach to Catalan 20th Century Literary Spaces’. Starting his presentation by reflecting on debates within the digital humanities about its geolinguistic diversity, he observed that ‘DH’ has, to some extent, “morphed” from a monolingual community to a bilingual (English and Spanish) one in global settings, but that this is more driven by a “sense of the utility” of languages spoken by large sections of this digital field than by “an interest in multilingual approaches”. Anchoring his argument in postcolonial approaches to digital humanities advocated by Risam and others (Risam, 2018), Nilsson-Fernàndez presented his own efforts to contribute to the reconstruction of the Catalan national literary space in the 20th century using an array of digital humanities methods, including text mining and GIS. By capturing spatial references in the work of a key Catalan author of the period, Manuel de Pedrolo, he explored how digital tools can challenge prevailing concepts of space. Nilsson-Fernàndez concluded by arguing that fostering multilingual approaches in research is a key element in diversifying its cultural reach, and that we should not just see multilingualism as a “focus of study” but that we should also practice multilingualism in our dissemination of research. In his words, in an “era of automatic translation engines and mass consumption of subtitled popular media on platforms such as Netflix, multilingual scholars are missing a great opportunity”. [Nilsson-Fernàndez]

Part two - future directions

Whereas the first part of the Disrupting Digital Monolingualism workshop was intended to reflect the state of the art in digital multilingualism, the second part aimed to bring together various experts to explore future strategic priorities and co-design creative responses (potentially including specification documents, prototypes, white papers, or manifestos) to specific multilingual challenges. The workshop as a whole principally aimed to address four themes, which can be summarised as follows:

- **Linguistic and geocultural diversity in digital knowledge infrastructures**
 - Exploring how digital infrastructures reinforce or redistribute linguistic and cultural influence or authority, and how can we challenge this authority.
- **Working with multilingual methods and data**
 - Examining to what degree digital data practices or research reflect linguistic and cultural difference. How does the growing datafication of culture and society affect the balance of world languages, and what steps can be taken to foster greater diversity and language awareness?
- **Transcultural and translingual approaches to digital study**
 - Investigating the new possibilities offered by digital media for studying translingual and transcultural dynamics. For example, how can we effectively study intercultural, plurilingual or transborder perspectives using digital tools?
- **Artificial intelligence, machine learning and NLP in language worlds**
 - Identifying the challenges which AI, machine learning and Natural Language Processing pose for language research fields and professions. What opportunities does AI provide to foster linguistic diversity or intercultural communication?

Once the event went online, we decided to create four **theme groups** around these themes, which would run for one week after the first part of the workshop, each led by two facilitators who would design and manage their group focus based on our initial proposal and prompts (see Appendix 2 for questions we set each group to get debate started). Different groups were at liberty to organise themselves as they saw fit over the course of the week 17th-24th June 2020, with a final meeting between the workshop organisers and the theme group facilitators on the 24th of June. Final decisions about everything else lay with the facilitators, but we proposed that each group establish its *modus operandi* early on, define any outcomes it wished to achieve (including reports on activity) and decide how public it wished to make its activity while the groups were still operative (including the possibility of publicly asking for contributions). Groups were welcome to narrow their focus, to collaborate with each other and to use social media (if that matched their objectives), and we encouraged groups to make the in-group conversations multilingual where possible. We also provided local contacts to support groups in case they needed practical assistance.

We summarise the composition of the theme groups below:

1. Linguistic and geocultural diversity in digital knowledge infrastructures

- [Focus chosen: requirement profile for multilingually enabled digital knowledge infrastructures with a special focus on non-Latin scripts]

- Facilitators: Cosima Wagner, Cornelis van Lit and David Wrisley
- Local contact: Paul Spence

1. Working with multilingual methods and data

- [Focus chosen: using multimodal corpora]
- Facilitators: Miguel Escobar and Darja Fišer
- Local contact: Kristen Schuster

3. Transcultural and translingual approaches to digital study

- Facilitators: Sender Dovchin and Emanuela Patti
- Local contact: Naomi Wells

4. Artificial intelligence, machine learning and NLP in language worlds

- Facilitators: Kalika Bali and Quinn Dombrowski
- Local contact: Gabriele Salciute-Civilienne

Final outcomes are published on the ‘theme groups section’ of the workshop website, but we will now briefly summarise how the focus of each group was shaped, and how their outcomes connected to other workshop discussions.

Theme group 1: Requirement profile for multilingually enabled digital knowledge infrastructures with a special focus on non-Latin scripts

The general topic for the first theme group was *Linguistic and geocultural diversity in digital knowledge infrastructures* and the questions we proposed to set off discussions were the following:

Figure 8: Disrupting Digital Knowledge Infrastructures: a Status Quo Survey

Disrupting Digital Knowledge Infrastructures: A Status Quo Survey

Many of us work in internationalized universities, in which researchers use different languages for research and scholarly communication. We know that our research infrastructures can fall short, however, in accommodating our linguistic and geo-cultural diversity.

As a follow up to the "Disrupting Digital Monolingualism" (DDM) conference at King's College London (16-17 June 2020) our working group aims to draft a "requirement profile" for digital platforms (i.e. repository, WordPress, library catalog, XML editor, e-journal, etc) in which people use different languages. By "requirement profile" we mean a list of minimal standards that platforms would support to be accessible to different users. We are particularly interested in standards supporting non-Latin scripts.

So, if you are using more than one language or non-Latin scripts in digital environments then we want to hear from you in this survey about issues, workarounds, and downright dealbreakers that you experience every day.

What we will do with these answers: We will be collating the different answers we get for the purposes of including them in a short white paper, hopefully to provide developers and infrastructure specialists with design specifications. The white paper might also be published as part of a conference report by the DDM conference organisers by the end of June 2020.

How we will acknowledge your contribution: You can offer your opinion anonymously OR give your name and email. If you provide your name, respond to the questions and agree we will include your name on the white paper as one of the contributors. Please list it as you would like it to appear in the white paper.

- In what ways do digital knowledge infrastructures embed multilingual or language-focused practice at present, in social and technical terms?
- How do digital infrastructures reinforce or redistribute linguistic and cultural influence

authority?

- What multilingual and localisation responses from industry, academia or the third sector have been most effective so far?

The theme group was led by three facilitators, all with ample experience in this area: Cosima Wagner's co-organisation (with Martin Lee) of workshops on digital humanities approaches to non-Latin scripts in 2019 (Lee and Wagner, 2019) has led to numerous other multilingual initiatives in the field; David Wrisley's involvement includes multiple events on Right-to-Left languages in DH (including the DHSI conference co-chaired with Kasra Ghorbaninejad in 2021); and Cornelis van Lit is founder of [The Digital Orientalist](#), which aims to make "the digital world more inclusive to non-Latin languages, ancient or modern, dead or alive" (van Lit et al., 2020).

After some debate, the group decided to focus on producing a "requirement profile for multilingually-enabled digital knowledge infrastructures with a special focus on non-Latin scripts," which would serve as a reference for organisations creating or managing digital scholarship infrastructure in the future. This 'minimum requirement list' was informed by various stakeholder perspectives (including researchers, library and information professionals, infrastructure policymakers and digital practitioners) and was designed to foster evaluation of the multilingual support provided by different digital research platforms (including those used for web authoring and editing, XML-based content creation, catalogues, digital publishing and repository management). The group launched a survey titled '[Disrupting Digital Knowledge Infrastructures: A Status Quo Survey](#)' just before the DDM workshop to help shape these guidelines, which asked the following questions:

1. What languages (other than English) and scripts do you use in digital environments in your daily work? List them in order of their importance to you.
2. What kinds of difficulties do you encounter when you use more than one language and/or non-Latin scripts in a digital environment?
3. What would you like to be able to do using your languages that you can't right now? If you can, specify the kind of digital environment (repository, WordPress or other web publishing platform, library catalog, XML editor, e-journal, etc).
4. Do you have workarounds to get a task done that would seem to be so simple but is actually not, due to difficulties described in the second question of this survey?

