Learning to find transliteration on the Web

Chien-Cheng Wu
Department of Computer Science
National Tsing Hua University
101 Kuang Fu Road, Hsin chu, Taiwan
d9283228@cs.nthu.edu.tw

Jason S. Chang
Department of Computer Science
National Tsing Hua University
101 Kuang Fu Road, Hsin chu, Taiwan
jschang@cs.nthu.edu.tw

This prototype demonstrate a novel method for learning to find transliterations of proper nouns on the Web based on query expansion aimed at maximizing the probability of retrieving transliterations from existing search engines. Since the method we used involves learning the morphological relationships between names and their transliterations, we refer to this IR-based approach as *morphological query expansion for machine transliteration*. The *morphological query expansion* approach is general in scope and can be applied to translation and transliteration, but we focus on transliteration in this paper.

Many texts containing proper names (e.g., “The cities of Mesopotamia prospered under Parthian and Sassanian rule.”) are submitted to machine translation services on the Web every day, and there are also service on the Web specifically target transliteration of proper names, including *CHINET* (Kwok et al. 2005) ad *Livetrans* (Lu, Chien, and Lee 2004).

Machine translation systems on the Web such as *Yahoo Translate* (babelfish.yahoo.com) and *Google Translate* (translate.google.com/translate_t.g) typically use a bilingual dictionary that is either manually compiled or learned from a parallel corpus. However, such dictionaries often have insufficient coverage of proper names and technical terms, leading to poor translation due to out of vocabulary problem. The OOV problems of machine translation or cross language information retrieval can be handled more effectively by learning to find transliteration on the Web.

Consider Sentence 1 containing three place names.

1. *The cities of Mesopotamia prospered under Parthian and Sassanian rule.*
2. *城市繁榮下parthian 達米亞、sassanian統治。*
3. 美索不達米亞城市在巴底亞和薩珊統治下繁榮起來。

Google Translate produce Sentence 2, leaving “Parthian” and “Sassanian” not translated. A good response might be a translation like Sentence 3 where all place names have appropriate transliterations (underlined). These transliterations can be more effectively retrieved from mixed code Web pages by extend each of the place names into a query (e.g., “Parthian NEAR 巴”). Intuitively by requiring one of likely prefix transliteration morphemes (e.g., “巴” or “帕” for “par-“ names), we can bias the search engine towards retrieving the correct transliterations (e.g., “巴底亞” and “帕提亞”) in snippets of many top-ranked documents.

The method involves pairing up the prefixing morphemes between name and transliteration in a set of train data, calculating the statistical association for these pair, and selecting pairs with a high degree of statistical association. The results of this training stage are morphological relationships between prefixes and postfixes of names and transliterations. At run time, a given name is automatically extended into a query with relevant prefixing morphemes, then the query is submit to some search engine. After retrieving snippets from a search engine, the system extract transliterations from the snippets based on redundancy, proximity between name and transliteration, and cross language morphological relationships of prefix and postfix.

We present a new machine transliteration system based on information retrieval and morphological query expansion. The system automatically

NAACL HLT Demonstration Program, pages 21–22,
Rochester, New York, USA, April 2007. ©2007 Association for Computational Linguistics

learns to extend the proper names into a query expected to retrieve and extract transliterations of the proper names. Consider the case of transliteration of "Parthian." The system looks at possible prefixes of the given name, including *p*-, *pa*-, *par*-, and *part*-, and determine determines the best *n* query expansions (e.g., "Parthian 巴," "Parthian 帕 "). These effective expansions automatically during training by analyzing a collection of 23,615 place names and transliterations pairs.

We evaluated the prototype system using a list of 500 proper names. The results show that 60% of the time there are sufficient relevant data on the Web to carry out effective machine transliteration based on IR and morphological query expansion. Of many results returned by the system, the top 1, two and three results are 0.88, 0.93, and 0.94. By performing query expansion, the system improves the recall rate from 0.48 to 0.60.

The results indicate that most names and transliteration counterparts can often be found on the Web and the proposed method are very effective in retrieving and extracting transliterations based on a statistical machine transliteration model trained on a bilingual name list. Our demonstration prototype shows alternative transliterations in use on the Web and snippets of such usage, so that the user can easily validate these transliterations.

