Institution: SOAS University of London

Unit of Assessment: 26 - Modern Languages and Linguistics

Title of case study: Digitization of Tibetan texts: a new resource for learners, teachers and Buddhist practitioners

Period when the underpinning research was undertaken: 2012–2015

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Nathan Hill	Reader in Tibetan and Historical Linguistics	2008–present
Ulrich Pagel	Professor of Languages and Religions of Tibet and Central Asia	1993–present
Edward Garret	Research assistant	2006–2020

Period when the claimed impact occurred: 2014–31 July 2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact (indicative maximum 100 words)

Research conducted at SOAS between 2012 and 2015 played a key role in supporting the conservation and strengthening of the Tibetan language – both Classical and Modern. Because of the limited commercial and legal status of Tibetan, the interest of major software companies in the development of language tools for Tibetan-speaking users – such as predictive keyboards, spell checkers, or voice recognition – and the creation of online learning and language conservation resources have been limited. The research provided the necessary building blocks to process and analyse Tibetan in digital environments. It consequently facilitated the development of language technologies by major multinational companies such as Google and Microsoft, and also assisted in the creation of educational materials for both Classical and Modern Tibetan by learning-focused organizations, thus benefiting Tibetan speakers and learners across the world.

2. Underpinning research (indicative maximum 500 words)

Tibetan is a language of significant heritage and great cultural value, particularly because of its role in transmitting Buddhist literature, but it is also used daily by some 6 million speakers in Tibet and an international diaspora. Before 2012, Tibetan speakers had effectively no modern digital technologies available to them and Tibetan learners had no learning resources that take advantage of modern technologies. In short, although in principle included in the Unicode encoding standard since the early 2000s, Tibetan was not fully 'digitized', endangering its preservation for future generations. 'Digitization' refers to the converting of information into a structured digital (i.e. computer-readable) format. This process has been applied to many languages. However, it has faced significant challenges for languages such as Tibetan, which are written in unique orthographic systems very different from the Roman script of English and other major European languages. Furthermore, computer tools take many of the features of Western languages for granted, including the use of spaces between words – which does not apply to Tibetan. These tools simply do not work when applied to languages that don't conform to these assumptions.

Research supported by an AHRC grant (GBP448,563) conducted at SOAS between 2012 and 2015 by Dr Nathan Hill (Reader, Tibetan and Historical Linguistics, at SOAS from 2008), Prof Ulrich Pagel (Professor, Languages and Religions of Tibet and Central Asia, at SOAS from 1999) and Dr Edward Garrett (Research Assistant in Tibetan in Digital Communication at SOAS from 2006) has been at the forefront of fully bringing Tibetan into the digital era.

In order to explore linguistic questions, such as the structure of the noun phrase [3.4], the team developed key applications and tools for analyzing Tibetan, together with a gold standard corpus of Tibetan texts. This corpus is a powerful resource for scholars and non-scholars working with language, since it offers ready access to and comparison across texts from different time periods, regions and genres. Solving the technical problems in how computers handle Tibetan resulted in the development of tools and resources that have contributed fundamentally to Tibetan Natural Language Processing (NLP), and have a wide application within and outside academia. The body of work, facilitated by additional analysis by SOAS PhD candidate, Abel Zadoks, Di Jiang (Chinese Academy of Social Science), Marieke Meelen (University of Cambridge), and additional expertise on web language by Adam Kilgarriff (Director, Lexical Computing Ltd) and Ravikiran Vadlapudi (Nuance Communications), led to the development of key Tibetan language software tools. These included:

1. A word breaker – (or tokenizer) that allows to decide what a word is in Tibetan language **[3.2, 3.3, 3.6]**;
2. A part of speech tagger – to assign each word a 'part of speech' tag, such as noun, verb, adjective, etc. **[3.1, 3.3, 3.5, 3.6]**;
3. A part of speech tagged corpus (of around 300,000 words) for Tibetan, which allows users to create further software tools of their own design **[3.6]**.

These components are prerequisites for digital communication technologies such as speech recognition software; optical character recognition and screen-reading devices for the blind; or cross linguistic communication, including machine-aided translation and inter-lingual tools for the internet. To encourage commercial use of their data and the component developed by the research, Hill and his colleagues published the related datasets with an open-source license **[3.7].**

3. References to the research (indicative maximum of six references)

3.1. Garrett, E. and Hill, N.W. (2015). 'A Constraint Grammar POS-Tagger for Tibetan'. Proceedings of the Workshop on 'Constraint Grammar – methods, tools and applications' at NODALIDA 2015, May 11-13, 2015. Vilnius: Institute of the Lithuanian Language, pp. 19–22. ISBN: 978-91-7519-037-2 https://eprints.soas.ac.uk/20172/ **Peer-reviewed**

3.2. Hill, N.W. and Di, J. eds. (2016). Himalayan Linguistics 15.1 (Special Issue on Tibetan Natural Language Processing). Berkeley, CA: University of California Press. https://doi.org/10.5070/H915131516 **Peer-reviewed**

