Names are of the essence in the media but display a great variability in spelling and even scripts through the variety in languages and each language’s individual pronunciation system. In an NWO Kiem project we built an app that focuses on recommending Dutch spellings for originally Cyrillic names to news reporters who more often than not first encounter new Slavic names in English, French or German newswire, rather than in the original Cyrillic script. The project’s user groups are ANP, NOS, Open Taal, VRT and the Dutch Language Union who all strive for more conformity in news reports. The CLAM-based Transcriptor web application is now online.

Main idea

The system is in part based on a series of rule-based transliterators. These emulate the transliteration systems used in the countries themselves (e.g. on signposts and passports), systems that are used in the scientific world and libraries, and transliterations intended for a general audience in other countries. These latter are mainly based on the pronunciation, and therefore differ by language. This means names are usually spelled differently across and even within languages.

Quick Search

Input for the Quick Search may be an English version of a name in Latin script or the Russian version in Cyrillic script.

Extended Search

The other part of the system: data-oriented querying.

Fuzzy, metadata guided database search

Anagram hashing performs a speedy fuzzy lookup through the names databases of JRCnames for personal names and Geonames for place names. Next to the rule-based systems, this gives the user corpus-based evidence and guidance.

The system

The Transcriptor web application/webservice is based on CLAM. Quick search relies on CLAM ‘Actions’, the extended search is based on CLAM ‘Projects’. The system incorporates the lexical variation look-up system of TICCL. The Quick Search transliterators have been implemented as Finite State Automata by means of FLEX.

Acknowledgements

The authors acknowledge the Netherlands Organisation for Scientific Research (NWO) for granting a ‘Kiem’ project to Nicoline van der Sijs. Also CLTS for providing the necessary funds to have the current interface developed. They wish to thank Matje van de Camp of De Taalmonsters for the actual interface development.