The prototype supports:

- Searching and extracting transliterations of a given term
- Listing alternative transliterations on the Web
- Listing alternative transliteration in a local dictionary
- Browsing of snippets containing for each alterative transliteration
- Saving transliterations in a local dictionary
- Selecting and saving transliteration in snippets to a local dictionary

The method explored here can be extended as an alterative way to support such MT subtasks as back transliteration (Knight and Graehl 1998) and noun phrase translation (Koehn and Knight 2003). Finally, for more challenging tasks, such as handling sentences, the improvement of translation quality probably will also be achieved by combining this IR-based approach and statistical machine translation. For example, a preprocessing unit may replace the proper names in a sentence with transliterations (e.g., mixed code text such as Sentence 4) *on the fly* or by looking up a local dictionary before sending it off to MT for finally translation.

4. *The cities of* 美索不達米亞 *prospered under* 巴底亞 *and* 薩珊 *rule*.

Morphological query expansion represents an innovative way to capture cross-linguistic relations in name transliteration. The method is independent of the bilingual lexicon content making it easy to adopt to other proper names such person, product, or organization names. This approach is useful in a number of machine translation subtasks, including name transliteration, back transliteration, named entity translation, and terminology translation.

References

Y. Cao and H. Li. (2002). *Base Noun Phrase Translation Using Web Data and the EM Algorithm*, In Proc. of COLING 2002, pp.127-133.

K. Knight, J. Graehl. (1998). *Machine Transliteration*. In Journal of Computational Linguistics 24(4), pp.599-612.

P. Koehn, K. Knight. (2003). *Feature-Rich Statistical Translation of Noun Phrases*. In Proc. of ACL 2003, pp.311-318.

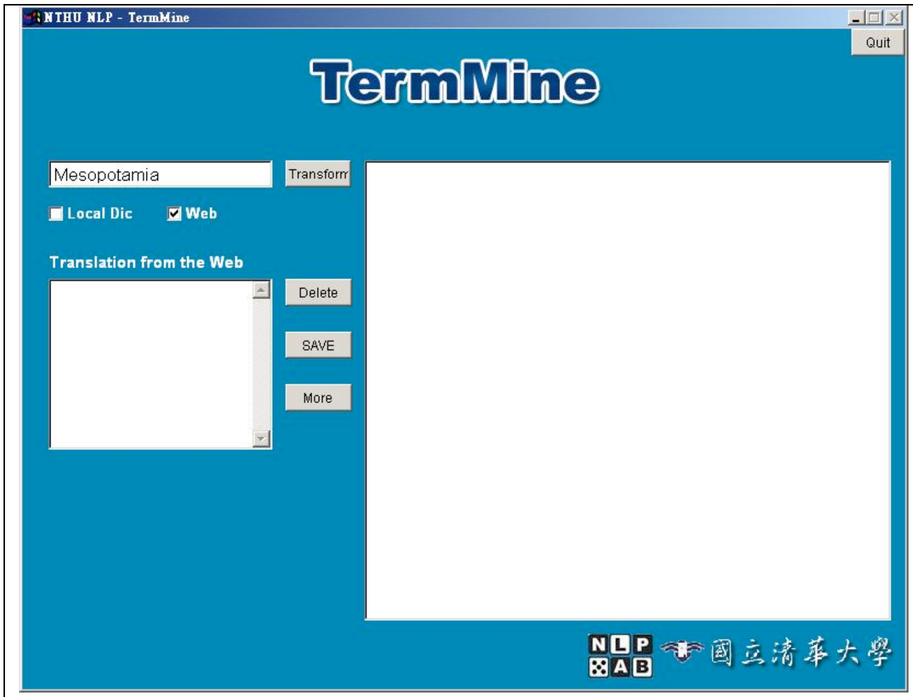
KL Kwok, P Deng, N Dinstl, HL Sun, W Xu, P Peng, *CHINET: a Chinese name finder system for document triage*. Proceedings of 2005 International Conference on Intelligence, 2005.

T. Lin, J.C. Wu, and J. S. Chang. (2004). *Extraction of Name and Transliteration in Monolingual and Parallel Corpora*. In Proc. of AMTA 2004, pp.177-186.

WH Lu, LF Chien, HJ Lee. *Anchor text mining for translation of Web queries: A transitive translation approach*. ACM Transactions on Information Systems (TOIS), 2004.

