BabelMeSH2 and PICO Linguist2: Combined Language Search for MEDLINE/PubMed

Fang Liu, Paul Fontelo, Michael Ackerman
National Library of Medicine, Bethesda, Maryland 20894

Abstract

BabelMeSH2 is a transparent interface for searching MEDLINE/PubMed in one or a combination of nine currently supported languages. The search algorithm automatically detects mix language entries, finds the English equivalent, and retrieves the relevant PubMed citations. We believe this is the first search application that allows mixed language entries.

Background

For journal citations added to MEDLINE/PubMed since 2000, around 47% are from U.S. publications with 90% published in English. About 79% have English abstracts by authors of articles [1]. Journals not published in English have English titles and abstracts. For those whose primary language is not English, the benefits of a tool that allows searching in their native language are obvious. It is easier to search using a language one is very familiar with. Sometimes finding English equivalents are challenging. We previously reported on BabelMeSH and PICO Linguist, cross-language search tools in nine non-English languages where the user had to select the appropriate language interface first before searching [2]. In BabelMeSH2 [3] and PICO Linguist2 [4], we present a unified interface, allowing multilingual users to search in combined languages.

Methods

The integrated interfaces are shown in Figs. 1 and 2. To identify language types used, the parser algorithm first detects Latin and non-Latin terms in the input. The translation process then divides into two pathways: for non-Latin based terms, the parser searches the Arabic, Chinese, Japanese and Russian databases. For Latin terms, the parser looks for matches in the French, German, Italian, Portuguese and Spanish databases. The search is optimized based on database size; smaller databases are searched first. Each database is searched until a match is found. If the input term is found in a database, that database is searched again no matter its size because of a high probability that the succeeding term will be in the same language. The process is repeated until all the search terms are matched or none is found. For unmatched terms, GSpell spellchecker is used to recognize English terms. Unrecognized terms are then removed. The English equivalents are sent to PubMed via e-utilities. Relevant citations will be returned to the user in (Figs 1 and 2).

Fig 1. Mixed Arabic/French search and result

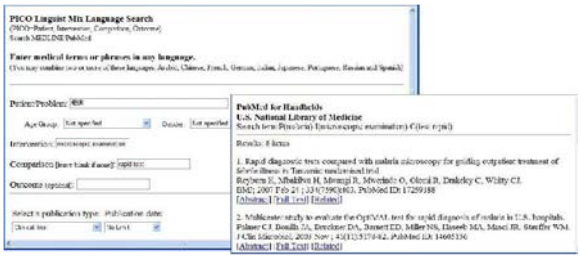

Fig 2. PICO Chinese/English search and result

Discussion

A two language combination of Latin and non-Latin based terms search may take two seconds or longer. As the search strategy gets more complex, involving three or more languages or more terms, the search may take longer. The results are comparable to previous version of BabelMeSH and PICO Linguist. Current development efforts are geared towards optimizing search speed.

Conclusion

BabelMeSH2 and PICO Linguist2 offer the multilingual user a unified search tool for searching MEDLINE/PubMed in several languages.

References

[1] MEDLINE Fact Sheet http://www.nlm.nih.gov/pubs/factsheets/medline.html
[2] Liu F, Fontelo P, Ackerman M J. BabelMeSH: Development of a Cross-Language Tool for MEDLINE/PubMed. Proc. AMIA 2006, p. 1012.
[3] http://babelmesh.nlm.nih.gov/index_mix.php
[4] http://babelmesh.nlm.nih.gov/pico_mix.php