The survey was open for five days and received 51 responses. The group's initial report is available on the workshop website, and will be followed by further publications, including the minimum requirements list itself, which will function as a "practical checklist" to facilitate content creation, analysis and discovery in "multilingually enabled knowledge infrastructures" (van Lit et al., 2020). Materials will eventually be published on Zenodo's '[Towards Multilingualism in DH](#)' community.

The group will report on its findings in full at a later date, but discussion included:

- Challenges with content generation, and support for multilingual input, display and interaction with information (in particular for non-Latin script languages). Specific issues mentioned included: consistency in actions such as copy and paste; different transcription conventions; punctuation; and texts containing multiple languages (or script directionality) in the same space.
- Within these challenges, there was a particular focus on limitations in current keyboard configuration – including the ability to switch between languages, scripts, and directions – and font representation/relative sizing.
- Search engine/discovery platform limitations in capturing NLS and Right-To-Left languages.
- Insufficient training data for data-driven approaches and the need for better NLP tools in many languages.
- The possibility of establishing a multilingual certificate for digital scholarship infrastructure, which could demonstrate that a given infrastructure meets commonly-agreed minimum

multilingual criteria/expectations.

“What are the technical, legal and ethical issues in building non-English multimodal corpora?”

Theme group 2: Multilingual challenges in the use of multimodal corpora

The second theme group was asked to reflect on challenges in *Working with multilingual methods and data*, and in so doing, to contemplate the following questions:

- What possibilities exist for ‘data-driven’ methods in languages-based research? What triggers and blocks influence their use?
- How well do digital data practices or research reflect linguistic and cultural difference?
How does the growing datafication of culture and society affect the balance of world languages, and what steps can be taken to foster greater diversity and language awareness?
- What new forms of digital, linguistic and cultural criticism are needed to interpret data-driven approaches to languages and multilingualism?

The group was facilitated by two people with widespread experience in working with digital methods and data: Darja Fišer, whose role in CLARIN (a European research infrastructure initiative for language as social and cultural data) has included maximising the impact of its research methods, tools and data, and Miguel Escobar Varela, who has widespread experience with a range of computational humanities and cultural analytics research methods, in particular related to South East Asian theatre, and who has recently explored how to communicate cultural specificity through digital interfaces (Escobar Varela, 2020).

The group opted to narrow the focus of their theme to *multilingual challenges in the use of multimodal corpora*, and the facilitators sought to build a team of researchers and practitioners dedicated to:

- Technical, legal, and ethical issues in building non-English multimodal corpora.
- Quantitative, qualitative, and mixed methods for researching multimodal corpora.
- Using multimodal corpora for pedagogical purposes and therapeutic/medical purposes (e.g., sign language, language impairments, brain damage, etc.).

As a preparatory step, participants were asked to summarise their experience (including their response to specific challenges), key research strategies and other resources for the group to draw on. Theme group members were asked to submit thought-pieces addressing these points and this was followed up by a two-hour virtual meeting held on June 22nd, 2020.

One of the key tasks the group set itself at the outset was to examine how researchers track concepts in multimedia corpora across languages and modalities, and this proved to be an ambitious enterprise. Even after narrowing the focus, the range of disciplines involved (such as linguistics, history, or social science-based fields) and their (at times vastly) differing research needs, meant that the group was faced with an array of practical and epistemological challenges, including:

- Challenges in multilingual archiving, including the need for standards to foster interoperability between multilingual corpora.
- Systems and standards for encoding gestures in multimodal and intercultural corpora.
- The contradiction between creating general standards-based approaches to facilitate interoperability across languages or knowledge communities, on the one hand, and the

need for cultural specificity and differentiation in research processes, on the other.

- Multimodality in language learning pedagogies, integrating linguistic and cultural content, and including multi-location telecollaboration.
- Ethical and legal issues which may restrict open access; even as open access content is much needed in this area.
- Measures to nurture multilingual communication within research groups themselves.
- Addressing the lack of training data and processing tools to annotate large corpora in many languages and the need to extend these tools to multimodal content.
- The desirability of greater collaboration between researchers in different fields (including linguists and NLP researchers) and between public and private sectors.
- Calls to give corpora greater recognition as a formal research output.

The group's summary concluded that, while epistemic differences need to be understood and respected, we nevertheless need to foster more "conversations across disciplinary boundaries" and to take further measures to port methods and data between different research communities in working with multilingual, multimodal corpora.

"What challenges and possibilities do digital media offer for studying translingual and transcultural dynamics?"

Theme group 3: Transcultural and translingual approaches to digital study

The third theme group explored the challenges and new possibilities which digital media offer for studying translingual and transcultural dynamics, with the following initial questions set to frame discussion:

- What new possibilities do digital media offer for studying translingual and transcultural dynamics? For example, how can we effectively study intercultural, plurilingual or transborder perspectives using digital tools?
- What is the role of digital practitioners (in research or industry) and language experts (such as translators, linguists, and modern languages researchers) in studying these interactions?

The facilitators for this group were Emanuela Patti, specialist in comparative cultural studies, who has published widely on digital cultural studies (especially in relation to Italian interart/intermedia representations), and Sender Dovchin, whose work focuses on the experience of culturally and linguistically diverse youth in Australia, contributing to the second language learning of youth living at the margins. Once constituted, the group set its own forum questions, which were debated in a live online discussion on June 18th, 2020.

- What challenges and possibilities do digital media offer for studying translingual and transcultural dynamics?
- How can we effectively study intercultural, plurilingual or transborder perspectives using digital tools?
- What is the role of digital practitioners (in research or industry) and language experts (such as translators, linguists and modern languages researchers) in studying these interactions?
- How can digital humanities and transcultural studies collaborate?

The theme group broadly integrated two areas of research, one focusing on Modern Languages/

transcultural approaches to digital and the other on sociolinguistics, with specialism in digital linguistic diversity and translanguaging. Both groups shared concerns about the challenges in using digital tools – with training and critical digital literacies being particular areas of concern. There was considerable consensus around the potential benefits of digital methods for studying transcultural/translingual dynamics (for example to study migration or identities in motion), but also concern that digital catalogues and archives were not multilingual enough, and generally do not offer balanced support across languages, with minority languages being greatly understudied and under-resourced. Funding, and different levels of infrastructure support for different languages were identified as problems, which reflect a “problematic relation between the politics of information and politics of representation of [minority] cultures and languages, demonstrating the powerful interconnection between digital infrastructures and agency”.

There was a strong sense that more dialogue was needed with digital practitioners and researchers to co-create research-driven tools and methods, so that they could “fully capture the complex data created by the mixture of multiple different linguistic resources and codes”. Finally, the group felt that digital studies would benefit from greater expertise in transnational and comparative perspectives encompassing a wider range of languages and geocultural points of reference.

“What concrete steps can be taken by technology leaders, academics, and users to ensure that language technology moves towards multilingualism?”

Theme group 4 - Artificial intelligence, machine learning and NLP in language worlds

Theme group 4 was tasked with analysing key challenges in the world of Natural Language Processing, including both academic and industry perspectives. We gave the group the following starter questions:

- What challenges do AI, machine learning and Natural Language Processing pose for language research fields and professions?
- What implications are there for translation and for studying or researching modern languages and their cultures?
- What opportunities does AI provide to foster linguistic diversity or intercultural communication?