M. Nagata, T. Saito, and K. Suzuki. (2001). *Using the Web as a bilingual dictionary*. In Proc. of ACL 2001 DD-MT Workshop, pp.95-102.

Demonstration Script

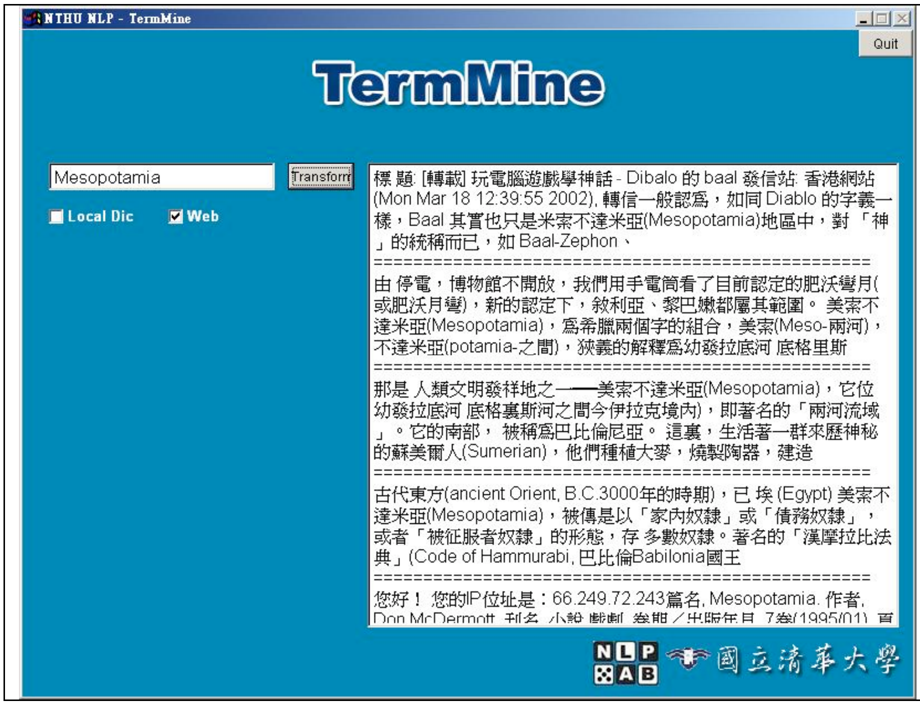


Mesopotamia

□ Local Dic ☑ Web

Translation from the Web

The system runs under Microsoft Windows as a local application program. It opens up a form and accepts a name for translation. The user can choose to search on the Web or to look up in the local dictionary for results previously obtained and edited.



Mesopotamia

□ Local Dic ☑ Web

標 題: [轉載] 玩電腦遊戲學神話 - Dibalo 的 baal 發信站: 香港網站 (Mon Mar 18 12:39:55 2002), 轉信一般認爲，如同 Diablo 的字義一樣，Baal 其實也只是米索不達米亞(Mesopotamia)地區中，對 「神 」的統稱而已，如 Baal-Zephon、

==

由 停電，博物館不開放，我們用手電筒看了目前認定的肥沃彎月(或肥沃月彎)，新的認定下，敘利亞、黎巴嫩都屬其範圍。 美索不達米亞(Mesopotamia)，爲希臘兩個字的組合，美索(Meso-兩河)，不達米亞(potamia-之間)，狹義的解釋爲幼發拉底河 底格里斯

==

那是 人類文明發祥地之一──美索不達米亞(Mesopotamia)，它位幼發拉底河 底格裏斯河之間今伊拉克境內)，即著名的「兩河流域」。它的南部， 被稱爲巴比倫尼亞。 這裏，生活著一群來歷神秘的蘇美爾人(Sumerian)，他們種植大麥，燒製陶器，建造

