Australian
National
University

Digital Approaches to Multilingual Text Analysis

January 27th 1:45pm – 7:30pm AEDT

Doe quam de Heere neder om de Stadt en Toren te beſien die de Menſchen-kinderen Bouwden, en ſeide; laat ons hare Sprake Verwerren op dat elk synen naaſten niet en hoore Alſoo wierden sy verſtrooit en hielden op de Stadt te bouwen. etc. Gen: 11. 5.7.8.

Image: Babylonische spraakverwarring, Rijksmuseum, http://hdl.handle.net/10934/RM0001.COLLECT.336773

Conveners

Joshua Brown Senior Lecturer and Convenor, Italian Studies, Australian National University

Katrina Grant Senior Lecturer, Centre for Digital Humanities Research, Australian National University

For more information contact: joshua.brown@anu.edu.au or katrina.grant@anu.edu.au

This online symposium brings together professionals and academics to discuss the complex dynamics of applying digital approaches in multilingual text analysis. Up until now, use of DH tools and methods have been applied across a variety of corpora but text-analysis of English language sources has dominated this field. These approaches are increasingly being used in languages and linguistics research for non-English corpora. At the same time, the integration of these tools has seen new research questions and possibilities emerge, including questions such as "Is there a non-Anglo digital humanities (DH), and if so, what are its characteristics" (Fiormonte 2016: 438). Recent studies have begun to examine aspects such as OCR for historical text analysis and data mining (Hill & Hengchen 2019; Goodman et al. 2018), multilingual computation analysis (Dombrowski 2020), semantic and sentiment analysis (Daems et al. 2019) and historical linguistics (Evans 2016), among others. We are particularly interested in the ways researchers have used digital tools to examine "big data" from the textual past, and how new tools are being developed to process multilingual texts.

Australian
National
University

Full program and abstracts

Session 1 | 1:45pm AEDT – 3:15pm AEDT

Keynote - Quinn Dombrowski (Stanford)
Non-English DH Is Not a Thing

Charbel El-Khaissi (Australian National University)
Syriac in the Digital Humanities: Successes and Challenges

Simon Musgrave (University of Queensland) and Peter Sefton (University of Queensland)
Infrastructure for Multilingual Text Analysis

Session 2 | 3:30pm-4:30pm

Samantha Disbray, Ben Foley (The University of Queensland), Shruti Rijhwani (Carnegie Mellon University), Meladel Mistica (The University of Melbourne)
Reading it Right: A Case Study in Pintupi-Luritja

Eunjeong Park (Sunchon National University)
The Affordances and Challenges of Using Learner Corpora to Multilingual Learners' Writing Instruction

Hua Tan (Central China Normal University) *Authorship attribution of Chinese Martial Art Fictions by Gu Long with Stylometry*

Session 3 | 5:30pm-6:30pm

Jonas Müller-Laackman (Freie Universität Berlin)
Arabic vernacular poetry - challenges in working with Arabic script, speech and (re)presentation

Diana Fabiola Zavala Rojas (University Pompeu Fabra), Danielly Sorato (University Pompeu Fabra), Lidun Hareide (Møreforsking AS), Knut Hofland (University of Bergen)
Multilingual Corpus of Survey Questionnaires

Yu Yan (Hubei University of Technology)
A Corpus-based Study of China's National Image in the English Translations of Government Report

Session 4 | 6:45pm-7:30pm

Joshua Brown (Australian National University)
Digital approaches to multilingual text analysis: the Dictionnaire as a code-intermediate space

Katrina Grant (Australian National University)
*Final remarks and lead into discuss*ion