The two facilitators for this theme group were Quinn Dombrowski, coordinator of a number of languages-centred initiatives including the [Multilingual DH network](#), who has recently explored the [history and future of linguistic diversity in the field of the digital humanities](#), and Kalika Bali, whose research on providing digital responses to lower resourced languages has been described earlier in this report, and whose [TED talk on the multilingual challenges for language technology](#) has been widely viewed.

They identified a mixture of NLP practitioners and academics, and asked them to attend an initial virtual meeting presenting the theme group aims and to then respond to a series of daily questions which functioned as prompts on key topics related to working with NLP tools in languages other than English (questions were also posted on Twitter in order to encourage a wider set of responses):

- Day 1: What are the biggest challenges that you’re facing in building or using NLP tools?
- Day 2: How does the dominance of English in NLP tools and technology affect work in other languages?
- Day 3: ‘NLP is about engineering and you don’t need to think about sociocultural aspects

of language technology.’ [Discuss]

- Day 4: What would you like to talk to someone in another community about (e.g. talking to an NLP technologist as an academic)? What perspective could you offer to someone in another community, based on your position?
- Day 5: What concrete steps can be taken by technology leaders, academics, and users to ensure that language technology moves towards multilingualism?

As with other theme groups at the workshop, the group was very diverse in terms of regions, languages, disciplines, and professional profiles, with people from academia, industry and start-ups. They included both people who build systems using NLP and others who use AI and NLP regularly (or analyse their social impact). The group will report on its progress separately in a white paper, but some of the issues identified included:

- A shortage of NLP tools, corpora, training data, tutorials, documentation, and examples for a large proportion of languages. A lack of experts to create datasets and appropriate collaboration spaces for multilingual NLP.
- The dynamic, but *ad hoc*, nature of tools available, and the fact that there is no single reference point for knowledge and tools - to understand what is available, in what language, and with what functionality.
- NLP systems often have limited exposure to different language typologies and assume that, because they work for English (or another highly resourced language), they work for all languages.
- There were specific issues around particular features of languages (for example, in using word count methods with morphologically agglutinative languages), the lack of written standardisation or the kinds of data available in particular languages (including data sparsity/conflicts or aligned translations).
- Recognising the importance of the “sociocultural context of language use when developing NLP”.
- A deficit in understanding about the wider issues facing world languages among technologists building NLP systems.
- The high concentration of NLP research on English (or a small number of other highly-resource languages) creates a particular set of benchmarks which undervalue the social, cultural or community context and make it difficult to get recognition (or funding) for NLP research focusing on other languages.
- The difficulties in fostering academic-industry or interdisciplinary collaborations in this area due to different credit or validation dynamics.
- These discussions habitually neglect the actual language communities – which in this case are largely absent from discussions on (and therefore have little agency over) the development of multilingual NLP.

The group concluded by highlighting the need for generating greater awareness around the requirements of multilingual NLP and advocating for further opportunities for future collaboration between scholars, industry professionals and language communities, which they explore in more detail in their White Paper.

See [all Theme Group Summaries](#) on the DDM workshop website.

“Digital Multilingualism requires input from numerous professional and academic fields”

Concluding the theme groups

The theme groups encompassed a wide range of professional and academic backgrounds,

including speech and language technology, library and information science, minority language research, technologies for low resource languages, AI and NLP, cultural analytics, spatial humanities, modern languages, area studies, visualization, sociolinguistics, critical media theory, second and foreign language learning, diasporic studies, web archive studies, science, and technology studies, inter-cultural studies, language justice/policy, digital humanities, research data management and open science.

At the end of the week of Theme Group activity a de-briefing meeting was held between organisers and the co-facilitators from the theme groups. Each group reported on how they had organised themselves and the outcomes they hoped to achieve going forward, and their individual conclusions are included in the Theme Group summaries above. In this final summary of the Theme Group work, the report aims to draw out some common themes and issues which existed across individual groups.

- **Time pressures.** The unique conditions of pandemic-induced lockdown meant that there was very little time for everyone to respond – in some cases the group composition was only finalised in the days before the event, which was challenging, although some argued that this was also an advantage because it led to more agile forms of working (!).
- **Discussion is still often fragmented or marginalised, but there is considerable interest in digital multilingualism right now.** One contributor related a feeling of marginalisation when discussing multilingualism (in her case at university), which typically happens across numerous (and largely disconnected) fields, and this tallied with feedback from the first part of the workshop, where numerous attendees reported that the biggest difficulty in disrupting monolingualism lies in finding other people to collaborate with on projects. At the same time, participants reported a widespread hunger to engage with the topic and this workshop showed the variety, and quantity, of people interested in disrupting digital monolingualism.
- **Digital Multilingualism requires input from numerous professional and academic fields.** The multiple roles, skillsets and disciplinary backgrounds make for a more rounded response to digital monolingualism, but also present challenges in how we talk across diverse perspectives, where the meaning of particular concepts and terms may be interpreted quite differently. It is not enough to just *want to* work across fields, we need to give thought to how to encourage mutually constructive discussion across communities.
- **Bridging academic and professional fields.** The experience of collaborating across professional and academic divides was seen as particularly valuable, and participants were keen to continue seeking opportunities for dynamic discussions/collaborations around digital multilingualism in the future.
- **Shared infrastructure.** This epistemic work is important if we want to achieve meaningful cross-usage of data and tools, and we need to find shared infrastructure which allows diverse fields to engage more meaningfully with digital multilingualism. This requires attention to metadata, standards, sustainable and scalable models for open data and shared workflows, while taking into account legal, ethical and privacy issues.
- **Training requirements.** How can we ensure that a broader group of researchers and professionals have access to the necessary skills to engage multilingually with digital methods and data?
- **Intercultural exchange.** It is not just diversity in epistemic culture which is a challenge; different geocultural contexts require a more rounded transcultural and translingual approach which goes beyond both the national/regional topology of research policies and the anglophone/‘big language’ bias of large digital media ecologies.
- **Digital multilingualism is not just about highly-resourced languages.** As Kalika Bali commented during the workshop, “making a tool available in English and five other languages does not make something multilingual”.
- **Cross-language work.** There are currently significant limitations in understanding and responding to how communities cross languages (which also affects translingual

research) and more work is needed to better understand this, and to facilitate cross-lingual interactions. For example, even where progress is made in developing language-specific tools and technologies, these may still reinforce bounded conceptualisations of languages that do not reflect the daily translingual practices of groups and individuals, and these in turn potentially further constrain translingual creative and research practices.

- **Involve language communities.** The community perspective is often overlooked – people designing digital research, tools and infrastructure need to actively engage with language communities.

Conclusion


The *Disrupting Digital Monolingualism* workshop sought to draw together a wide range of stakeholders active in confronting the current language bias in most of the digital platforms, tools, algorithms, methods, and datasets which we use in our study or practice, and to reverse the powerful impact this bias has on geocultural knowledge dynamics in the wider world. The workshop aimed to describe the state of the art across different academic disciplines and professional fields, and foster collaboration across diverse perspectives around four points of focus: Linguistic and geocultural diversity in digital knowledge infrastructures; Working with multilingual methods and data; Transcultural and translingual approaches to digital study; and Artificial intelligence, machine learning and NLP in language worlds.

As noted earlier, transforming the workshop from a primarily on-land face-to-face affair into an online virtual event at short notice proved challenging, despite the creativity, agility and commitment shown by contributors and local team alike. Although the programme was internationally diverse both in geographic representation and languages covered, the fact that the initial plan was for an event in London/UK meant that there was a strong bias towards presentations in English, and this raised wider issues about how to organise a virtual, global, and multilingual event, including attending to time zone differences and multilingual communication at the event itself. Nevertheless, the main objective of the workshop was met, which was to increase awareness and connection between different fields which often do not have opportunities to communicate fluidly between each other, in an online and inclusive discussion space which brought together languages, communities, culture and digital practice across multiple time zones. The switch to an online only event also allowed many people to be involved who likely would not have been able to attend an in-person event.