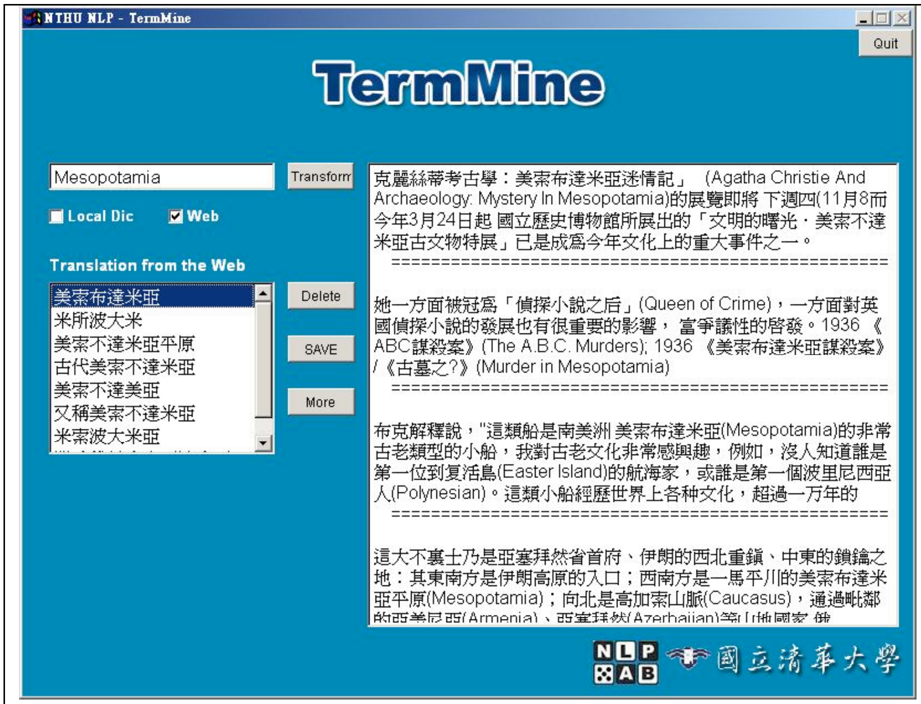
==

古代東方(ancient Orient, B.C.3000年的時期)，已 埃 (Egypt) 美索不達米亞(Mesopotamia)，被傳是以「家內奴隸」或「債務奴隸」，或者「被征服者奴隸」的形態，存 多數奴隸。著名的「漢摩拉比法典」(Code of Hammurabi, 巴比倫Babilonia國王

==

您好！ 您的IP位址是：66.249.72.243篇名. Mesopotamia. 作者. Don McDermott. 刊名. 小說 戲劇. 卷期／出版年月. 7卷(1995/01). 頁

The system now expanded the query and sending them to a search engine to retrieve snippets containing transliterations.



Mesopotamia

□ Local Dic ☑ Web

Translation from the Web

美索布達米亞
米所波大米
美索不達米亞平原
古代美索不達米亞
美索不達美亞
又稱美索不達米亞
米索波大米亞

克麗絲蒂考古學：美索布達米亞迷情記」 (Agatha Christie And Archaeology: Mystery In Mesopotamia)的展覽即將 下週四(11月8而今年3月24日起 國立歷史博物館所展出的「文明的曙光·美索不達米亞古文物特展」已是成爲今年文化上的重大事件之一。

==

她一方面被冠爲「偵探小說之后」(Queen of Crime)，一方面對英國偵探小說的發展也有很重要的影響， 富爭議性的啓發。1936 《ABC謀殺案》(The A.B.C. Murders); 1936 《美索布達米亞謀殺案》/《古墓之?》(Murder in Mesopotamia)