Questions and final discussion

Abstracts

Quinn Dombrowski (Stanford) | *Non-English DH Is Not a Thing*

Within the broad international digital humanities community, two partially-overlapping foci of activity in the last decade have drawn attention to language. One has called into question the Anglophone hegemony in scholarly communications (e.g. conferences, publications, and less-formal spaces such as mailing lists; see Fiormonte, Gil, Grandjean, and Ortega); the second has emphasized the need for corpora, algorithms, and pedagogical approaches to support computational research in languages other than English (see Skorinkin, Dombrowski & Burns). While there is wide variation in the terminology to refer to these kinds of efforts, both "multilingual DH" and "non-English/Anglo-American DH" have seen use.
This talk argues that "non-English" is an unhelpful framing for these issues. Monolingual Anglophone scholars misinterpret the phrase, reifying it rather than treating it as a placeholder that represents an extremely diverse set of languages, literatures, cultures, and practices. "Non-English" DH is not a thing one can "support" with pedagogy, corpora, or tools -- one can only adopt a linguistically-inclusive perspective and choose a very small subset of languages beyond English to actively engage with, perhaps in different ways for each language. "Non-English" suggests a problem with the possibility of a singular solution, when in reality, response depends both on the language and the context. Drawing on experiences teaching a course framed as "non-English DH", and collaborating on a wide variety of projects in a modern languages department, this talk will offer a set of practical strategies for how to draw feasible boundaries around the set of languages one can meaningfully support.

References
Dombrowski, Quinn and Patrick Burns. "Language is Not a Default Setting". Forthcoming in Debates in the Digital Humanities 2021, ed. Lauren F. Klein and Matt Gold.
Fiormonte, Domenico. "Towards a Cultural Critique of Digital Humanities". Debates in the Digital Humanities 2016, ed. Lauren F. Klein and Matt Gold.
Gil, Alex and Élika Ortega. "Global Outlooks in Digital Humanities". In Crompton, C., Lane, R.J., and Siemens, R. eds. Doing Digital Humanities: Practice, Training, Research. London: Routledge.
Grandjean, Martin. "Le Rêve Du Multilinguisme Dans La Science : L'exemple (Malheureux) Du Colloque #DH2014.". 2014. http://www.martingrandjean.ch/multilinguisme-dans-la-science-dh2014/
Ortega, Élika. "Whispering/Translating During DH 2014: Five Things We Learned". July 21, 2014. https://elikaortegadotnet.wordpress.com/2014/07/21/dhwhisperer/
Skorinkin, Daniil. "Digital Humanities in Russia: A View from the Inside". El'Manuscript keynote, 2021. https://danilsko.github.io/slides/elmanuscript21/elmanuscript_keynote#/

Charbel El-Khaissi (Australian National University) | *Syriac in the Digital Humanities: Successes and Challenges*

This presentation discusses digital approaches in Aramaic, with a focus on the Late Aramaic dialect of Syriac (2ndc.—13th c. CE). The discussion begins by tracing the origins of Syriac in the Digital Humanities (SGH) in the mid-20th-century via the pioneering computational efforts of George Kiraz (1965—) and Sebastian Brock (1938—), which situate Syriac as an 'early adopter' of DH. Then, an overview of key characteristics in SGH are demonstrated, including digital archives, textual corpora and a 'cyberinfrastructure for classical philology' (Crane, Seales and Terras, 2009). The presentation concludes by summarising two ongoing issues in SGH relating to scripts and fonts, and automated part-of-speech (POS) tagging. I connect these two DH challenges to gaps in broader linguistic research—specifically, the connection between fonts (or lack thereof) and 'mixed languages' (e.g. Arameo-Arabic), and persistent blind spots of Natural Language Processing when dealing with Aramaic (and Semitic) morphology. When these insights are considered in light of Fiormonte's (2016: 438) question of *is there a non-Anglo DH?*, Aramaic demonstrates that in spite of some challenges, non-Anglo DH is an otherwise thriving cross-sectional discipline.

References

Crane, G., Seales, B. W., & Terras, M. (2009). Cyberinfrastructure for Classical Philology. *Digital Humanities Quarterly, 3*(1). Retrieved from http://www.digitalhumanities.org/dhq/vol/3/1/000023/000023.html

Fiormonte, D. (2016). Toward a Cultural Critique of Digital Humanities. In Matthew K Gold & Lauren F Klein (eds.). *Debates in Digital Humanities* (pp. 438-458). Minneapolis: University of Minnesota Press.

Sources

Benardou, A., Champion, E., Dallas, C., & Hughes, L. M. (2017). Introduction: a critique of digital practices and research infrastructures. In A. Benardou, E. Champion, C. Dallas, & L. M. Hughes (Eds.), *Cultural Heritage Infrastructures in Digital Humanities* (1 ed.). London: Routledge.