Since the workshop

How has the wider outlook for digital language diversity changed since the workshop? Language technology continues to evolve swiftly, and techno-centric agendas dominate, to a large extent, public debate and strategic discussions about the language-related digital media systems we use and which now dominate our lives. A recent report on ‘The Dawn of the Human-Machine Era: A forecast of new and emerging language technologies’ by the [LITHME](#) network contended that the increasing integration of technology with our senses will lead to major transformations in how we engage with languages in the coming years. The report forecasts changes in how we speak *through* technology and how we speak *to* technology. In speaking *through* technology they assert that human/technology boundaries will blur further and our interactions with machines will be expanded to include responses to minor finger movements and brainwaves. According to the report, “intelligent eyewear and earwear” will allow us to hear instant translation of what someone else is saying in another language and will manipulate our perception of their facial movements with visual overlays so that it will look like they are actually using the translated words. In speaking *to* technology, they claim chatbots and other AR/VR experiences will produce lifelike virtual characters who will be able to take on certain language-centred communicative and learning acts, which poses interesting opportunities and challenges for professional sectors such as language learning. The report also recognises some of the social and ethical risks and limitations, in particular for those dependent on sign languages and those using under-resourced languages. In one section the authors examine models to decolonise speech and language technology (Sayers et al., 2021).

The controversy in December 2020 around the departure of Timnit Gebru from the Google

Ethical AI team which she co-led resulted in a campaign of support from various co-workers at Google, diversity advocates and the wider research community, and led to expressions of concern over the state of ethical approaches to AI in large digital media companies. The trigger for the incident was a co-authored paper ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ’ which analysed the environmental, economic, financial and diversity risks of the ever-larger language models which predominantly dictate the priorities of current NLP research, and which in turn underpin key functionality such as online search and translation in many digital systems. The paper was unusual, within digital studies, in situating linguistic diversity within wider issues of (gender, race, ethnicity, and disability status-related) bias and, while we do not have space here to reflect on the wider discriminatory or hegemonising dangers the paper identifies, for the purposes of this report it clearly articulates the manner in which language data generated by these language models (and the language technologies they enable) may be unrepresentative and exclude languages and communities which are poorly represented online (Bender et al., 2021).

“The current trend in building language technology is designed to work on languages with very high resources in terms of data and infrastructure”

Joshi et al. 2019

Confirming views expressed both in this article and during the DDM workshop, Joshi et al have asserted that “the current trend in building language technology is designed to work on languages with very high resources in terms of data and infrastructure”. This approach, they argue, neglects one of the core values of NLP which is to “build systems that add value to its users” (Joshi et al., 2019). As they note, this tendency in language technology research, dependent on large data models, consolidates existing social, economic and geographic vectors of exclusion.

A survey of European language technology researchers and professionals published in 2017 found that 40% of respondents identified “insufficient research being done for minority languages and dialects” when asked about technologies for specific languages, and reported being hampered by poor professional incentives and “limited funding for low-resourced languages” (Rehm and Hegele, 2018). Nevertheless, Europe is relatively well served in this area, with [CLARIN](#) (“Common Language Resources and Technology Infrastructure”) offering a widely used platform integrating open digital linguistic resources in Europe and beyond for research support in the humanities and social sciences. CLARIN has a long trajectory in this area, giving access to a range of textual and audio-visual data containing written, spoken and signed language. CLARIN increasingly stands out as a model for access to high quality tools and services for language-based analysis, which, moreover considers language “not only as an object of inquiry, but also as a carrier of cultural content, as a means of communication, and as a component of identity” (Renckens et al., 2019: 11). Meanwhile, the [European Language Grid](#), building on historic language diversity projects such as [META.NET](#), aims to overcome the fragmented nature of the language technology landscape across both commercial and non-commercial stakeholders in creating an “architecture and components for a public, open and interoperable grid connecting resources and tools, sharing and combining resources to support effective development and deployment of [LT] (software and services) across Europe” (Rehm, 2019). Lastly, a new campaign by the [European Language Equality](#) project aims to develop “a roadmap for achieving full digital language equality in Europe by 2030”.

Small transformations in various fields of digital study also offer some room for optimism here. Returning to NLP, recent work by Joshi et al. analyses the relationship between language types, language resources and publication patterns in NLP conferences, as part of wider appeals within the Computational Linguistics (CL) community to pay attention to language diversity in both NLP systems in general (what the article calls a linguistic “typology echo-chamber” created by the supremacy of a few language families in language technology systems) and in the CL community’s

own practices (focusing on specific measures it can take to address the linguistic resource divide). The article points to the enormous influence of NLP conferences on language technology development and the strategic agendas of industry and government alike. The authors close with some recommendations for how to make NLP/CL conferences more language-inclusive in future (Joshi et al., 2021).

In a similar vein, Equality, Diversity and Inclusion (EDI) was a major feature of [EACL 2021](#), the conference of the European Chapter of the Association for Computational Linguistics, which built on EDI initiatives at other CL events in recent years, and which, at EACL 2021, had a particular focus on linguistic diversity. EACL organisers announced a range of [inclusion measures](#), which included ‘Birds of a Feather (BoF) Sessions’ for attendees to share research and generate new collaborative relationships, mentoring support sessions, affinity group social events organised around a range of topics, many of which were language community-centred ([African NLP](#), North Africans in NLP, Arabs in NLP and LatinX in AI) and a ‘Language diversity panel and games’. The panel and games sessions were designed to “celebrate language diversity” and to foster concrete actions the NLP research community can take to bridge linguistic divides in the languages it studies, the datasets it creates, its language(s) of communication and its evaluation mechanisms. The panel scrutinised the “great disparity” between the numbers of speakers of given languages and their relative mentions in publication platforms of the Association for Computational Linguistics (ACL); this linguistic imbalance is evident in the fact that some languages with millions of speakers are barely mentioned in the ACL anthology, whereas other, smaller languages have a relatively high mention rate. Attendees were polled on a series of questions to make the conference more linguistically inclusive, and panellists proposed measures such as improved credit mechanisms in the field around resource creation for low resourced languages, and naming the language you are working on (following the so-called ‘[Bender rule](#)’).

How are other fields of digital study addressing these disparities? One initiative in 2021-22, at the intersection between NLP and the digital humanities (DH), consists of a one year, three-workshop course bringing together researchers from various fields under the title ‘[New Languages for NLP: Building Linguistic Diversity in the Digital Humanities](#)’. Here, the course participants are working on ten current or historic language variants and over the course of the programme they will be taught to create their own datasets and statistical language models using NLP tools. In so doing, they will also contribute valuable data to open-source repositories, and lessons will be drawn on how to address situations where current culture-driven NLP tools do not work, in order to diversify the language resources available to people carrying out similar NLP research in future.

The course materials will later be published on an open educational platform called [DARIAH-CAMPUS](#).

The last decade has seen increasing attention in the digital humanities to the challenges of digital linguistic diversity and while early interventions tended to focus on theoretical issues or the field’s own communication practices –summarised by Élika Ortega in an MLA presentation (Ortega, 2015) – recent efforts have increasingly aimed to build a community of practice around the label ‘Multilingual DH’. This report earlier described how the 2019 workshops led by Cosima Wagner and Martin Lee looked at the specific infrastructural and methodological challenges facing researchers working with non-Latin scripts, while the ‘Multilingual DH’ group, a loose network of researchers interested in the challenges of carrying out digital humanities research with non-English languages, has started to build an open and shared repository of bibliographic material, tools, corpora and resources (with an accent on multilingual NLP) in order to facilitate multilingual DH research. Specific language communities (and/or research communities focused on particular languages/language families) have also started to consolidate: [Network for Digital Humanities in Africa](#) or East Asian Digital Humanities. And finally, the [Programming Historian](#) has expanded its language coverage to four languages for its widely used and community-led “novice-friendly, peer-reviewed tutorials that help humanists learn a wide range of digital tools, techniques, and workflows to facilitate research and teaching”.

