

通往過去的鑰匙—古文解碼

Shou-de Lin

古文解碼，是一個橫跨自然、社會與應用科學領域的題目。解碼的本身不但涉及歷史及語言學，也常需藉助數學和統計的概念；電腦的掘起更為現代的解碼者提供了一個新的方向。本文嘗試從跨領域的角度，來探討古文解碼的過去、現在與未來。

源於古埃及的謎題

We will discover the nature of our particular genius when we stop trying to conform to our own or to other peoples' models

Shakti Gawain

古埃及象形文是世上最古老的文字之一，又名聖書體，代表獻給神明的文字。它由一群容易分辨的圖像組成（如圖一），經常被書寫成碑文或是刻在建築物上。



圖一：古埃及象形文（有如鳥，山，河的符號）

這個古老的文字直到十九世紀時才被完全破解，後世的人們將其歸功於英國物理學家楊格（Thomas Young）跟法國語言學家商博良 (Jean-François Champollion)，還有一個叫做羅塞塔的石碑 (Rosetta Stone)。

故事得從羅賽塔這座號稱世上最有名的石頭說起。1799 年法國遠征軍在埃及一個名為羅賽塔的地方，發現了這個重達七百多公斤的黑色大石，上面刻滿密密麻麻無人能理解的文字。當時的指揮官意識到這個石碑的重要性，於是將之拓印交與法國的埃及研究所收藏。這塊石頭的特別之處在於其上刻有三種不同的文字，最下方的是古希臘文，是三種文體中唯一在當時已被瞭解的文字；中間刻的是古埃及蝌蚪文，據信是平民所用的通俗體；而最上端部分毀損的即是埃及象形文（如圖二）。因為埃及曾經臣服於亞歷山大帝國之下，所以這三種文字極可能互為譯文，也因此羅賽塔石碑被認為是譯解古埃及文的關鍵物。



圖二：羅賽塔石碑

楊格與喬博良不約而同對羅賽塔石產生興趣。由於他們分屬於兩個當時敵對的國家，埃及文的解碼也形成英法之間的另一場競爭。在法國戰敗後，英國欲接管羅賽塔石，但是遭到法國學者反對，宣稱不惜玉「石」俱焚。最後達成的協議是法國人可以留下拓本，羅賽塔石則運到英國，成為大英博物館埃及分館的鎮館之寶。

羅賽塔石的所有權之爭落幕，其實正是楊格與喬博良競爭的開始。他們兩人雖然來自不同的背景，此時卻擁有對等的競爭條件，也就是羅賽塔石碑上面的文字。首先打破僵局的是楊格，他發現在古埃及文中有一些符號每每被放在一個橢圓框中，於是他假設橢圓框是對於君王的敬稱，裡面寫的即是埃及君主的名字。利用這個線索，再加上本身精通古希臘文，楊格得以一步步由古希臘文慢慢解譯出古埃及蝌蚪文。最後他在西元 1814 年的論文中發表了古埃及蝌蚪文的讀音與文法規則的解譯。楊格也對象形文提出他的看法，他認為那些外來的君王名稱如托勒密等，是無法以表意的象形文來表達的，因為造象形字時根本就不知道那些外來的名字，所以只能用拼音。對照之下他發現象形文中一類漩渦狀的圖像可能就代表著音標，繼而猜測出表示托勒密的符號。

楊格在這場英法解碼戰的第一回合獲得勝利。不過喬博良並未放棄，這個 19 歲就在大學教書的語言天才，知道象形文的解碼將會是長期戰。跟其他的解碼者比較，他擁有一個重要的優勢：便是他精通十多種古語，尤其是一種和古埃及語系非常接近的科普特語。他首先計算羅賽塔石上面的符號，發現希臘文中總共有四百多字，但是碑上的象形文符號卻有三倍之多。再加上受到楊格找出托勒密拼音符號過程的啟發，他開始懷疑一個非常根本的問題：古埃及的象形文真的是「畫成其物，隨體詰詘」的象形文嗎？有沒有可能不只是如楊格說的外來語，其實所有的看似有

意義的埃及象形符號其實就像是注音符號一般，也代表著某種讀音？這個想法在現在看起來也許不怎麼偉大，在當時卻是劃時代的想法，因為它跟大師級人物荷拉被隆 (Horapollo) 所提的觀念背道而馳，也跟一般學者所認知的象形文（如中國象形文字）大相逕庭。