Ishac, E. A. (2020). From Ancient Manuscripts to Digital Screens: Syriac Liturgy in Digital Humanities. In M. Tomić, M. Willer, & N. Tomašević (Eds.), *Empowering the Visibility of Croatian Cultural Heritage through the Digital Humanities* (pp. 148–159). United KIngdom: Cambridge Scholars Publishing.

Michelson, D. A. (2016). Syriaca.org as a Test Case for Digitally Re-Sorting the Ancient World. In C. Clivaz, P. Dilley, & D. Hamidović (Eds.), *Ancient Worlds in Digital Culture* (pp. 59-85). Online: Brill.

Walters, J. E. (2020). The Digital Syriac Corpus: A Digital Repository for Syriac Texts. *Zeitschrift für Antikes Christentum / Journal of Ancient Christianity, 24*(1), 109-122. doi:10.1515/zac-2020-0018

Simon Musgrave (University of Queensland) and Peter Sefton (University of Queensland) | *Infrastructure for Multilingual Text Analysis*

Small scale projects involving digital analysis of texts in languages other than English can manage their data locally. For research at the scale implied by the term 'big data' using large datasets which can be reusable resources, large-scale infrastructure is needed to support and enable the research and to work towards data management in line with FAIR principles and the Australian Code for the Responsible Conduct of Research (ARC et al. 2018).

The Language Data Commons of Australia (LDaCA) is an initiative to provide infrastructure for language-based research in Australia aiming to provide access to materials which record language use in Australia. LDaCA is not a repository in its own right (although finding vulnerable content and ensuring its preservation is within the scope of the project), but it will federate access to existing storage facilities to improve accessibility and reusability. Given the diverse linguistic environment of Australia, the architecture and interfaces have to be capable of handling data in various languages and writing systems.

In this paper we will introduce the overall architecture of LDaCA and demonstrate its capability in handling multilingual data by exploring a small parallel corpus. This collection contains the English original of information sheets prepared by a federal government agency along with their translations in Arabic, Chinese, Turkish and Vietnamese. Our demonstration will show that the LDaCA interface can discover documents on the basis of metadata (e.g. a title) in a language other than English, and that text searches inside a non-English document can also be carried out.

Reference

Australian Research Council, National Health and Medical Research Council and Council of Australian Universities 2018. *Australian Code for the Responsible Conduct of Research*. https://www.nhmrc.gov.au/about-us/publications/australian-code-responsible-conduct-research-2018
FAIR https://www.go-fair.org/fair-principles/

Samantha Disbray, Ben Foley (The University of Queensland), Shruti Rijhwani (Carnegie Mellon University), Meladel Mistica (The University of Melbourne) | *Reading it Right: A Case Study in Pintupi-Luritja*

Languages with small numbers of users can benefit from digital tools. In a recent project, a set of community newsletters in Pintupi-Luritja language were digitised and made available on Trove, a database of Australian library content. Pintupi-Luritja is used by approximately 1000 people, and has a rich literary tradition due to its history of bilingual education and use in church. While the digitisation makes the original content available and shareable, the quality of the OCR output is unreliable. This limits information search and retrieval, and provides a poor model for standardising spelling of the language.

To remedy this, our project proposes a three stage approach:

1. Investigate methods to improve the OCR output quality.
2. Evaluate and correct errors post-OCR using a language-model based spelling correction tool.
3. Work with Trove through its ‘voluntrove’ program to revise the plain texts.

The questions we ask are why and for whom we are creating these language resources.

1. How important is this kind of tool to the language community?
2. How valuable is this tool for the research community?
3. How will/should we engage the language community?

To the community, it is not obvious that a spell checker is needed, and there has been no request from the community for such a tool. The project will investigate potential benefits for the community of the downstream effects, and whether these lead to increased interest in the tool. Information that has been effectively hidden away could be found and retrieved by the community based on themes within the documents.