In the introduction to this report, attention was drawn to concepts such as ‘language indifference’ or ‘language insensitivity’, which, in the context of digital multilingualism, serve to underscore the need to raise the visibility of languages-related work, to situate languages-based study in their broader social and cultural context and to increase the agency of the *languages* perspective in digital languages study. In a recent article, Renata Brandão and Paul Spence argued for a holistic approach to language diversity in the field of the digital humanities, which includes improving awareness of the dynamics of linguistic diversity in digital research, developing “languages-sensitive” approaches to digital study, greater attention to the creation of multilingual infrastructure and “playing a more significant role in the languages-related cultural questions of our era in future” (Spence and Brandão, forthcoming).

“The thing about linguistic discrimination is that the discrimination isn't really about language, but rather, people.”

[Vijay Ramjattan 2021](#)

Final thoughts

These initiatives are important because they address a number of issues raised at the DDM workshop, such as the importance of including cultural and social perspectives in discussions of digital multilingualism, the need for input from a wider range of skillsets and stakeholders (beyond Big Tech) and re-centring digital research & innovation on the needs of language communities. Although some speakers and participants at the DDM workshop were members of language communities, language communities are one of the key missing stakeholder groups in discussions about digital multilingualism and we need to look more at how we involve their voices and contributions more directly in similar initiatives in the future. Nowhere was this more evident than in the case study on how linguistic diversity has affected responses to COVID-19 presented by Mandana Seyfeddinipur discussed earlier.

In their presentation of ‘Learnings from Technological Interventions in a Low Resource Language: A Case-Study on Gondi’, Mehta et al. made the point that it is rare to see “a holistic view of technology being used to help revive usage of an endangered or vulnerable language” (Mehta et al., 2021). This is a wider problem within digital multilingualism as a whole, which is currently fragmented across disciplines and professional areas. Numerous respondents highlighted the importance of connecting digital multilingual initiatives better with the experience of language communities and designing tool or technology development based on how these communities actually use language in practice. There is also a need for improved cross-disciplinary connections on various levels (e.g. between those researching translanguaging practices and those creating the technologies). The challenge is not just “solving an engineering problem” (to borrow the words of Mehta et al. again), and the intention of this workshop report is to contribute to the wider theoretical and practical scaffolding which is currently being developed to support more rounded and linguistically-inclusive approaches to digital study and practice.

The Disrupting Digital Monolingualism workshop was designed to address ongoing (and largely unresolved) multilingual challenges facing digital research and practice, namely that digital methods and infrastructures too frequently gloss over (or actively ignore) linguistic and cultural diversity, and that they still tend to favour monolingual and highly-resourced languages (in particular English) to the detriment of others. It aimed to bring together a number of different dialogues, happening in a number of different fields, and which include debates about endangered languages, biocultural diversity, the role and future of language disciplines and professions, multilingual digital research infrastructure, localisation in digital media, decolonising the internet, multilingual/multicultural perspectives on the role of Artificial Intelligence and machine translation, and global digital humanities.

Our goal was not only to give testament to the sheer breadth of perspectives which need to come into play when exploring digital linguistic and cultural diversity, but also to forge new collaborations between key stakeholders who do not necessarily have enough opportunities

to engage with each other in exploring these questions. We also wanted to promote alliances between academic, commercial, third sector and language community respondents in addressing multilingual challenges drawing together infrastructure, digital methods and data, transcultural/translingual perspectives and AI/NLP. Discussion around digital multilingualism has historically been fragmented or marginalised, but the need to engage with new forms of multilingual and transcultural knowledge creation has never been clearer and there is evidence of a growing interest in addressing the challenge from multiple points of departure. We believe that the initiatives presented at the workshop can be an important source of inspiration for future debate and action over digital multilingualism.

This report demonstrates the extraordinary dynamism of the research and practice represented by those who contributed to the Disrupting Digital Monolingualism workshop, which was evident in the enthusiastic feedback we received about the event afterwards. We have attempted to capture both the key messages and the spirit of the workshop in this report, and our analysis has aimed to find connections and avenues for future collaboration rather than presenting fixed 'solutions'. Some of those involved in the theme groups may publish their own findings and proposals going forward, and we hope to collaborate with some of you on other activities 'disrupting digital monolingualism' in the future.

References

- Anon (2020) *Multilingual DH*. Available from: <http://multilingualdh.org/> (Accessed 15 October 2020).
- Arriaga, E. & Villar, A. (2021) *Afro-Latinx Digital Connections*. Gainesville: University Press of Florida.
- Bender, E. M., Gebru, T., McMillan-Major & A., Shmitchell, S. (2021) 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. 3 March 2021 New York, NY, USA: Association for Computing Machinery. pp. 610–623. Available from: <https://doi.org/10.1145/3442188.3445922> (Accessed 2 August 2021).
- Burdett, C., Burns, J., Duncan, D. & Polezzi, L. (2018) *Transnationalizing Modern Languages: Reframing language education*. Available from: <http://www.bristol.ac.uk/media-library/sites/policybristol/PolicyBristol-report-35-Sept18-transnationalizing-modern-languages.pdf> (Accessed 15 October 2020).
- Cebero Berger, K., Gurrutxaga Hernaiz, A., Baroni, P., Hicks, D., Kruse, D., Quochi, V., Russo, I., Salonen, T., Sarhimaa, A. & Soria, C. (2018) *Digital Language Survival Kit: The DLDP Recommendations to Improve Digital Vitality*. Available from: http://www.dldp.eu/sites/default/files/documents/DLDP_Digital-Language-Survival-Kit.pdf.
- Danet, B. & Herring, S. C. (eds.) (2007) *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford ; New York: Oxford University Press, U.S.A.
- The Digital Language Diversity Project* (2019) *Project*. Available from: <http://www.dldp.eu/en/content/project> (Accessed 15 October 2020).
- Escobar Varela, M. (2020) 'Emic interfaces: UX design for cultural specificity'. *2020 Global DH Symposium*. Available from: <https://www.youtube.com/watch?v=tCOVPtKwQG8> (Accessed 15 October 2020).
- Forsdick, C. (2017) 'Translating Cultures', *translating research*. Available from: <http://www.meits.org/dialogues/article/translating-cultures-translating-research> (Accessed 15 October 2020).
- Held, M. (2010) "'The People Who Almost Forgot": Judeo-Spanish Web-Based Interactions as a Digital Home-Land'. *El Prezente: Studies in Sephardic Culture*. 83–102.
- Humanistica (2014) *Humanistica: Présentation*. Available from: <http://www.humanisti.ca/presentation/> (Accessed 15 October 2020).
- Joshi, P., Santy, S., Budhiraja, A., Bali, K. & Choudhury, C. (2021) 'The State and Fate of Linguistic Diversity and Inclusion in the NLP World'. *arXiv:2004.09095 [cs]*. Available from: <http://arxiv.org/abs/2004.09095> (Accessed 2 August 2021).
- Joshi, P., Barnes, C., Santy, S., Khanuja, S., Shah, S., Srinivasan, A., Bhattamishra, S., Sitaram, S., Choudhury, M. & Bali, K. (2019) 'Unsung Challenges of Building and Deploying Language Technologies for Low Resource Language Communities'. *arXiv:1912.03457 [cs]*. Available from: <http://arxiv.org/abs/1912.03457> (Accessed 14 September 2020).
- Lee, C. (2016) *Multilingualism Online*. 1 edition. London ; New York: Routledge.
- Lee, M. & Wagner, C. (2019) *Towards Multilingualism In Digital Humanities: Achievements, Failures And Good Practices In DH Projects With Non-latin Scripts (Workshop)*. Available from: <https://multilingualdh.org/en/dh2019/>.
- van Lit, C. et al. (2020) *DDM Workshop: Theme Group 1 Biographies* [online]. Available from: <https://languageacts.org/digital-mediations/event/disrupting-digital-monolingualism/Themes/themeone/biographiesOne/> (Accessed 3 August 2021).
- van Lit, C. et al. (2020) *DDM Workshop: Theme Group 1 Summary* [online]. Available from: <https://languageacts.org/digital-mediations/event/disrupting-digital-monolingualism/Themes/themeone/summary/> (Accessed 3 August 2021).