商博良默默將這個想法放在心上。直至有一天，他在另一塊同時刻有埃及象形文和古希臘文的碑上，看到用希臘文寫的托勒密王和王后克麗奧佩脫拉的名字。他發現在象形文中有一個字，構成它的第二、第四和第五個符號，和已知的「托勒密」的象形文在第四、第三和第一個符號相符。這時他突然想起，克麗奧佩脫拉與托勒密的讀音在相關的音節位置正好也有這種相似性。在那瞬間，他清楚理解到他已經找到那把可解開整個謎團的關鍵鑰匙。從這幾個字開始，商博良利用其他埃及君王的名字，逐個音節將埃及象形文重建成一個拼音文體，並在 1822 年發表了一篇劃時代的論文，將古埃及象形文解碼的結果公諸於世。

能夠跳脫過去觀念的桎梏，思索埃及象形文的音標特性，是商博良能夠成功解譯古埃及文的重要關鍵。而藉由他對埃及象形文的解譯，日後的埃及學學者得以正確地研究古埃及文明，也為他贏得了埃及學之父的名號。

伏尼契手稿 (Voynich Manuscript)

“What is important is what you believe.”

Quote from *Da Vinci Code*

古埃及文的成功解密，給後來的解碼者很大的鼓舞。之後的幾個世紀人們也勢如破竹地解出楔形文字、線形文 B，和馬雅象形符號等文字。古文解碼似乎不再那麼遙不可及。

直到伏尼契手稿的出現。

這個號稱史上最神秘的手稿，從 20 世紀初現世至今，可能創下一項世界紀錄：就是被宣告破解次數最多的謎題。

能夠被宣告破解這麼多次，表示從來沒有任何一種解譯是被公認正確的。

伏尼契手稿是一個厚達兩百多頁、圖文並茂的手抄稿¹。裡頭的插圖大部分是無法辨識的植物與奇怪的天文構圖；所用的文字則是無人知曉的符號。沒有人知道它的作者是誰，最早的紀錄顯示它是十六世紀魯道夫大帝用高價購得，之後幾經易手，終在 1912 年被舊書商伏尼契買下並公諸於世。手稿中的文字有許多有趣的現象，例如它的單字頻率雖然跟其他語言一樣都符合極夫定律（zipf's law），但是它比一般的歐洲語言有更多連續重複出現的字。此外，在伏尼契文中，字母長度的分佈跟一般拼音語言有明顯的不同。最常見的長度是五或六個字母，並不像英文中許多較常見的字都只有很短的字母（如 a, is）；但是它的熵數（entropy）² 卻又跟英文及拉丁文類似。

這些分析的結果顯示伏尼契文雖然某種程度像是一般語言，但又有其特異之處，這個模糊地帶的存在也引起許多截然不同的解碼觀點。不時有傳聞聲稱「解開」了伏尼契。

最早做此宣稱的是 1912 年一名叫做紐伯德的教授。他稱伏尼契文本身其實沒有意義，但是把字母放大來看之後，就會發現其中有數個微小的標誌。這些標記是古希臘速記的寫法，內容是十三世紀的哲學家培根描述他如何發明顯微鏡。這個論調在短時間內就被推翻，因為在驗證他的結果時，人們發現他對相同的記號常有不同的翻譯，而且原來的字母順序也常被任意排列以求得有意義的拉丁文。

¹ 請參照 <http://www.voynich.net/images.html>。

² 如何計算？

另有一派人士，包括美國國安局的解碼專家，認為伏尼契文是當時某種歐洲語言，經過加密後（如字母取代、母音去除、斷字等等）轉換成另一種讓人看不懂的文字。這個理論最大的問題在於：經過複雜的加密轉換後，一些自然語言常見的統計現象如極夫定律通常會被破壞，不太可能仍保留在轉譯後的語言中。

還有一些學者認為伏尼契文之所以迄今未解碼成功，是因為它本身並非「某種」語言的編碼，而是一種混合體語言或外來語。這當中有一派說法認為這是亞洲來的語言，因為在中文或是越南文裡出現疊字的頻率較高。此外，它的天文圖像裡將一年分成 24 個單位，正好與中國農曆的 24 節氣相符；而它的黃道十二宮始於雙魚座，也正好對應農曆的正月。但是反對人士質疑伏尼契的圖畫中從來沒出現過任何跟亞洲文化相關的符號。

