isif asia

fast facts

Project: Web based Transliteration and Translation System between Urdu and Hindi Languages
Lead Organization: Punjabi University
Country: India

situation

The problem of communication between Hindi and Urdu languages has long been a social barrier between the Muslim populations of India and the majority Hindi public. Although the languages share more than 70% of their words, and follow nearly identical grammatical conventions, the writing systems are completely distinct.

solution

The aim of this project is to facilitate electronic and written communication between people living in India and Pakistan through the development of a bi-directional web based Hindi-Urdu Language Transliteration/Translation Tool. The target groups will be media organizations, literary and literacy promotional organizations, writers and NGOs involved in dissemination activities in areas where the two languages coexist. This project will develop a complete machine translation and transliteration system between Hindi and Urdu languages as facilitate the use of these technologies on the web, thus enhancing networking between India and Pakistan.

This translation tool enables Hindi and Urdu literate people to convert Urdu websites to Hindi and vice versa, allowing each user to read content in their primary language. The interface is designed to allow the user to see the meaning of Urdu words by placing the mouse cursor over any word about which they may be confused or curious. This allows the text to remain visually "clean" (without distracting pop-ups and unnecessary information) as only relevant information is shown at any point in time.

One of the main challenges in the research process has been the lack of useful pre-existing research about linguistic equivalents which could be easily integrated into the system. Instead, the project team found it necessary to perform its own preliminary research; a process which has proved to be longer and more time intensive than anticipated. Fortunately, however, the team has been successful in resolving the majority of the issues.

Initially the team attempted to develop an option for translating Urdu text into Hindi but realized that in order to perfectly implement this task they would need much more time than was available. Instead, the team chose to focus on improving the output of the existing translation system, "Sampark" and how it might be improved in the new tool. By the end of development, the project team's translation software was able to provide a near- perfect transliteration at accuracy of more than 98%, whereas the transliteration accuracy of Sampark is just 64%.

The current iteration of the tool consists of four main modules: A Urdu-Hindi electronic dictionary, transliteration/translation rules and mapping tables, parallel Hindi/Urdu corpus, and a tool to convert any Unicode based Hindi website to Urdu and reverse. User research confirms that the outputs of the team's Urdu-Hindi translation system are far easier to understand than those of any other existing software seeking to provide the same service. It is currently considered to be the best transliteration system for Urdu text available.

broader impact

As a result of the research and development necessary to complete this project the project team has been able to develop new skills in the areas of Machine Transliteration and building Internet-based systems using Unicode, an industry Standard designed to allow text and symbols from all of the writing systems of the world. The research center itself has also increased its capacity in terms of general advancement of Urdu and Hindi. The development of Unicode-based system components is expected to have wide and long-term impacts on their future projects. By promoting and developing the use of Urdu script the team is engaging in the active development and promotion of a marginal minority group language.

Any future technical developments and enhancements made by the center will be managed by the director of the Advanced Centre for Technical Development of Punjabi Language, Literature & Culture in collaboration with other center staff and students, thus ensuring the long-term sustainability of the project.

project contact

Gurpreet Singh Lehal
Email Address: gslehal@gmail.com
Advanced Centre for Technical Development of Punjabi Language, Literature & Culture. Punjabi University
Punjabi University, Patiala (Punjab)
India 147 002.
+91 1 75 3046171
www.learnpunjabi.org