- Maaya Network (2012) *NET.LANG: Towards the multilingual cyberspace*. Caen, France: C & F Editions. [online]. Available from: http://net-lang.net/externDisplayer/displayExtern/_path_/netlang_EN_pdfedition.pdf (Accessed 15 October 2020).
- Maffi, L. (2005) Linguistic, Cultural, and Biological Diversity. *Annual Review of Anthropology*. [Online] 34 (1), 599–617.
- Mehta, D., Santy, S., Kommiya Mothilal, R., Lal Srivastava, B.M., Sharma, A., Shukla, A., Prasad, V., Venkanna U, Sharma, A. & Bali, K. (2021) Learnings from Technological Interventions in a Low Resource Language: A Case-Study on Gondi. arXiv:2004.10270 [cs]. [online]. Available from: <http://arxiv.org/abs/2004.10270> (Accessed 30 July 2021).
- META-NET (2012) *META-NET White Paper Series*. [online]. Available from: <http://www.meta-net.eu/whitepapers/overview> (Accessed 15 October 2020).
- Nekoto, W. et al. (2020) Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. arXiv:2010.02353 [cs]. [online]. Available from: <http://arxiv.org/abs/2010.02353> (Accessed 3 August 2021).
- Network for Digital Humanities in Africa (2020) Resources & past events. Network for Digital Humanities in Africa [online]. Available from: <https://dhafrica.blog/resources/> (Accessed 15 October 2020).
- Noble, S. U. (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. Illustrated edition. New York: NYU Press.
- Ortega, É. (2015) *Multilingualism in DH*. [online]. Available from: <http://web.archive.org/web/20210424073656/https://www.disruptingdh.com/multilingualism-in-dh/>.
- Patti, E. (2018) Digital culture studies: National and transnational perspectives in Modern Languages. *Explorations in Media Ecology*. 17 (3), 259–262.
- Pitman, T. & Taylor, C. (2017) *Where's the ML in DH? And Where's the DH in ML? The Relationship between Modern Languages and Digital Humanities, and an Argument for a Critical DHML*. 11 (1), . [online]. Available from: <http://www.digitalhumanities.org/dhq/vol/11/1/000287/000287.html> (Accessed 20 April 2017).
- Rehm, G. (2019) *European Language Grid: An Overview*. [online]. Available from: <https://www.european-language-grid.eu/wp-content/uploads/2019/10/00-03-ELG-Overview-Georg-Rehm.pdf>.
- Rehm, G. & Hegele, S. (2018) 'Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs', in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. May 2018 Miyazaki, Japan: European Language Resources Association (ELRA). p. [online]. Available from: <https://aclanthology.org/L18-1519> (Accessed 2 August 2021).
- Renckens, E. et al. (eds.) (2019) *CLARIAH: a digital research infrastructure for humanities researchers*. Amsterdam: Spinhuis-Huygens ING-CLARIAH-Bureau.
- Risam, R. (2018) *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Evanston, Illinois: Northwestern University Press.
- Sapir, E. (1924) The Grammarian and His Language H. L. Mencken & George Jean Nathan (eds.). *The American mercury*. 1149–155.
- Sayers, D. et al. (2021) *The Dawn of the Human-Machine Era: A forecast of new and emerging language technologies*. [Online] [online]. Available from: <https://jyx.jyu.fi/handle/123456789/75737> (Accessed 2 August 2021).
- Schneider, F. (2015) Searching for 'Digital Asia' in its Networks: Where the Spatial Turn Meets the Digital Turn. *Asiascape: Digital Asia*. [Online] 2 (1–2), 57–92.
- Soria, C., Russo, I., Quochi, V., Hicks, D., Gurrutxaga, A., Sarhimaa, A. & Tuomisto, M. (2016) 'Fostering digital representation of EU regional and minority languages: the Digital Language Diversity Project', in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. May

2016 Portorož, Slovenia: European Language Resources Association (ELRA). pp. 3256–3260. Available from: <https://www.aclweb.org/anthology/L16-1518> (Accessed 15 October 2020).pp. 3256–3260. [online]. Available from: <https://www.aclweb.org/anthology/L16-1518> (Accessed 15 October 2020).

Spence, P. & Brandão, R. (forthcoming) Towards language sensitivity and diversity in the digital humanities. *Digital Studies / Le champ numérique*.

de Swaan, A. (2002) *Words of the World: The Global Language System*. 1 edition. Cambridge, UK ; Malden, MA: Polity.

Taylor, C. et al. (2017) Modern Languages and the Digital: The Shape of the Discipline Claire Taylor & Niamh Thornton (eds.). *Modern Languages Open*. [Online] [online]. Available from: <http://www.modernlanguagesopen.org/articles/10.3828/mlo.v0i0.156/> (Accessed 9 June 2017).

Vierthaler, P. (2020) Digital humanities and East Asian studies in 2020. *History Compass*. [Online] 18 (11), e12628.

Vrana, A., Sengupta, A., Pozo, C., & Bouterse, S. (2020) *Decolonizing the Internet's Languages – Summary Report*. Available from: <https://whoseknowledge.org/resource/dtil-report/> (Accessed 29 July 2021).

Appendices

Appendix 1: About the organisation of the event

The local team

This workshop was led by the Language Acts and Worldmaking project with the support of the Cross-Language Dynamics: Reshaping Community project, both funded by the AHRC under its Open World Research Initiative. It was co-convened by Paul Spence, Renata Brandão and Naomi Wells with the support of our local team (Felicity Roberts, Nayana Dhavan, Gabriele Salciute Civiliene, Kristen Schuster and Beth Martin).

The two projects hosting this event - Language Acts and Worldmaking, which hosted the event and Cross-Language Dynamics, which provided generous support – were funded by the UK's Arts and Humanities Research Council under its Open World Research Initiative to demonstrate the importance of languages in understanding today's transnational, translingual and transcultural interactions. In the Digital Mediations strand of the Language Acts and Worldmaking project, Paul Spence and Renata Brandao have attempted to map interactions between modern languages and digital culture and analyse the challenges which arise. In a series of collaborations under the [Digital Modern Languages](#) label, Paul Spence and Naomi Wells have brought together research and teaching in Modern Languages which engages with digital culture, media and technologies.

Thanks

We would like to thank the many people who provided advice or inspiration for this event; they include Cosima Wagner, Quinn Dombrowski, David Wrisley, Gimena del Río, Ernesto Priani, Alex Gil, Élika Ortega, Claire Taylor, Padmini Ray Murray, Mandana Seyfeddinipur, Jamies Smithies, Arianna Ciula and Shawn Kelly. Thanks to our respective projects, Language Acts and Worldmaking and Cross-Language Dynamics: Reshaping Community project, and to the UK Arts and Humanities Research Council, whose Open World Research Initiative funding made this all possible. We have immense gratitude for the work of our theme groups facilitators, who responded with agility and creativity to our proposal - L. W. Cornelis van Lit, Cosima Wagner, David Joseph Wrisley, Miguel Escobar, Darja Fišer, Sender Dovchin, Emanuela Patt, Kalika Bali and Quinn Dombrowski. We would also like to thank Kristen Mapes and the Global Digital Humanities team at MSU, who generously provided technical advice on on-line event management. Finally, we would like to thank all contributors to the workshop itself, who made the event what it was.

