13 January 2025

English

United Nations Group of Experts on Geographical Names
2025 session
New York, 28 April – 2 May 2025
Item 5 (c) of the provisional agenda*
Technical expertise: Writing systems and pronunciation

Report of the Working Group on Romanization Systems: current status of United Nations romanization systems for geographical names

Submitted by the Working Group on Romanization Systems

* GEGN.2/2025/1.
**The full report was prepared by Catherine Cheetham, United Kingdom of Great Britain and Northern Ireland, Convenor of the Working Group on Romanization Systems.

The report of the Working Group on Romanization Systems (WGRS) outlines the main developments in respect of the status of romanization systems on the agenda of the Working Group on Romanization Systems; it also considers innovative methods of romanization and calls for input from UNGEGN experts on the romanization systems used in their domestic contexts.

Progress on romanization systems, UNGEGN's Strategic Plan, Action 1-i-4

In Action 1-i-4, the WGRS undertakes to promote, agree and disseminate romanization systems for national and global use. This is the core role of the WGRS, but progress depends on the member states providing updates on their efforts to develop and/or implement romanization systems. Progress on specific writing systems and romanization systems is recorded here:

Kazakh

The WGRS has continued to monitor the status of the introduction of a Roman-script alphabet for the Kazakh language. As reported in 2023, the proposed alphabet discussed at the 2019 UNGEGN Session had undergone some further changes, and the timeframe given for its adoption as the main alphabet for the Kazakh language, replacing Kazakh Cyrillic, in Kazakhstan is now reported as 2023-2031. The WGRS has attempted to learn of developments, though has not had been successful in receiving confirmation from Kazakhstan regarding the status or implementation of the new alphabet.

Lao

As reported at the 2023 Session, the WGRS had been invited to review a draft of the Lao Romanization System prepared by the National Geographic Department of the Lao People's Democratic Republic. The WGRS had made certain recommendations and were keen to resume the discussion with the Lao government in order to finalise the system and consider it for adoption as a UN system. As of July 2024, the WGRS learnt that Lao PDR was in the process of finalizing a draft of toponymy guidelines, and it is hoped that the system will be put forward to the UNGEGN plenary for consideration at the 2027 Session.

Sinhala

As reported at the 2021 session, Sri Lanka's Cabinet of Ministers officially accepted the proposed romanization system for Sinhala in 2018, although some modifications to the system were made subsequently. The WGRS remains keen to discuss with Sri Lanka the system's being put forward to the UNGEGN plenary. With the approval of a system for Sinhala, the UN and international community would have a recognized and agreed method of representing Sinhala geographical names. Until now, the unapproved 1972 'Sharma' system has been applied in such instances, in the absence of anything more official. Once the system is submitted to UNGEGN, the WGRS will be able to consider its adoption, noting that, as well as being scientifically sound, UNGEGN needs to see evidence of its implementation.

Uzbek

As reported at the 2023 Session, the WGRS is aware that although Uzbekistan adopted the Roman script for the Uzbek language in the 1990s, the adoption and implementation of the Roman script has not been complete, and there have frequently been proposals to alter the alphabet used. In 2022 an alternative alphabet was proposed, with a stated intention to introduce the altered alphabet in 2023. Since the 2023 Session, in September 2023, the WGRS has learnt that an alternative proposal

was introduced[1]. The WGRS has not been able to verify the introduction of the altered alphabet and has not received further information from Uzbekistan. The WGRS will continue to monitor this.

The WGRS also calls for further information on the projects relating to romanization system development, particularly for those member states that do not use the Roman script and do not yet have an UNGEGN-approved method of transliteration.

Innovative methods of romanization, Action 1-iii-10

In Action 1-iii-10, the WG Romanization Systems has undertaken to investigate and report on innovative methods of romanization. With the development of digital technologies including Artificial Intelligence (AI) and Large Language Models (LLMs), tools that can automate the process of romanization, potentially dealing with large amounts of data very rapidly, while also improving usability, accuracy, or efficiency have already and will continue to become more readily available.

Already, even freely available LLMs can apply scientific romanization systems quite successfully; in many cases, they are able to understand structured rules and can apply letter-by-letter mappings accurately. Large texts can therefore be managed efficiently with perhaps more accuracy than humans. As the tools develop, they will also make increasingly good judgments in cases such as the unwritten short vowels of Arabic, that are not encoded in the original script; with access to many place names and linguistic patterns, these tools will often infer missing short vowels based on word context, grammar or knowledge of a specific place name.

Admittedly, if not specifically trained and without adequate oversight, some errors may be introduced. For example, LLMs are trained on real-world text, which means they sometimes favour common transliterations or conventional names over strict official rules (for instance, the UNGEGN system for Ukrainian renders *Zaporizhzhia*, but many sources use *Zaporizhia*, so an LLM might default to the more common spelling).

Some tools, such as Google's Input Methods, allow a two-way transliteration approach whereby users can type in the Roman script and have it intelligently converted back into the native script. Such tools are not scientific romanization tools, rather they take an adaptive and flexible approach that combines elements of formal systems with real-world typing habits and phonetic approximations. They rely on machine learning and user data, looking at common user input patterns, rather than fixed, rule-based romanization systems. However, with a widely-used romanization system, it is likely that the tools will use that system, for instance for Chinese Pinyin.

There are also some phonemic and simplified approaches that prioritize phonetic simplicity over script accuracy. For instance, informal romanization (so-called 'chat language') of Arabic (Arabizi) or Greek (Greeklish) is based on how native speakers might approximate sounds in Roman script letters. However, though the tools are of value to casual users, these will not be successful in applying the strict scientific romanizations needed for the unambiguous representation of geographical names. It will be of interest with the increasing multilingualism of the digital world, as tools and applications improve for non-Roman script languages, whether the use of such approaches wanes.

However, for UNGEGN's purposes, specifying the application of a particular romanization system in a LLM (feeding it with the content of a system) can also have very successful results. The WGRS continues to call for the experience of UNGEGN's experts on their use of any innovative and automated methods of romanization or script transfer.

[1] These changes bring the system closer to the Turkish alphabet, the most recent (in 2023) being to alter: O' o' to Õ õ, G' g' to Ğ ğ, Sh sh to Ş ş and Ch ch to C c.

Call for information on romanization systems used in domestic contexts

A request for information is also made to UNGEGN experts on the use of romanization systems in their domestic contexts. Information on the use of UNGEGN-approved, or other non-UNGEGN systems, is solicited. WGRS would like to collect data on romanization systems used, e.g. in a nationally-produced world atlas? And if these are not UNGEGN systems, WGRS would like to collect information on the reasoning for adopting alternatives.

UNGEGN romanization systems are now available on the Working Group's webpages, which can be found in subpages of the UNGEGN site: https://unstats.un.org/unsd/ungegn/working_groups/wg5.cshtml. They remain archived at the former WGRS website http://www.eki.ee/wgrs. \

Points for discussion

The Group of Experts is invited to:

(a) Express its views on the progress made by the Working Group in the intersessional period and on its suggested activities in the coming intersessional period;
(b) Provide case studies on how Large Language Models (LLMs) can support the development and use of romanization systems; and,
(c) Provide input on its requests for information on the use of romanization systems in their domestic contexts.