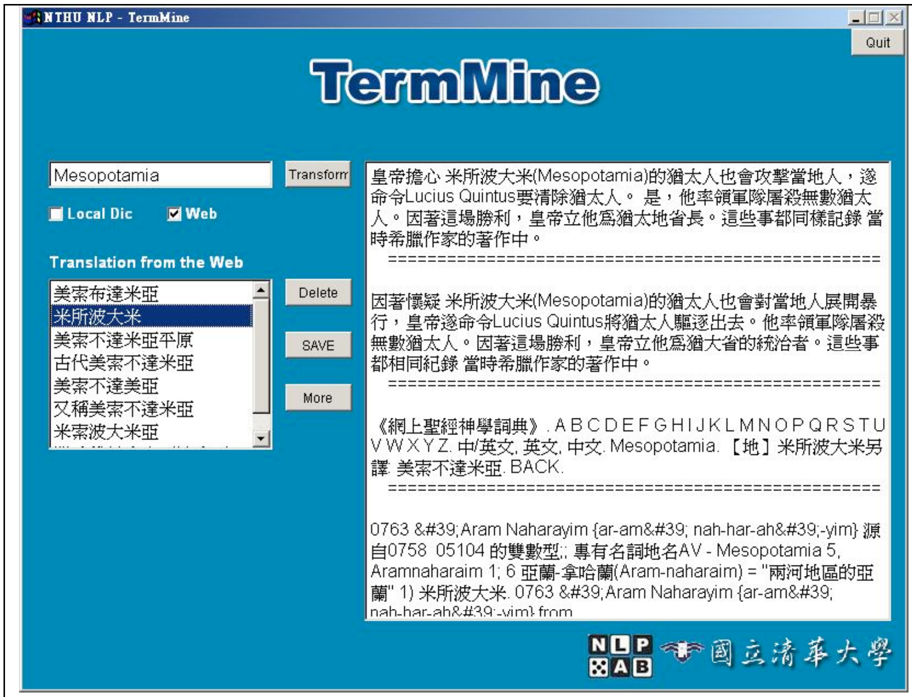
==

布克解釋說，"這類船是南美洲 美索布達米亞(Mesopotamia)的非常古老類型的小船，我對古老文化非常感興趣，例如，沒人知道誰是第一位到复活島(Easter Island)的航海家，或誰是第一個波里尼西亞人(Polynesian)。這類小船經歷世界上各种文化，超過一万年的

==

這大不裏士乃是亞塞拜然省首府、伊朗的西北重鎮、中東的鎖鑰之地：其東南方是伊朗高原的入口；西南方是一馬平川的美索布達米亞平原(Mesopotamia)；向北是高加索山脈(Caucasus)，通過毗鄰的亞美尼亞(Armenia)、亞塞拜然(Azerbaijan)等山地國家 俄

The system extracted transliterations from the snippets and showed the list of most likely answers.



Mesopotamia

□ Local Dic ☑ Web

Translation from the Web

美索布達米亞
米所波大米
美索不達米亞平原
古代美索不達米亞
美索不達美亞
又稱美索不達米亞
米索波大米亞

皇帝擔心 米所波大米(Mesopotamia)的猶太人也會攻擊當地人，遂命令Lucius Quintus要清除猶太人。 是，他率領軍隊屠殺無數猶太人。因著這場勝利，皇帝立他爲猶太地省長。這些事都同樣記錄 當時希臘作家的著作中。

==

因著懷疑 米所波大米(Mesopotamia)的猶太人也會對當地人展開暴行，皇帝遂命令Lucius Quintus將猶太人驅逐出去。他率領軍隊屠殺無數猶太人。因著這場勝利，皇帝立他爲猶大省的統治者。這些事都相同紀錄 當時希臘作家的著作中。

==

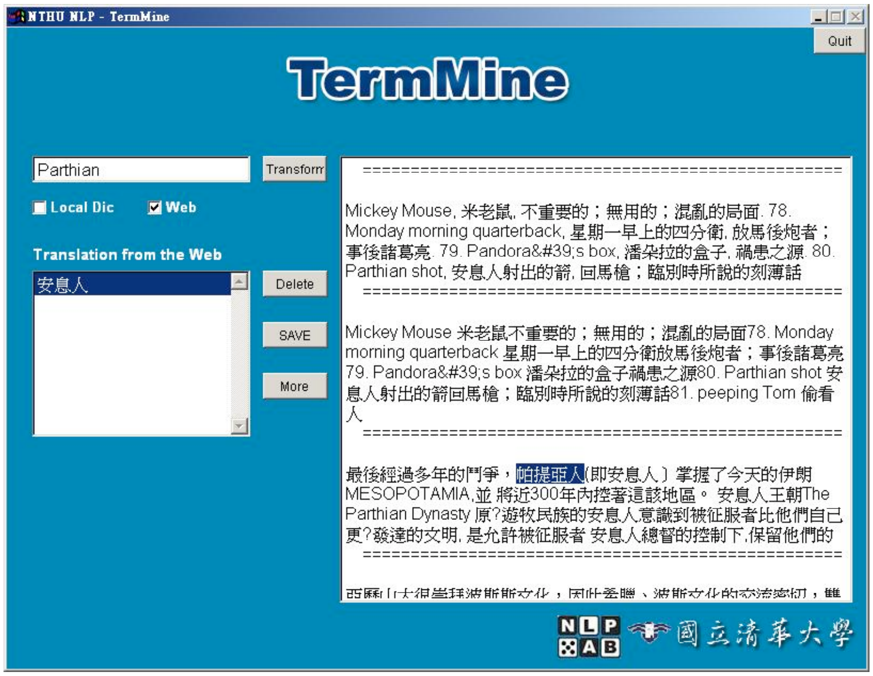
《網上聖經神學詞典》. A B C D E F G H I J K L M N O P Q R S T U V W X Y Z. 中/英文, 英文, 中文. Mesopotamia. 【地】米所波大米另譯: 美索不達米亞. BACK.