其實，上述的解譯基本上都犯了一個很大的通病，就是「自由心證」。人們先相信了自己的假設，然後從這角度來看伏尼契，再加上一些人為雕琢解釋，自然就認為自己解開了迷團。有些學者更諷刺地指出，很多為解譯伏尼契所創造出的方法，把它們用在一個三歲小孩隨便畫的圖上，也可以得到有意義的解譯。

既然從語言學和密碼學的角度破解伏尼契手稿進展都不大，另外有一些學者開始思考另一個可能：會不會伏尼契文根本沒有意義，只是有人為了惡作劇或斂財編造出來的？將這個說法集大成的是英國的心理學家魯格（Gordon Rugg）。他發現西元 1550 年發明的一種加密工具「卡丹格」（Cardan grille），利用這個中間有小空格的版子蓋在文件上，抓出小段字母拼在一起，就可以做出一種看起來跟伏尼契文非常相似的文體。這個發現在 2003 年發表時被媒體斷章取義，紛紛報導伏尼契終於被破解，而且是個騙局。這個結論傷了很多伏尼契迷的心，就好像告訴幽浮迷墜落在羅茲威爾的其實不過是一個氣球。但是這個結論還言之過早，因為魯格所謂「很像」伏尼契的文字，只有外觀類似而已，其統計與機率分布都跟真正的伏尼契文並不符合。連魯格自己也承認：「證明當時的人有能力捏造出伏尼契文，並不表示伏尼契文本身是捏造的。」

所以，伏尼契手稿至今仍是各執一詞，信者恆信的謎。那些不願輕信這是個騙局的人，仍不停地為它找尋一個浪漫的詮釋。



伏尼契手稿的一頁

全域翻譯 (Universal Translation)

A discovery is said to be an accident meeting a prepared mind.

