Ấn bản Thư viện AI tiếng Việt

AI giải mã chữ viết cổ đại

Ghi chép cổ được AI hồi sinh

Luật sư Kim Kyung-jin

kimkj.com · Thư viện AI · 2026

AI giải mã chữ viết cổ đại

Ghi chép cổ được AI hồi sinh

Luật sư Kim Kyung-jin

Những cuộn giấy bị lửa đốt thành than, những mảnh gỗ bị chôn trong bùn đất, và những bảng đất sét vỡ vụn nằm rải rác từng là những vật không thể đọc trong một thời gian dài. Cuốn sách này lần theo mười hai sự kiện đã phá vỡ sự im lặng ấy. Nội dung đi từ mở cuộn ảo, tức đọc các cuộn giấy Herculaneum bị chôn dưới tro núi lửa mà không cần mở ra, sách Lê-vi tại En-Gedi được khôi phục bằng X-ray, việc giám định Cuộn Isaiah bằng nét chữ để phân biệt số người chép, chất kết dính số nối lại hàng chục nghìn mảnh vỡ, việc tái dựng Sử thi Gilgamesh, dịch sổ sách tiếng Akkad, đóng tập ảo 3D cho giáp cốt văn, đọc chữ thảo trên thẻ gỗ Silla, ngôn ngữ học tính toán xử lý chữ tuyến tính A, chữ viết Ấn Hà và rongorongo, mô hình ngôn ngữ điền phần thiếu trong bia khắc Hy Lạp bị vỡ, cho đến kho lưu trữ đa phổ vẫn còn lại dù bản gốc biến mất. Mỗi chương được cấu thành theo thứ tự hồ sơ vụ việc, cuộc truy tìm của đội điều tra khoa học, và sự thật đã được làm sáng tỏ. Các thuật ngữ được giải thích để người không có kiến thức nền vẫn có thể đọc. Mỗi tiểu chương có kèm tài liệu tham khảo đã được xác nhận. Phụ lục tập hợp các cổng dữ liệu công khai mà độc giả có thể tự mở tại nhà. Cuốn sách này được viết dựa trên 251 tài liệu đã được tập hợp trong ghi chú NotebookLM “Bộ sưu tập nghiên cứu giải mã ngôn ngữ cổ đại và cổ văn hiến dựa trên AI (2024-2026)”.

Kho lưu trữ tài liệu nghiên cứu công khai

Cuốn sách này được sắp xếp và biên soạn bằng AI. Đây là kết quả thu thập và tổng hợp tài liệu bằng các ngôn ngữ chính trên thế giới. NotebookLM không giới hạn ngôn ngữ đặt câu hỏi theo ngôn ngữ viết của tài liệu. Độc giả có thể đặt câu hỏi bằng ngôn ngữ của mình trong cùng kho lưu trữ tài liệu nghiên cứu đã được dùng để chuẩn bị cuốn sách này. Điều đó vẫn đúng dù một số tài liệu bằng tiếng Hàn, tiếng Anh hoặc ngôn ngữ khác. Nếu mở phần cài đặt trong NotebookLM và đặt ngôn ngữ đầu ra thành ngôn ngữ mong muốn, câu trả lời và kết quả tạo sinh sẽ xuất hiện bằng ngôn ngữ đó.

Kho lưu trữ NotebookLM công khai đã được dùng khi chuẩn bị cuốn sách này (Bộ sưu tập nghiên cứu giải mã ngôn ngữ cổ đại và cổ văn hiến dựa trên AI 2024-2026, 251 tài liệu):
https://notebook.google.com/notebook/cf32a6a1-66cc-473d-99d8-ed5a51596eb3

Bản gốc tiếng Hàn của cuốn sách này nằm trong AI Library của kimkj.com:
https://kimkj.com/ai-library/?mod=document&uid=12641

Mục lục

Nhấn vào tiêu đề từng chương để chuyển đến phần nội dung tương ứng. Nếu muốn đọc từ đầu, hãy chọn Chương 1.

Chương 1. Thư viện của nhà triết học bị tro núi lửa nuốt chửng: cuộn giấy bị carbon hóa ở Herculaneum

1.1. Hồ sơ vụ việc

Ngày núi lửa phủ lên thư viện

Năm 79 sau Công nguyên, núi lửa Vesuvius ở miền nam nước Ý phun trào. Núi lửa phun khí nóng và tro núi lửa ra cùng lúc. Dưới chân núi, bên bờ biển, có một thành phố tên là Herculaneum. Đó là thành phố nghỉ dưỡng nơi giới giàu có La Mã đến nghỉ hè. Thành phố này cũng bị chôn vùi dưới đất chỉ trong một ngày[1].