==

0763 &#39;Aram Naharayim {ar-am&#39; nah-har-ah&#39;-yim} 源自0758 05104 的雙數型;; 專有名詞地名AV - Mesopotamia 5, Aramnaharaim 1; 6 亞蘭-拿哈蘭(Aram-naharaim) = "兩河地區的亞蘭" 1) 米所波大米. 0763 &#39;Aram Naharayim {ar-am&#39; nah-har-ah&#39;-yim} from

The user can click each answer and view a list of snippets where the transliteration appears.

The user finds and selects “帕提亞人” in the snippets for “安息人”



Parthian

☐ Local Dic ☑ Web

安息人

==

Mickey Mouse, 米老鼠, 不重要的；無用的；混亂的局面. 78. Monday morning quarterback, 星期一早上的四分衛, 放馬後炮者；事後諸葛亮. 79. Pandora&#39;s box, 潘朵拉的盒子, 禍患之源. 80. Parthian shot, 安息人射出的箭, 回馬槍；臨別時所說的刻薄話

==

Mickey Mouse 米老鼠不重要的；無用的；混亂的局面78. Monday morning quarterback 星期一早上的四分衛放馬後炮者；事後諸葛亮79. Pandora&#39;s box 潘朵拉的盒子禍患之源80. Parthian shot 安息人射出的箭回馬槍；臨別時所說的刻薄話81. peeping Tom 偷看人

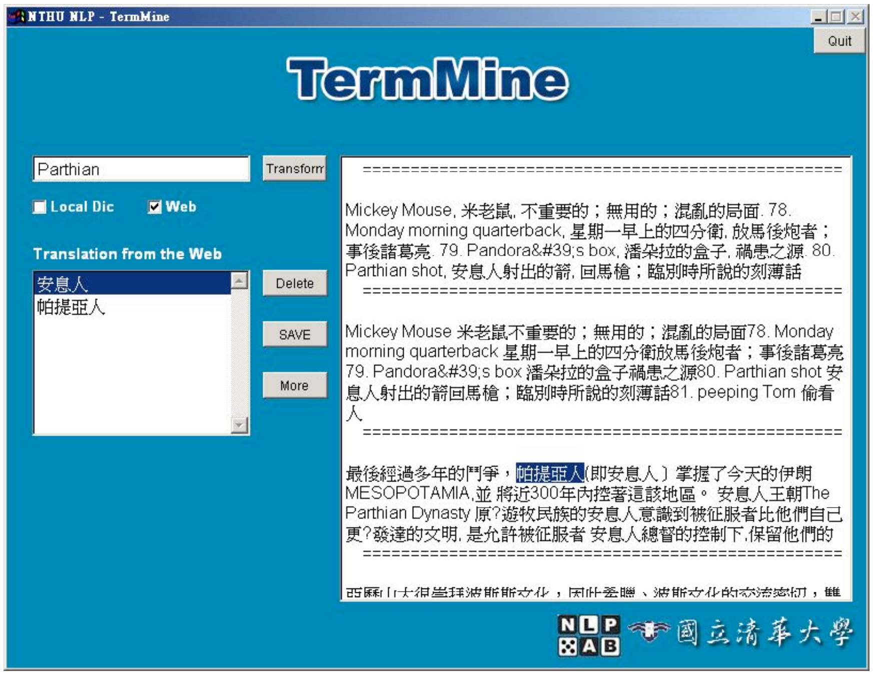
==

最後經過多年的鬥爭，帕提亞人(即安息人）掌握了今天的伊朗MESOPOTAMIA,並 將近300年內控著這該地區。 安息人王朝The Parthian Dynasty 原?遊牧民族的安息人意識到被征服者比他們自己更?發達的文明, 是允許被征服者 安息人總督的控制下,保留他們的