Despite the 'lost year' due to the pandemic it seemed important to leave a testament to the contributions at this workshop, but these are my own observations about the event, and any imprecisions here are all my responsibility. I have avoided an overly synthetic approach, which risked losing individual voices, in order to stay faithful to the structure and individual presentations as far as possible.

Finally, thanks go to the six people who reviewed the text of this report and made vital contributions for which I am very grateful. They included: Renata Brandão, Peter Chonka, Nayana Dhavan, Quinn Dombrowski, Kristen Schuster and Naomi Wells.

Appendix 2: Aims and themes

Aims

- To map the current state of multilingualism in digital theory and practice through, and across, languages and cultures
- To identify areas of linguistic (especially anglophone) bias and 'language indifference' in digital methodologies and infrastructure
- To discuss the value and role of languages in digital theory and practice and their implications for language study and professions

- To bring together experts in languages-driven digital study and practice to discuss priorities for future action and potential collaboration
- To explore emerging models for linguistic diversity and languages-aware digital practice in academia, education and private/third sectors and to document best practice

General theme: Multilingualism in digital theory and practice

Challenges

- How well do current digital media ecologies respond to multilingualism? How do digital research practices reflect/enact linguistic and geocultural diversity?
- What constitutes 'language indifference' in digital practice, and how can it be overcome?
- Which critical perspectives, models and best practices might be most effective in enabling multilingual responses?

Theme 1: (Linguistic and geocultural diversity in digital knowledge infrastructures)

Challenges

- In what ways do digital knowledge infrastructures embed multilingual or language-focused practice at present, in social and technical terms?
- How do digital infrastructures reinforce or redistribute linguistic and cultural influence or authority?
- What multilingual and localisation responses from industry, academia or the third sector have been most effective so far?

Theme 2 (Working with multilingual data)

Challenges

- What possibilities exist for 'data-driven' methods in languages-based research? What triggers and blocks influence their use?
- How well do digital data practices or research reflect linguistic and cultural difference? How does the growing datafication of culture and society affect the balance of world languages, and what steps can be taken to foster greater diversity and language awareness?
- What new forms of digital, linguistic and cultural criticism are needed to interpret data-driven approaches to languages and multilingualism?

Theme 3 (Transcultural and translingual approaches to digital study)

Challenges

- What new possibilities do digital media offer for studying translingual and transcultural dynamics? For example, how can we effectively study intercultural, plurilingual or transborder perspectives using digital tools?
- What is the role of digital practitioners (in research or industry) and language experts (such as translators, linguists and modern languages researchers) in studying these interactions?

Theme 4 (Artificial intelligence, machine learning and NLP in language worlds)

Challenges

- What challenges do AI, machine learning and Natural Language Processing pose for language research fields and professions?
- What implications are there for translation and for studying or researching modern languages and their cultures?
- What opportunities does AI provide to foster linguistic diversity or intercultural communication?

Responses

How can we best imagine and create language-sensitive digital tools, methods and infrastructure in the future? What are the implications for digital policy and strategy, and what critical frameworks and toolkits are required? What existing models and research can we draw on?

We propose two kinds of response at the workshop to the challenges listed above. One will focus on the digital policies and conceptual frameworks required to foster multilingual practices. The other will focus on a practical outcome of digital language frameworks and toolkits.

We propose possible outcomes from these two strands, but ultimately the outcomes will be defined by participants on the second day.

Response strand 1 (Digital policies and multilingualism)

Possible outcome

- A collaborative document outlining a conceptual framework for multilingualism and geocultural diversity in digital research created during (and possibly after) the event

Response strand 2 (Digital language framework and toolkits)

Possible outcome

- Publication of frameworks and toolkits demonstrated at the workshop
- New collaborations, prototypes or toolkits

Appendix 3: List of presentations/contributions

Lightning talks

- Bunty Avieson - University of Sydney. *Wikipedia as pharmakon: poison and cure for minority languages*
- Andiswa Bukula - South African Centre for Digital Language Resources. *Multilingualism in the South African context*
- Matteo Dutto - Monash University, School of Languages, Literatures, Cultures and Linguistics. *#YouthintheCity: Re-mapping Transcultural Spaces through the Voices of Multilingual Migrant Youth*
- Sarah McMonagle - University of Hamburg. *Which (mis)perceptions matter in minority language media research? Reflections on/from an enquiry of digital language practices among Sorbian adolescents*
- Pascal Belouin and Sean Wang - Max Planck Institute for the History of Science. *RISE and SHINE: An API-based Infrastructure for Multilingual Textual Resources*
- Michael Castelle - University of Warwick. *Multilingual Transformers: Linguistic Relativity for the 21st Century*
- Leonore Lukschy - SOAS University of London. *Making endangered languages archives linguistically accessible*
- Ernesto Priani Saisó - Universidad Nacional Autónoma de México. *Challenges of not using English as the dominant language in DH international projects*
- Carlos Yebra Lopez - New York University. *How to Use Digital-Homelands in order to Revitalise Diasporic Languages*
- Isabelle Zaugg - Columbia University's Data Science Institute. *Let's Talk About Scripts*
- Peter Chonka – King's College London. *Search as research in African indigenous languages: potentials and problematics of enquiry into auto-complete predictions/suggestions for Af Soomaaliga*
- Anna Jørgensen - University of Amsterdam. *Newswork on Wikipedia – the Case of the Coronavirus*
- Pedro Nilsson-Fernández - University College Cork, Ireland. *Digital Peripheries: A Postcolonial Digital Humanities Approach to Catalan 20th Century Literary Spaces*
- Elizabeth Marie Thaut - SOAS, University of London. *Language documentation and description with(in) a digital diaspora: The Sylheti Project - SOAS in Camden*
- Cosima Wagner - Freie Universität Berlin. *Challenging research infrastructures from a multilingual DH point of view - a short overview*

Panel

- Anasuya Sengupta - Whose Knowledge? campaign
- Eduard Arriaga - University of Indianapolis
- Cosima Wagner - Freie Universität Berlin
- Kalika Bali - Microsoft Research India

Demos

- Andrew Janco - Haverford College. *Cadet: A Tool to Add New Language Models to spaCy*
- Ethem Mandić - Faculty of Montenegrin Language and Literature (FCJK). *Digitization of Montenegrin language skills in the global multicultural society*
- Caoimhín Ó Dónaill - Ulster University. *CLILSTORE: An open online platform for multimedia language learning*

Posters

- Jessica Green - British Library. *Around the British Library in 40 Languages: Engaging with a Different Community Each Week #AToUnknown*

- Stuart Prior - Wikimedia UK. *Wikidata and Languages: Building a multilingual internet*

Invited talk

- Mandana Seyfeddinipur - SOAS. *If it's not written it did not happen: The written bias in the digital world*

Mini-workshop

- Mini-workshop on *Multilingual data in ELTeC: enacting European literary traditions* led by Ioana Alexandra Lionte and Roxana Patraş - "Alexandru Ioan Cuza" University of Iaşi.

Appendix 4: Speakers and contributors

Speakers

There were 25 speakers with institutional affiliation in 12 different countries over two days for the synchronous event (Part One), and the theme groups included 41 participants based in 16 countries for Part Two.

Audience

The first part of the workshop received over 300 registrations from all around the world and attendance was extended with up to 230 people watching the presentations on an additional YouTube stream. The workshop brought together people from academia, language professions, secondary education, local and global digital media companies, the cultural heritage sector, the Galleries Libraries Archives Museums (GLAM) sector, funding agencies, international policy organisations and the creative arts sector. The programme was internationally diverse both in geographic representation and languages covered, and while the programme would have been constructed differently if this had been planned as a virtual event from the outset, general feedback was extremely positive.