References

Anita Auer, Moragh Gordon, and Mike Olson. 2016. English urban vernaculars, 1400-1700: Digitizing text from manuscript. In Mar´ıa Jose´ Lo´pez-Couso, Bele´n Me´ndez Naya, Paloma Nu´ n˜ez Pertejo, and Ignacio M. Palacios Mart´ıinez, eds., *Corpus Linguistics on the Move. Exploring and Understanding English through Corpora*. Brill, Leiden.

Joke Daems, Thomas D'haeninck, Simon Hengchen, Tecle Zere, and Christophe Verbruggen. 2019. 'Workers of the World'? A Digital Approach to Classify the International Scope of Belgian Socialist Newspapers, 1885- 1940. In *Journal of European Periodical Studies 4.(1)*, pages 99–114.

Quinn Dombrowski. 2020. Preparing Non-English Texts for Computational Analysis. In *Modern Languages Open*, volume 45.(1), pages 1–9.

Domenico Fiormonte. 2016. Toward a cultural critique of digital humanities. In Matthew K Gold & Lauren F Klein, eds., *Debates in Digital Humanities*, pages 438–458. University of Minnesota Press, Minneapolis.

Michael Wayne Goodman, Ryan Georgi, and Fei Xia. 2018. PDF-to-text reanalysis for linguistic data mining. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mark J Hill and Simon Hengchen. 2019. Quantifying the impact of dirty ocr on historical text analysis: Eighteenth century collections online as a case study. In *Digital Scholarship in the Humanities 34.(4)*, pages 825–843.

Eunjeong Park (Sunchon National University) | *The Affordances and Challenges of Using Learner Corpora to Multilingual Learners' Writing Instruction*

Educators would agree that we need to provide effective instruction for ESL/EFL students' language development and improvement. Long's (1983) study found that formal instruction has advantages to multilingual learners. Second and foreign language researchers have investigated the preliminary impact of input that has been purposefully maneuvered to implementing language instruction (Sharwood Smith & Truscott, 2014). The purpose of this study is to examine the effectiveness of the lexical bundle interventions. Lexical bundles were extracted from the learner corpus of multilingual students' placement test essays. This study analyzed the functional and structural patterns of lexical bundles and utilized them to the second language writing instruction of both intentional and incidental language learning contexts. Findings showed that corpus-aided instruction has some potential to increase multilingual learners' writing skills. However, there were some challenges of using corpus-based materials. First, learner corpora are specialized, so it is often challenging to generalize the findings. Second, it takes considerable time and effort for corpus researchers to understand various types of corpus annotation or markup with refined corpus methods in order to develop and analyze corpus data. Third, there is an issue of how to treat a variety of unwanted or unidentified items. Despite the challenges, this presentation would help ESL/EFL educators and teachers improve awareness of lexico-grammar along with the knowledge and information of corpus linguistics. It is also hoped that the audience can build corpus literacy (i.e., the ability to use the technology of corpus linguistics for language development) to support their multilingual learners by developing 21st-century skills. Based on the preliminary findings, suggestions and implications are discussed.

Hua Tan (Central China Normal University) *Authorship attribution of Chinese Martial Art Fictions by Gu Long with Stylometry*

Gu Long is a known as one the three greatest writers of Wuxia novels (martial arts fictions) in the contemporary era. He was a productive writer, who was said to own authorship for more than 70 Wuxia novels. Yet, many readers and scholars claim that some of his novels were actually produced by other writers, or least partially produced by others. However, such claims were merely speculations without sound evidences. With the development of stylometry, it is possible now to examine through quantitative analysis the linguistic features of the literary works to distinguish different stylistic features, thus attributing the “right” works to the “right” authors. Author attribution is thus widely employed to solve such controversial issues. Usually, authorial analysis turns to such indicators as TTR, STTR, AWL, ASL to attribute authorship. In the present study, we conduct an author attribution analysis to examine the controversial authorship of Wuxia novels credited to Gu Long. We will take a more indicators of different levels into account, so as to have a more scientific and convincing analysis. The indicators examined in our study include vocabulary diversity, syntactic complexity, and dependency relation, which covers such counts as AWL (average word length), MATTR (moving average type-token ratio), ASL (average sentence length), KWL (key words list), MFC (most frequent characters), POS (part of speech) dispersion, DD (dependency distance), DR (dependency direction).