3.3. Hill, N.W. and Meelen, M. (2017). 'Segmenting and POS tagging Classical Tibetan using a memory-based tagger'. *Himalayan Linguistics,* 16(2), pp. 64–86. https://doi.org/10.5070/H916234501 **Peer-reviewed**

3.4. Garrett, E. and Hill, N.W. (2015). 'Constituent order in the Tibetan noun phrase'. SOAS Working Papers in Linguistics, 17, pp. 35–48. https://eprints.soas.ac.uk/20872

3.5. Garrett, E., Hill, N.W. and Zadoks, A. (2014). 'A Rule-based Part-of-speech Tagger for Classical Tibetan'. *Himalayan Linguistics*, 13(1), pp. 957. https://doi.org/10.5070/H913224023 **Peer-reviewed**

3.6. Garrett, E., Hill, N.W, Kilgarriff, A., Vadlapudi, R. and Zadoks, A. (2015). 'The contribution of corpus linguistics to lexicography and the future of Tibetan dictionaries'. *Revue d'Études Tibétaines*, 32, pp. 51–86. https://eprints.soas.ac.uk/19777/ **Peer-reviewed.**

3.7. Tibetan natural language processing (2017). Open-source dataset. Zenodo. https://zenodo.org/communities/tibnlp/

4. Details of the impact (indicative maximum 750 words)

Dr Hill, Prof Pagel and Dr Garret's research supported the availability of Tibetan resources online and improved the learning of Modern and Classical Tibetan in Tibet and around the world. It made learning Tibetan more relevant to younger generations and helped ensuring that the cultural heritage of Tibet is preserved in the digital domain and handed on to future generations. To achieve this, the SOAS team, particularly Hill, engaged both with multinational technology companies, such as Google and Microsoft, and with Tibetan preservation and learning-focused

organizations, which target the Tibetan diaspora, students of Tibetan, translators, religious communities and communities interested in learning and preserving Tibetan.

Engaging with multinational technology companies to improve Tibetan digitization
Multinational technological companies such as Microsoft have also faced challenges in digitizing Tibetan, thus limiting their ability to provide comprehensive services to Tibetan audiences and learners on a par with other languages. From 2018, Hill engaged with **Microsoft** and helped to solve the problem by providing a larger corpus of digitized Tibetan texts, which enabled Microsoft developers to train language models covering a larger vocabulary and offering more linguistically accurate predictions in the Tibetan version of their SwiftKey predictive keyboard for smartphones. The Natural Language Processing (NLP) Technical Lead at Microsoft indicated how fundamental Hill's support was to develop the ability to process Tibetan: 'in a language like Tibetan, where words are not delimited by spaces, we were restricted to offering predictions of syllables instead of words, which is an extremely poor experience for the user. Dr. Hill provided us with a Tibetan tokenizer . . . that has enabled us to move from syllable predictions to word predictions, putting our Tibetan offer on a par, in terms of quality, with other languages . . . Dr. Hill's word breaker did not just save us time; we simply wouldn't have been able to build a comparable one' **[5.1].**

The Tibetan Swiftkey keyboard had 1,128 active monthly users as of July 2020 **[5.8]**. Microsoft reported that this resulted in a net increase in the quality of Microsoft prediction engine for Tibetan (an average 27% increase in typing speed, as measured on 6600 Tibetan users). Qualitative feedback was also extremely positive, both from users and from partners pre-installing the Tibetan keyboard on smartphones in China [5.1]. A civil servant from the Huangnan Tibetan Autonomous Prefecture, Qinghai Province, China stated: 'I love using the [Tibetan keyboard] auto-completion when type on my phone . . . the slide typing is great, it's the first slide typing keyboard for Tibetan in the world, ever'. A pastoralist from Mgo mang town, Hainan Tibetan Autonomous Prefecture, Qinghai Province, China described how 'recently I found out about the SwiftKey, it is a better keyboard simply because it has auto-completion function. I have using it for a few months now and I am satisfied with it. Hope more people will use it so that we will have more words predictions as auto-completion suggestions' **[5.2]**.

For similar challenges, Hill's support was also fundamental to **Google's** efforts to improve its product offers – including Translate, Chrome, Maps, YouTube, the Google Assistant – in Tibetan language. A Google Speech Analytical Linguist explained how, from March 2016, Hill 'provided consultation, advice and insight into the Tibetan language and script which has greatly benefited our work on Tibetan. His tokenization method for segmenting text into words has been a great source of inspiration, and his comments and help with our spelling-to-pronunciation mapping rules have significantly improved their quality and accuracy . . . [his research project] has significantly advanced our understanding of the complexities of the Tibetan language and script, and greatly impacted research and development of natural language processing solutions for the language . . . to better support our Tibetan-speaking users' **[5.3]**.