Bên ngoài thành phố có một dinh thự lớn. Ngày nay, các học giả gọi nơi này là Biệt thự Giấy cói. Chủ nhà có lẽ là Lucius Calpurnius Piso Caesoninus. Ông là cha vợ của Julius Caesar. Nhà triết học Hy Lạp Philodemus từng lui tới ngôi nhà này. Philodemus là học giả thuộc trường phái Epicurus. Một phía của dinh thự có thư viện cất giữ các cuộn giấy. Vì tác phẩm của ông xuất hiện nhiều khác thường, nơi này được xem là thư viện gắn với ông[1].

Giấy cói là loại giấy của thế giới Địa Trung Hải cổ đại. Nó được làm từ loài cây mọc bên sông Nile ở Ai Cập. Người ta thái mỏng phần ruột thân cây rồi xếp ngang thành dải dài. Trên đó, họ lại đặt thêm một lớp theo chiều dọc. Khi rưới nước và ép xuống, nhựa cây đóng vai trò như keo dán. Khi khô, nó trở thành một tờ giấy. Vì vậy, trên bề mặt giấy cói còn lại các thớ đan ngang dọc. Những thớ này về sau trở thành manh mối quan trọng[7].

Hôm đó, khí nóng tràn vào thư viện. Các cuộn giấy bị nung chín nguyên vẹn mà không có ngọn lửa. Giấy cói từng đóng vai trò như giấy đã biến thành than đen. Việc cháy thành than như vậy được gọi là carbon hóa. Các cuộn giấy co lại và bề mặt phồng lên[2]. Phía trên chúng, trầm tích núi lửa chất cao hơn 20 mét. Trầm tích núi lửa là lớp tro núi lửa và bột đá tích lại rồi đông cứng[1].

Khi bị chôn sâu dưới đất, không khí không chạm tới chúng. Khi không khí không chạm tới, sợi thực vật không dễ mục nát. Thảm họa đã đóng vai trò như chiếc nắp bảo vệ di vật. Các cuộn giấy đã trụ lại như vậy suốt 2.000 năm[2].

Những thanh đen được đưa ra từ đường hầm dưới đất

Năm 1738, hoàng gia Napoli bắt đầu đào khu vực này. Năm 1750, một công nhân đào giếng bắt gặp sàn khảm nhiều màu. Vị trí của dinh thự đã lộ ra như vậy. Karl Weber phụ trách khai quật. Ông là kỹ sư quân sự người Thụy Sĩ[1].

Weber đào một đường hầm thẳng xuống trong lớp trầm tích núi lửa đã đông cứng. Loại hầm này được gọi là giếng đứng. Sau đó, ông đào các đường hầm hẹp như mê cung. Bên trong đường hầm tối và không khí xấu. Nguy cơ tường sập luôn hiện hữu. Ngay trong điều kiện đó, Weber vẫn vẽ sơ đồ mặt bằng của dinh thự rất tỉ mỉ. Nhờ bản đồ ấy, người ta biết vật gì được tìm thấy ở phòng nào[1].

Từ năm 1752, công nhân bắt đầu đưa ra những vật giống như các thanh đen. Ban đầu, họ tưởng đó là những cục than. Sau đó mới biết đó là các cuộn giấy cói. Số cuộn giấy lấy ra từ dinh thự này được biết là khoảng 1.100 cuộn. Đây được biết là trường hợp duy nhất mà một thư viện của thế giới cổ đại còn lại gần như nguyên vẹn[1].

Trong thư viện này, tác phẩm của Philodemus nhiều khác thường. Bản hoàn chỉnh, bản sao và bản nháp được đặt cùng nhau. Các học giả cho rằng ông đã viết và chỉnh lý tác phẩm trong ngôi nhà này[1]. Trong các cuộn giấy được mở ra rất khó nhọc có những bài viết về triết học, tu từ học và đạo đức. Tu từ học là

ngành nghiên cứu cách thuyết phục người khác bằng lời nói và chữ viết. Không ai biết các cuộn chưa mở chứa nội dung gì[2].

Những cuộn giấy bị phá hỏng khi cố mở ra

Các học giả muốn đọc chữ bên trong chúng. Vấn đề là các cuộn giấy sẽ vụn ra khi chạm vào[2]. Năm 1753, linh mục Antonio Piaggio chế tạo một chiếc máy. Đó là thiết bị dùng chỉ và quả nặng để kéo và mở cuộn giấy từng chút một[3]. Chiếc máy này cứu được vài cuộn, nhưng cũng làm rách các lớp bên trong[2].

Đến thế kỷ 19, nhà hóa học người Anh Humphry Davy tham gia. Ông dùng khí clo, hơi iốt và axit trên các cuộn giấy. Đó là cách dùng hóa chất để tách các lớp ra. Kết quả không tốt. Hóa chất làm giấy cói mềm đi và vỡ nát[3].