Jonas Müller-Laackman (Freie Universität Berlin) | *Arabic vernacular poetry - challenges in working with Arabic script, speech and (re)presentation*

Dealing with Arabic Script in a digital, Latin-Script-dominated environment requires the mastering of different challenges, the most obvious one being to work with RTL-Script in an LTR context. As part of my doctoral thesis1 on Libyan Arabic concentration camp poetry, I aimed to process the sources I gathered in a machine-readable way, that is, in TEI-based XML to facilitate further research. The poems originated in a mostly unknown mode, supposedly less written than oral. They were transmitted and edited in Arabic Script as part of a pro-nationalist campaign without specific, comprehensible or even accessible editing guideline. Thus, there is not only the problem of dealing with Arabic Script, but with oral dialect transcribed in Arabic Script, which is insufficient for an adequate phonetic rendering of speech. I needed to come up with a solution that allowed me to structure the text and to code entities and rhetorical figures without completely destroying text order or human readability of the XML, as well as to acknowledge the fact that the poem's oral performance is not represented in the Arabic Script. Based on my own far from perfect solution, I would like to discuss other approaches that address similar issues with Non-Latin Script and the processing of NLS-data for computational analysis. I also aim to address the problem of epistemic violence in using Latin-Script based workarounds like mine to deal with the issues of coding Arabic Script in XML and discuss ways of acknowledging this problem as an integral part of the methodology.

Reference

Müller-Laackman, Jonas: *Ein leises Geräusch, wie ein Gefühl des Sehnens. Dichtung und Zeugenschaft zum faschistischen Konzentrationslager in* Libia Coloniale. Reichert-Verlag. Wiesbaden. (publication pending)

Diana Fabiola Zavala Rojas (University Pompeu Fabra), Danielly Sorato (University Pompeu Fabra), Lidun Hareide (Møreforsking AS), Knut Hofland (University of Bergen) | *Multilingual Corpus of Survey Questionnaires*

The dawn of the digital age led to increasing demands for digital research resources, which shall be quickly processed and handled by computers. Due to the amount of data created by this digitization process, the design of tools that enable the analysis and management of data and metadata has become a relevant topic. In this context, the Multilingual Corpus of Survey Questionnaires (MCSQ) contributes to the creation and distribution of data for the Social Sciences and Humanities (SSH) following FAIR (Findable, Accessible, Interoperable and Reusable) principles, and provides functionalities for end users that are not acquainted with programming through an easy-to-use interface.

The Multilingual Corpus of Survey Questionnaires (MCSQ) is the very first publicly available multilingual database comprised of international survey texts. Its latest version (Rosalind Franklin) is composed of 306 distinct questionnaires comprising approximately 766.000 sentences and includes new annotations and datasets to the corpus. The MCSQ is compiled from the questionnaires from the European Social Survey, the European Values Study, the Survey of Health, Ageing and Retirement in Europe, and the Wage Indicator Survey.The surveys are available in English language and their translations into different languages: Catalan, Czech, French (localized language varieties for France, Switzerland, Belgium and Luxembourg), German (localized for Austria, Germany, Switzerland and Luxembourg), Norwegian (Bokmål), Portuguese (localized for Portugal), Spanish (localized for Spain) and Russian (localized for Belarus, Estonia, Israel, Latvia, Lithuania, Russia and Ukraine).

The MCSQ database can be used for such purposes as linguistic research (e.g., analyzing linguistic patterns of survey items, creating bilingual dictionaries of survey terms), questionnaire design (e.g., comparing survey items), multilingual resources for domain specific machine translation (e.g., creating translation memories, translation verification).