==

亞歷山大很崇拜波斯斯文化，因此希臘、波斯文化的交流密切，雖

The transliteration “帕提亞人” is added to the list



Parthian

☐ Local Dic ☑ Web

安息人
帕提亞人

==

Mickey Mouse, 米老鼠, 不重要的；無用的；混亂的局面. 78. Monday morning quarterback, 星期一早上的四分衛, 放馬後炮者；事後諸葛亮. 79. Pandora&#39;s box, 潘朵拉的盒子, 禍患之源. 80. Parthian shot, 安息人射出的箭, 回馬槍；臨別時所說的刻薄話

==

Mickey Mouse 米老鼠不重要的；無用的；混亂的局面78. Monday morning quarterback 星期一早上的四分衛放馬後炮者；事後諸葛亮79. Pandora&#39;s box 潘朵拉的盒子禍患之源80. Parthian shot 安息人射出的箭回馬槍；臨別時所說的刻薄話81. peeping Tom 偷看人

==

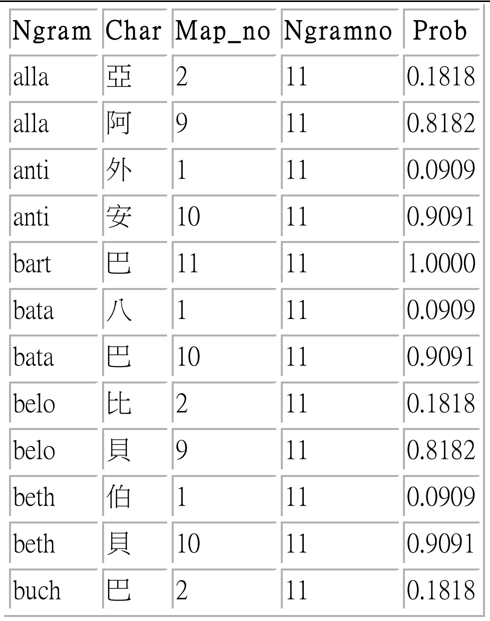
最後經過多年的鬥爭，帕提亞人(即安息人）掌握了今天的伊朗MESOPOTAMIA,並 將近300年內控著這該地區。 安息人王朝The Parthian Dynasty 原?遊牧民族的安息人意識到被征服者比他們自己更?發達的文明, 是允許被征服者 安息人總督的控制下,保留他們的

==

亞歷山大很崇拜波斯斯文化，因此希臘、波斯文化的交流密切，雖

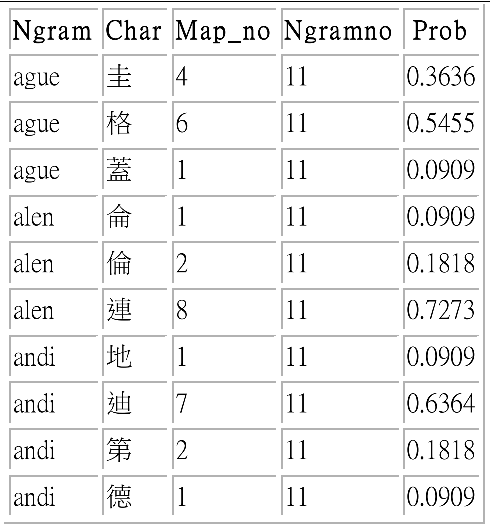
The user can do a number of things with the transliteration list:

1. View the snippets associated with each transliteration
2. Delete a transliteration
3. Save the list along with the snippet
4. Select a string in the snippets and add it to the answer list by clicking the right bottom