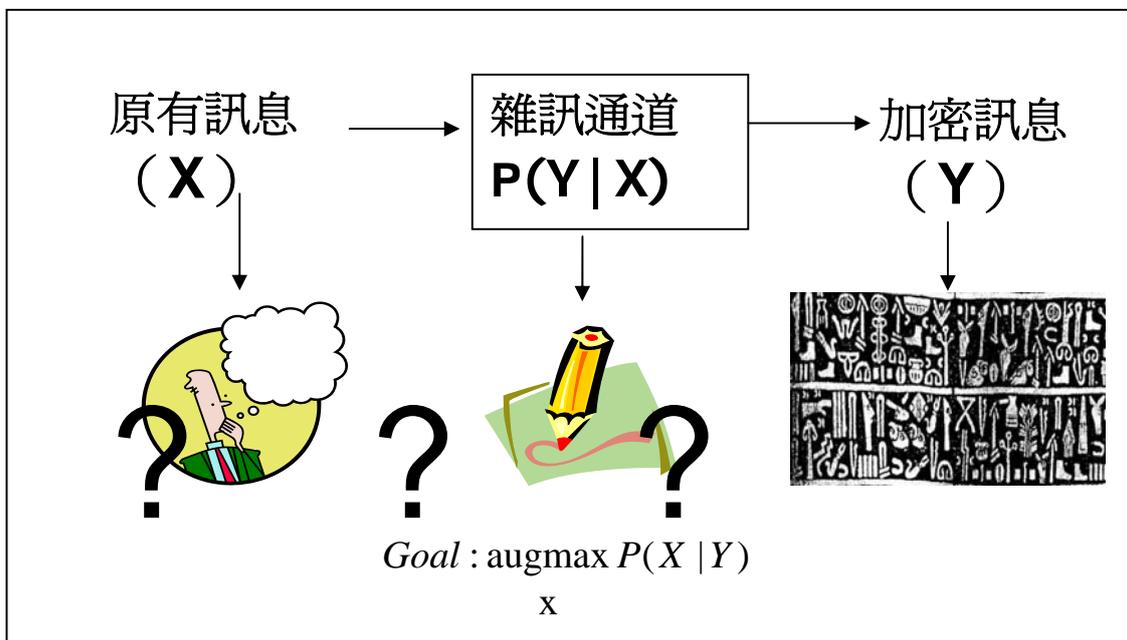
Szent-Gyorgyi

二十世紀末的古文解碼，因為電腦的介入，進入一個新的世代。南加州大學資訊科學院的奈特(Kevin Knight) 博士與他的學生，也是全球電腦翻譯研究的指標性團隊，正用另一種角度看待古文解碼的問題。他們思索古文解密與自動翻譯的相似性，目的是找出背後共通的規則，利用電腦的運算能力來完成所謂「全域翻譯」。也就是對任何一種語言，都可以利用人工智慧的軟體做某種程度的解譯成爲另一種語言。這其中，當然也包含了把古文翻譯成現代文。奈特曾利用語言模組（language model）的概念計算伏尼契文中各個詞彙的相關性，進而駁斥某些研究者提出伏尼契文是烏克蘭語的假說，因為這兩種語言產生的語言模組差異太大。這個電腦在考據學上的小小勝利，讓他們進一步去思索人工智慧是否可以進一步被應用在古文解密上面。

古文解譯可以分成幾大類。第一類是以未知的符號撰寫已知的語言，第二類是用已知的符號撰寫未知的語言，最難的則是如喬博良解古埃及文般，面對的是未知的符號撰寫未知的語言。而古文解密也可以分成幾個「次級任務」。最終目的雖是瞭解它的意義，近中程目標則可以先設定在瞭解古文的閱讀順序，以及這些古文的讀音與文法。而這些任務都可以藉由電腦的幫助達成。

他們的基本假設是電腦可以結合訊息理論之父夏農（[Claude Shannon](#)）在 20 世紀中提出的「雜訊通道模型」（noisy channel model）與 1977 年丹普司特（Dempster）的「預估最佳化」（Expectation-Maximization，簡稱 EM）的演算法，來進行古文翻譯。發現這個方法的靈感，其實是由觀察人類解碼過程產生：通常人類爲古文解碼時，第一步是先建立一些假設，然後利用它們去做另外一些預

測，之後再從比較合理的預測結果去推出更進一步的結果，如此一步步推進；如果中間碰到瓶頸，就放棄或修改某些先前的推測。例如商博良先猜測托勒密符號的位置，再以之推論出克麗奧佩脫拉的譯法，再利用這兩個譯法漸漸推衍出剩下符號的讀音，這樣反覆推衍了四年才解開埃及文的奧密。奈特發現這樣的辯證過程可以用雜訊通道模型併以預估最佳化來模擬。更精確的說，上述人類解碼的過程，其實是一種把 EM 演算法運用在雜訊通道模型的簡化特例。



上圖簡單介紹電腦解碼的概念：古文其實可以被模擬成是雜訊通道的產物：過去的人在寫下古文之前，心中一定已經有想表達的事（圖中的 X），書寫的本身，從訊息理論的角度來看，可以當成把雜訊加入原有訊息的過程（用條件機率來看就是 $P(Y|X)$ ），最後我們所看到的古文，即是加入雜訊後的加密結果（Y）。不幸的是數千年過後，古人的智慧、語言以及他們書寫的方式都已失傳，也就是說 X 跟 $P(Y|X)$ 都變成了未知，唯一能知道的是我們現在看到的古文

Y。古文解譯則可以當成上述過程的反推。身為後人的我們，希望能夠藉由 Y 來找出原來他們的語言書寫方式以及想表達的意念。從機率上來說，上述的解碼可以用 $\mathop{\text{arg max}}_x P(X|Y)$ 這個式子來表示，也就是說解碼的目的是在觀測到加密訊息 Y 的條件之下找到一個機率最大的解譯 X。

很多古文的解碼過程其實都有著雜訊通道模型的影子。回頭來看看楊格的兩個重要猜測：外來君主是以橢圓框及拼音書寫。對照到圖 5 即是對於雜訊通道的一種假設。而喬博良進而利用這種假設來解出古埃及文的過程，就好像是利用雜訊通道的資訊以及古文 Y 來還原 X。