Yu Yan (Hubei University of Technology) | *A Corpus-based Study of China's National Image in the English Translations of Government Report*

The official international discourse makes contribution to constructing a country's image. As an authoritative official document of government of People's Republic of China, the Government Report expresses wishes of the country and people and also constructs China's national image and provides opportunities for the international community to fully and objectively understand China and the Chinese government. Previous research on government reports was mostly focused on exploring the characteristics or changes of government functions. Some scholars also used critical discourse analysis to explore the national image presented by the diachronic changes in the collocation characteristics of key words in the Government Reports. However, in general, there is a lack of research on the national image presented by the translated version of the Government Report. Thus, in light of critical discourse analysis, the present paper, using corpus methodology, describes the linguistic features of the English translations of Government Report(2016-2021), and analyzes the national image of China in the reports from the perspectives of high-frequency words and key words. The high-frequency verbs in the English translation of Government Report revealed that the Chinese government is pragmatic and diligent and also shaped an enterprising and promising image of China. Besides, the high-frequency nouns highlighted the achievements of China in the past few years and constructed the image of China as a powerful country with self-reliance and unremitting effort. Additionally, the keywords reflected China's strong national governance capacity and sense of responsibility, and strengthened the image of China as a big socialist country giving priority to people's well-being. Therefore, it is argued that the research on official international discourse is conducive to telling China's story well, eliminating negative stereotypes and promoting the construction of a positive international image.

Joshua Brown (Australian National University) | *Digital approaches to multilingual text analysis: the Dictionnaire as a code-intermediate space*

The first responds to the ongoing development of multilingual digital humanities as it relates to multilingual texts. Specifically, it complicates the question of what "multilingual" and "text" mean, in situations of language contact that render both terms ambiguous. In some recent evaluations of work done in digital humanities and area studies, linguistic approaches have been deliberately (and notably) left aside (Armstrong & Patti 2020). As Dombrowski (2020) has noted, "most methods for computational text analysis involve doing things with 'words'". But in many cases of language contact, it is difficult (if not impossible) to ascribe a particular 'word' to any one linguistic variety in a categorical way: code-mixing between Italian and English, for example, leads to data being produced such as *fensa* 'fence', with an English nominal stem, but Italian morphological inflection. How is it best to proceed we try to analyze digitally varieties of languages that do not belong to any one linguistic taxonomy? The need for computers to be able to assign binary values complicates a vast swathe of linguistic production which cannot easily be interpreted as being Italian, French, English, etc. This is particularly true for textual materials which are purported to contain evidence of 'mixed' varieties of languages, such as lingua francas.

The second part conceptualizes this broader framework to one specific document from the past: the *Dictionnaire de la langue franque* of 1830. This dictionary, written by an anonymous author in Marseille, purports to record the most comprehensive and complete lexical entries for a Mediterranean trade language used throughout the early modern period. It provides entries of 'lingua franca' in the left-hand column, and a French correspondent in the right-hand column. This section introduces an ongoing project to create a database of the forms purported to be in 'lingua franca' (Brown 2021). The question of how best to represent linguistic forms in a digital space ultimately leads to further questions about the corpus itself. In particular, I canvas the orthographical and morphological issues arising in text encoding and lemmatizing in preparing the non-English elements for computational analysis, expanding on the methodological issues discussed by Vanetik & Litvak (2019). A brief conclusion reflects further on Fiormonte's question about whether a non-Anglo DH is possible (Fiormonte 2016).

References

Armstrong, G. and E. Patti (2020). "Italian Studies and the Digital." *Italian Studies* **75**(2).

Brown, J. (2021). "On the existence of a Mediterranean lingua franca and the persistence of language myths". *Language Dynamics in the Early Modern Period. Volume 1.* Ed. by K. Bennett and A. Cattaneo. London, Routledge.

Dombrowski, Q. (2020). "Preparing Non-English Texts for Computational Analysis." *Modern Languages Open* 45(1): 1-9.

Fiormonte, D. (2016). “Toward a Cultural Critique of Digital Humanities”. *Debates in Digital Humanities.* Ed. by M. K. Gold and L. F. Klein. Minneapolis, University of Minnesota Press**:** 438-458.

Vanetik, N. and Litvak, M. (2019). “Multilingual Text Analysis: History, Tasks, and Challenges”. *Multilingual Text Analysis. Challenges, Models, and Approaches*. Ed. by M. Litvak and N. Vanetik. Singapore: World Scientific: 1-29.

Katrina Grant (Australian National University) | *Final remarks and lead into discus*sion