Improving learning technologies and resources to promote and preserve Tibetan language among interested audiences
Esukhia is an NGO dedicated to the professional development of Tibetan Language Resources for the use of scholars, translators and Buddhist practitioners both in Asia and the West. Esukhia largely teaches students through online courses around the world. It also provides face-to-face courses in Dharamsala (India), and short programmes in Ladakh and Goa (India). More than 2000 students enrolled with Esukhia from 2012. Hill worked closely with Esukhia from May 2014, providing the foundation for Esukhia's own research and development of Tibetan language NLP. Esukhia considered that 'simply put, without [Hill's] research and his support, we would not have [been] able to create these resources' **[5.4]**. More specifically, Hill provided the technology and tools that underlay many important projects by Esukhia, including the planned definitive edition of the Tibetan Buddhist Canon – a key publication for academic studies, traditional monastic contexts as well as for large-scale translation projects, such as 84000: Translating the Words of the Buddha. Esukhia also graded texts and created learning resources for beginning readers – including Tibetan-speaking children and foreigners wanting to learn Tibetan as well as more

professional and innovative Tibetan language textbooks. Esukhia indicated that the beginner-level textbook had already reached nearly 1,000 users across 63 countries by July 2020 [5.4], filling a gap in availability of learning resources for language learners in UK and USA. Esukhia also reported that feedback from independent users had been 'overwhelmingly positive in helping new students gain a foothold in a language many have trouble learning'. Users expressed their gratitude for these new learning resources as compared to others they have tried: one user described how they now 'enjoy learning Tibetan!'; another commented how easier it was compared to previous attempts, describing it as an 'exciting way to learn which WORKS!!!' **[5.4]**.

The **Buddhist Digital Resource Center (BDRC)** is a Boston-based non-profit organisation dedicated to seeking out, preserving, organizing and disseminating Buddhist literature through the digitisation and preservation of Tibetan texts in a free online library used by scholars, translators and representatives of the tradition, especially Tibetan lamas and monks. The online library has over 5,000 active monthly users from over 15 countries – over half of the hits coming from Asia (mainly China, India, Nepal and Japan) where the Tibetan diaspora lives. From July 2013, Hill worked with BDRC to ensure that their internal search engine could search through Tibetan text, improving user experience and making a more sophisticated database. BDRC stated that its 'technical team has embraced the breakthroughs introduced by Dr. Hill and have been exploring ways to build upon his momentous contribution to this key resource in Tibetan Studies', including planning for the development of new improvements such as search ranking, named entity recognition, automated line alignment between multiple editions of a scripture, and lexicography tools, thus enriching the work of scholars and translators worldwide. BDRC was also 'proud to have amplified [Hill's] innovations to a global audience' **[5.5]**.

Nettle is the only open-source, open-access comprehensive learning resource developed for online delivery of instruction in Classical Tibetan language. It was developed by Professor Frances Garrett, University of Toronto based upon the tokenizer developed by Hill. Garrett said that without the research undertaken by SOAS, Nettle would have been different and not as sophisticated as it is **[5.6]**. Nettle has 2 main user groups, Toronto-based students and Classical Tibetan language learners around the world. From 2016 when Nettle was developed and introduced, allowing for more flexibility through online classes, approximately 25 to 40 University of Toronto-based students took Tibetan each year – compared to around 5 students in previous years. Student feedback was positive, with class evaluations ranging from average to high. Nettle can also be used by anyone online, although it does not track online, non-student users, Garret received many emails from around the world expressing gratitude for the online resources. Several came from India and Nepal, but also from Europe **[5.6]**.

In May 2019, Hill was contacted by Christian Steinert, a software developer working for the **Foundation for the Preservation of the Mahayana Tradition (FPMT)**. Steinert developed a Tibetan-English dictionary application for mobile phones; he explained that he very much benefitted from Hill's research, with which he had enhanced the app substantially. After the app was revised thanks to Hill's research, the userbase of the dictionary almost tripled and in 2019 the app was used regularly by over 4,500 users all around the world. User feedback was also overwhelming positive: 'I now use this as my primary dictionary, replacing both Monlam and The Tibetan Translation Tool . . . The inclusion of the Verb Lexicon is very helpful, as verb tense is not included in complete form anywhere else. Thank you for this dictionary'; 'The results are much more detailed than the other dictionaries I've used. It even shows you tenses and their correct spellings'. Hill's research also helped Steinert as an FPMT tutor providing non-academic online courses about Buddhist philosophy **[5.7]**.

5. Sources to corroborate the impact (indicative maximum of 10 references)

5.1. Letter from Technical Lead, Natural Language Processing, Microsoft. Feb 2019

5.2. Email from Tibetan Buddhist Resource Centre, March 2020

5.3. Letter from the Analytical Linguist, Google Speech, October 2019

5.4. Letter from Esukhia

5.5. Letter from the Buddhist Digital Resource Center, December 2018

5.6. Letter from Frances Garrett, Associate Professor, University of Toronto, March 2020

5.7. Letter from Christian Steinert, Foundation for the Preservation of the Mahayana Tradition, May 2019

5.8. Update email from Technical Lead, Natural Language Processing Microsoft, Dec 2020