

A universal transliteration approach to web input methods

Venkat Ramdass
www.linguaseek.com
Santa Clara, CA
venkat@linguaseek.com

ABSTRACT

While English is undoubtedly the dominant language when it comes to content on the web, non-English content is growing, albeit slowly. There appears to be a circular situation that is behind this delay. The lack of proper input tools to generate content is causing potential content generators towards English. This lack of content is then forcing users (content consumers) to look only for English content. Lack of tools for users to query content is also a primary driver in users losing interest (and the ability) in querying for non-English content. This paper presents a tool and an approach to solving both of these problems: non-English content generation and content querying. It solves the problem by creating a single 'universal transliteration tool' that can help transliterate from English to virtually every electronically available language without the need for complex user interfaces or having to learn complicated keystroke combinations.

Keywords: non-English, transliteration, multi-lingual

1 INTRODUCTION

Linguaseek is a simple solution to a major problem today. It provides content generators and users a tool to facilitate the generation and querying for non-English content without the need for custom keyboards and complex solutions. The site provides transliteration to virtually every electronically available script through a single and simple interface. Most languages are transliterated using phonetic spellings, so there is no need to learn custom key combinations. Some languages support well documented and understood standards, so users don't have to learn new schemes. Another major enhancement linguaseek provides a user interface that allows users to edit machine generated transliterations using variants provided on screen. This saves a significant amount of time that could be lost trying to figure out accurate phonetic spellings. It also provides a gradual learning curve to become more familiar with the system.

While linguaseek.com provides the ability to search the web and also a page to use as a writing pad, these are primarily a demonstration of the capabilities behind the system. The technology can be applied to any web page/form to enable multi-lingual input. Be it a blog post, sign-up form, a web query, a site query or any other kind of input.

Linguaseek has a few modes of operation. Input fields can operate in a single language mode or a multi-lingual mode. The default home page for linguaseek currently operates in a single language mode. The language you pick is applied to the entire content in the field. The content generation page ('write' tab) on linguaseek operates in a multi-lingual mode. In this mode, you can switch between any number of languages between words. You can, for example, start a sentence in English, switch to Arabic, then to Russian, then to Chinese and so on.

Linguaseek also provides its services as a simple web service using a simple URL format. The documentation is provided on the site. This is for those who might want an occasional transliteration and also might want to limit the number of languages.

Linguaseek has prototype versions of java applets and complete client-side JavaScript implementations as an experiment and for potential commercial licenses. These versions speed-up transliterations by eliminating the round-trip overhead incurred for every keystroke in the current linguaseek.com implementation.

2 SYSTEM OVERVIEW

Linguaseek's transliteration is a combination of client-side UI application and a server-side transliteration engine. The transliteration engine was built from the ground up. There are no third-party applications or libraries providing any language capabilities. National and international standards for Romanization (language to English) were leveraged and enhanced to apply transliteration in the reverse direction. When available, standards such as Pinyin for Chinese were used.

Server-side is a J2EE application running on JBoss application server behind an Apache HTTP server. Client-side is HTML and JavaScript. The client-side script leverages several current technologies such as XMLHttpRequest and CSS to provide a rich user interface, and therefore only supported on Firefox 2.x and IE 7.x. Other browsers might work, but have not been tested.

3 LIMITATIONS

Linguaseek currently assumes that users have necessary fonts and language support enabled on their systems.

Some documentation could be provided on how to accomplish this for a few operating systems.

4 FUTURE WORK

Linguaseek started out primarily as an experiment and as a vehicle to pursue an interest in languages. It has now evolved to a stable product and has components that can be plugged into other web sites. Evaluating the potential for publishing HTML code snippets that can be embedded into any webpage to automatically enable multi-lingual input. Other ideas are welcome.

PRODUCT INFORMATION

URL: <http://www.linguaseek.com/>

Feedback: feedback@linguaseek.com