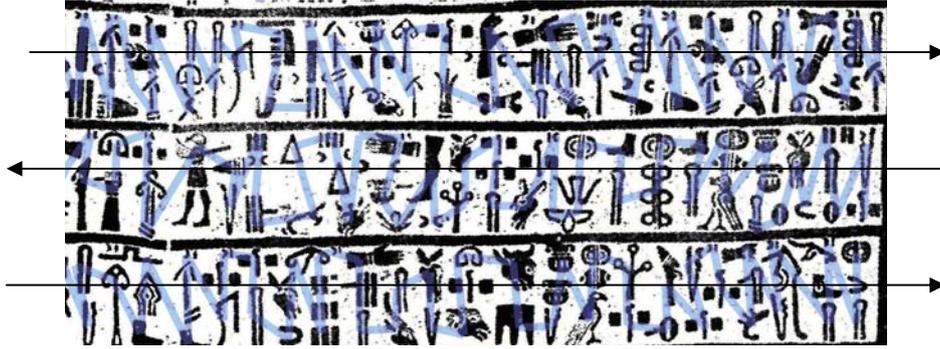
雜訊通道模型告訴我們古文解譯在數學上可以轉換成一個條件機率最佳化的問題。但是它並沒有告訴我們如何能夠找到這個最有可能 X。這個問題其實可以利用被通稱為 EM 的「預估最佳化」演算法解決。EM 分成兩個獨立部分，第一個部分是預估 (E-step) 第二個部分是最佳化 (M-step)。E-step 的精神在於利用已觀測到的結果經過既有的理論來猜測其他未知的部分，對應於上圖，即是利用雜訊通道及 Y 來預測 X。而 M-step 主要的目的是在利用 E-step 預測出來的結果對於既有的理論做修正及最佳化，也就是利用被預測出來的 X 來修正原先代表雜訊通道的參數。在 EM 演算法中，E-step 跟 M-step 必須要輪流地一直重複修正直到結果趨於穩定。這樣的過程非常系統化，可以輕易利用遞迴的觀念寫成電腦程式。

這個能夠自動改進既有理論的方法，其實跟從前古文解碼的精神若合符節。通常人類為古文解碼時，第一步是先建立在一些已知的假設上、利用它們對於未知的符號做出一些預測(M-step)，之後再從比較合理的預測結果修正原先的一些假說

(E-step)，再進一步推論新的結果 (M-step) ...如此循環不斷。這樣反覆推衍得花費人類很長的時間（比如喬博良花了至少四年才解開埃及文的奧秘），但是如果用電腦來進行這個步驟，可能只需要數天的時間。

他們在 2006 年發表的一篇論文中，說明電腦如何利用上述方法破解一種名為魯汶古文 (Luwian) 的閱讀順序。這種象形古文並非如一般文字是由上往下或是由左往右排列整齊，而是看似無規律的二維陣列。而程式執行結果顯示，魯汶文在局部大部分是由上往下，但各列之間方向相反（如下圖）。為了檢驗這個結果是否正確，他們隨機找了一些人，請他們用自己會的語言，任意把一段隨機選取的文章寫成二維分佈的文本，然後把這些寫出來的文本輸入給破解魯汶文的電腦程式處理。結果發現不管這些人用的是甚麼語言（包括英文、阿拉伯文、西班牙文、拉丁文，甚至還有人故意用電腦程式語言 Java），也不管他們用甚麼方法寫出這個二維文本，這個程式都能夠還原出可被理解的線性文章。他們也發展出了一種可以讀出未知語言的系統，對於即使尚無法瞭解意義的古文，程式仍可將子音母音分開，再利用人類發聲的習性，讓電腦能「讀」出這些文字最有可能的讀音。

這些研究成果顯示出除了知識擷取（如搜尋引擎）以及自動學習（如語音辨識）外，電腦也可以被賦予更深一層的智慧，嘗試處理如古文解碼般人類一時無解的問題，這便是被稱為「機器發現」（machine discovery）的新興領域。更有其他學者利用類似的觀念去研究海豚發出的聲音甚至來自外太空的訊號，希望有一天也能夠解譯這些不屬於人類的語言。



魯汶古文的線性讀法

古文，是每個人都可以挑戰的謎題

The cure for boredom is curiosity. There is no cure for curiosity.

Dorothy Parker

從古埃及文到近代的小說達文西密碼的風行，處處可見人類對歷史謎題懷抱的好奇心與浪漫憧憬。古文解碼最迷人之處，就在於它不是任何學科的專利，如同楊格、商博良等人分別擁有迥異的背景。他們唯一相同的，是一份對未知的好奇心與不服輸的毅力。

所以，麥克阿瑟或許不會介意自己的名言如此被轉譯：

「古文不死，只是漸漸被解譯。」