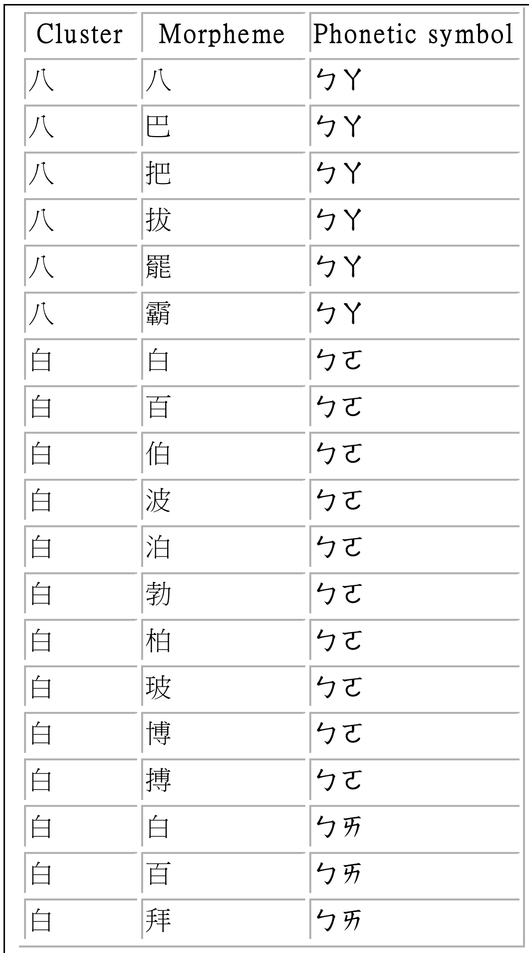
Ngram	Char	Map_no	Ngramno	Prob
alla	亞	2	11	0.1818
alla	阿	9	11	0.8182
anti	外	1	11	0.0909
anti	安	10	11	0.9091
bart	巴	11	11	1.0000
bata	八	1	11	0.0909
bata	巴	10	11	0.9091
belo	比	2	11	0.1818
belo	貝	9	11	0.8182
beth	伯	1	11	0.0909
beth	貝	10	11	0.9091
buch	巴	2	11	0.1818

This slide shows the relationships between the prefixes of names and translations.



Ngram	Char	Map_no	Ngramno	Prob
ague	圭	4	11	0.3636
ague	格	6	11	0.5455
ague	蓋	1	11	0.0909
alen	侖	1	11	0.0909
alen	倫	2	11	0.1818
alen	連	8	11	0.7273
andi	地	1	11	0.0909
andi	迪	7	11	0.6364
andi	第	2	11	0.1818
andi	德	1	11	0.0909

This slide shows the relationships between the postfixes of names and translations.

<table>
<tr>
<td>

<table>
<tr><th>Cluster</th><th>Morpheme</th><th>Phonetic symbol</th></tr>
<tr><td>八</td><td>八</td><td>ㄅㄚ</td></tr>
<tr><td>八</td><td>巴</td><td>ㄅㄚ</td></tr>
<tr><td>八</td><td>把</td><td>ㄅㄚ</td></tr>
<tr><td>八</td><td>拔</td><td>ㄅㄚ</td></tr>
<tr><td>八</td><td>罷</td><td>ㄅㄚ</td></tr>
<tr><td>八</td><td>霸</td><td>ㄅㄚ</td></tr>
<tr><td>白</td><td>白</td><td>ㄅㄛ</td></tr>
<tr><td>白</td><td>百</td><td>ㄅㄛ</td></tr>
<tr><td>白</td><td>伯</td><td>ㄅㄛ</td></tr>
<tr><td>白</td><td>波</td><td>ㄅㄛ</td></tr>
<tr><td>白</td><td>泊</td><td>ㄅㄛ</td></tr>
<tr><td>白</td><td>勃</td><td>ㄅㄛ</td></tr>
<tr><td>白</td><td>柏</td><td>ㄅㄛ</td></tr>
<tr><td>白</td><td>玻</td><td>ㄅㄛ</td></tr>
<tr><td>白</td><td>博</td><td>ㄅㄛ</td></tr>
<tr><td>白</td><td>搏</td><td>ㄅㄛ</td></tr>
<tr><td>白</td><td>白</td><td>ㄅㄞ</td></tr>
<tr><td>白</td><td>百</td><td>ㄅㄞ</td></tr>
<tr><td>白</td><td>拜</td><td>ㄅㄞ</td></tr>
</table>
</td>
<td>The performance can improved by clustering transliteration user can click each answer and view a list of snippets where the transliteration appears.</td>
</tr>
</table>