Registrations came in from the following countries, spanning six continents:

Africa: Ethiopia; Malawi; Morocco; Nigeria; Senegal; South Africa; Zambia.

Asia: Bangladesh; China (mainland and Hong Kong SAR); India; Indonesia; Japan; Kazakhstan; Malaysia; Nepal; Philippines; Saudi Arabia; Taiwan.

Europe: Austria; Belgium; Czech Republic; Denmark; Estonia; France; Germany; Great Britain; Greece; Hungary; Ireland; Italy; Montenegro; Netherlands; Poland; Portugal; Romania; Russian Federation; Spain; Turkey.

Australasia: Australia; New Zealand.

North America: Canada; Mexico; United States of America.

South America and the Caribbean: Argentina; Brazil; Chile, Colombia; Paraguay; Trinidad and Tobago.

It is important to note that participants at the workshop did not just consist of academics but included many professionals with a strong focus on language(s) working in digital media companies, in other commercial settings or in the third sector.

People registering for the workshop listed a wide variety of roles, including:

- Academic/education roles: student, PhD student, professor, lecturer, researcher or teaching fellow.
- Academic support roles: Academic Librarian; Digital Scholarship Librarian.
- Digital roles: Data Scientist; Enterprise Architect; Head of Digital Research; Software Engineer; Systems Designer.
- Academic or industry leadership roles: Business and STEM Dean; Coordinator of Academic Programs at a funding agency; Director of Modern Languages Centre; Head of Digital Research; Head of languages and Cultures; Programme manager at internet study centre; Director of Digital Liberal Arts Research; Vice dean.
- Other roles: Artist; Associate Editor; Curator; International Development; Theatre Director; Translator.

Academic fields listed included: digital pedagogy; language education; modern languages; area studies; linguistics; digital humanities; art; media and communication; world and comparative literature; diasporic studies; migration studies; postcolonial studies; library and information science; digital curation; data science; international development; translation studies; sociolinguistics; science and technology studies; theatre studies. This list largely obscures (professional or academic) digital fields of those attending, which were not generally requested or specified in any depth.

Registrants were asked why they joined the workshop, and a wide variety of responses was provided. These responses/framings have been grouped and summarised below:

- **General understanding of linguistic landscape online**
 - To improve understanding of the linguistic landscape online, and factors affecting facing linguistic

and epistemological diversity.

- To explore the representation of non-European languages in digital spaces and openness to other scripts, languages, and cultures.
- To discuss the topic with stakeholders from a variety of sectors.
- Several respondents noted the importance of having a focus on language, which often seems missing in the rush to digitize – and sought the opportunity to explore emerging models for languages-aware practice in academic, education and private/third sectors.
- **Greater multilingual perspectives on digital study and practice**
 - To learn more about the role of languages in digital theory and practice.
 - To overcome monolingualism in the digital representation of cultures and their heritage.
 - To study the digital humanities in the non-Western world.
 - One respondent explained that they work with mixed (philological and digital) methods in the context of East Asia, which is still an underrepresented area in digital studies, and they sought to learn more about best practices and the state of the field from like-minded scholars.
 - Another participant was researching mediated communication via chat apps in India, a highly multilingual country and wished to explore data collection challenges.
- **Multilingual approaches to Artificial Intelligence (AI), Natural Language Processing (NLP) and text mining**
 - To learn more about how groups are working towards NLP models for language diversity and the challenges of multilingualism in text mining.
 - To study multilingual approaches to NLP and discourse analysis or to predictive AI for social semiotics and Online Talk.
 - One participant wished to understand the challenges and research directions for NLP for long-tail languages.
- **Language technologies**
 - To confront monolingualism in the voice industry.
 - To incorporate digital tools in language research.
 - To research multilingualism in voice tech and social media.
 - To study how digital technologies can be used to teach, disseminate, and preserve languages.
 - To improve support for right to left languages.
- **Low-resourced languages**
 - To develop strategies and tools/infrastructure for enabling endangered, minoritized, heritage or indigenous languages in digital spaces.
- **Translation and intercultural exchange**
 - Digital translation tools and methods.
 - Transcultural and translingual approaches to digital study.
 - How to deal with issues of translation and transcultural exchange in online academic publishing.
- **Intersections and borders**
 - Intersections of linguistics, multilingualism, and digital humanities/digital technology.
 - Gaining insight about how to navigate an increasingly “disconnected/fractured” internet.
 - Cross language dynamics and transcultural / translingual approaches to digital study.
 - How to approach multiple languages in regions divided by a geopolitical border and applying digital humanities practices to expose these dynamics.

digital humanities practices to expose these dynamics.

- Digital practice in migrant communities.

- **Language education, Modern Languages and Area Studies**

- The reinvention of Modern Languages as a discipline.
- New tools and approaches to language learning.
- Technology mediated second language learning.

- **Language activism and justice**

- Deconstructing the dominance of one language (English) online.
- Development of multilingual technologies and language justice in the digital sphere.
- Community building and planning further actions to promoting the disruption of digital monolingualism.
- Translation and adaptation of monolingual digital spaces.
- One respondent noted that as someone living “in a global south country where digital censorship is often being imposed by the government, internet makes it more difficult for sexual and religious minorities to find diverse information in their native language due to the majority group dominating the digital space and the dominance of English as the main language on the internet”. They saw the workshop as an opportunity to raise the voices of marginalized minorities and to diversify the language dynamics of the internet.

- **Practical experience.**

A small but important minority saw the workshop as an opportunity to improve their practical knowledge:

- To learn how to use multilingual data.
- To gain practical experience with multilingual methods and tools.
- To learn more about working with mixed-language data.

Theme group analysis

The second (asynchronous) part of the workshop was organised at very short notice due to the constraints imposed by the global pandemic, but facilitators for the four groups (two or three per group) nonetheless managed to create internationally diverse ‘theme groups’ under significant time pressures brought about by the evolving situation. The facilitators for each group had to balance diversity across numerous factors, but participants (37 in total) in the theme groups represented a wide variety of fields of digital and/or language-based study and practice across academia, the private sector and third sector, and had institutional affiliation in 17 countries: Australia; Bangladesh; Czech Republic; Denmark; France; Germany; Hungary; India; Japan; Mexico; Netherlands; Singapore; Slovenia; South Africa; United Arab Emirates; United Kingdom; United States.

Appendix 5: Key links

Disrupting Digital Monolingualism website <https://languageacts.org/digital-mediations/event/disrupting-digital-monolingualism/>

There is an archive of this website in the Internet Archive at <https://web.archive.org/web/20210417121615/https://languageacts.org/digital-mediations/event/disrupting-digital-monolingualism/>

Disrupting Digital Monolingualism YouTube channel https://www.youtube.com/watch?v=G1oGc8tYnCM&list=PLfaB2f0CyBduOaUww1H2HQMewtq_HiviP

Digital Mediations strand of Language Acts & Worldmaking project <https://languageacts.org/digital-mediations/>

[DOI: 10.5281/zenodo.5743283](https://doi.org/10.5281/zenodo.5743283)

This report forms part of a series of reports produced by the Digital Mediations strand of the Language Acts & Worldmaking project, in this case in collaboration with the translingual strand of the Cross-Language Dynamics project (based at the Institute of Modern Languages Research), both funded by the UK Arts and Humanities Research Council's Open World Research Initiative. Digital Mediations explores interactions and tensions between digital culture, multilingualism and language fields including the Modern Languages.

Contact: paul.spence@kcl.ac.uk | <https://languageacts.org/digital-mediations/event/disrupting-digital-monolingualism/>