Phần lớn các nỗ lực mở bằng phương pháp vật lý đều làm hỏng di vật. Những mảnh vụn có chữ viết rơi vãi xuống sàn rồi biến mất. Cuối cùng, giới học thuật dừng việc thử mở cuộn. Hiện nay vẫn còn hàng trăm cuộn chưa được mở[2]. Những di vật này được chia ra bảo quản tại các cơ quan lưu giữ ở Napoli, Paris và Oxford[4].

Bức tường carbon chặn tia X

Các nhà khoa học tìm cách đọc mà không mở. Công cụ đầu tiên họ nghĩ tới là chụp cắt lớp vi tính. Công nghệ này thường được gọi là CT. CT là thiết bị đo sự khác biệt về mật độ bên trong vật thể mà không cần cắt vật thể. Vật chất có mật độ cao chặn nhiều tia X. Vì vậy, nó hiện sáng trên ảnh mặt cắt[5].

Mực có pha thành phần kim loại sáng trắng trên ảnh CT. Ở châu Âu thời trung cổ, mực chứa sắt được dùng rộng rãi. Cuộn giấy bị cháy tìm thấy ở En Gedi, Israel cũng tương tự. Di vật này là bản chép tiếng Hebrew của Israel cổ đại. Có vẻ mực có pha kim loại nên để lại dấu vết trên ảnh. Vì vậy, năm 2016, nó được đọc mà không cần mở[5].

Các thư lại ở Herculaneum có hoàn cảnh khác. Thư lại là người chép chữ bằng tay. Họ đốt gỗ lấy muội rồi trộn với nước và nhựa cây để làm mực đen. Thành phần chính của mực đen là carbon. Nhưng giấy cói bị nung bởi sức nóng núi lửa cũng đã trở thành carbon. Giấy và chữ đã trở thành cùng một vật chất[6].

Khi mật độ giống nhau, tia X đi qua hai nơi như nhau. Trên ảnh mặt cắt không xuất hiện đường ranh giới. Các học giả gọi bức tường này là vấn đề mực carbon. Năm 2015, nhóm nghiên cứu Ý và Pháp thử một phương pháp khác. Đó là kỹ thuật chụp tương phản pha. Kỹ thuật này dùng tính chất tia X hơi đổi hướng khi đi qua ranh giới vật chất. Dù mật độ giống nhau, tín hiệu vẫn khác ở mặt ranh giới. Bằng phương pháp này, nhóm nghiên cứu nắm được đường nét của vài chữ bên trong cuộn giấy đã cuộn lại. Nhưng vẫn chưa đủ để đọc thành câu[6]. Trong thời gian dài, người ta tin rằng không thể vượt qua bức tường này[2].

Tài liệu tham khảo

[1] Herculaneum Society. The Villa of the Papyri. University of Oxford. https://www.herculaneum.ox.ac.uk/index.php/the-villa-of-the-papyri/ (truy cập ngày 3 tháng 9 năm 2026)

[2] Vesuvius Challenge. FAQ. scrollprize.org. https://scrollprize.org/faq (truy cập ngày 3 tháng 9 năm 2026)

[3] Villa of the Papyri. Wikipedia. https://en.wikipedia.org/wiki/Villa_of_the_Papyri (truy cập ngày 3 tháng 9 năm 2026)

[4] University of Kentucky. (2026, 6월 25일). The day the Herculaneum scrolls began speaking again. UKNow. https://uknow.uky.edu/research/day-herculaneum-scrolls-began-speaking-again (truy cập ngày 3 tháng 9 năm 2026)

[5] Seales, W. B., Parker, C. S., Segal, M., Tov, E., Shor, P., & Porath, Y. (2016). From damage to discovery via virtual unwrapping: Reading the scroll from En-Gedi. Science Advances, 2(9), e1601247. https://doi.org/10.1126/sciadv.1601247

[6] Mocella, V., Brun, E., Ferrero, C., & Delattre, D. (2015). Revealing letters in rolled Herculaneum papyri by X-ray phase-contrast imaging. Nature Communications, 6, 5895. https://doi.org/10.1038/ncomms6895

[7] Papyrus. Encyclopaedia Britannica. https://www.britannica.com/topic/papyrus-writing-material (truy cập ngày 3 tháng 9 năm 2026)

1.2. Truy vết của đội điều tra khoa học 1: bóc lớp vỏ hành ảo

Ý tưởng đọc mà không mở

Giáo sư Brent Seales của Đại học Kentucky, Hoa Kỳ, chọn một con đường khác. Ý tưởng là không mở di vật, mà mở nó trong máy tính. Hiện vật được giữ nguyên và chỉ thu ảnh 3 chiều. Sau đó, trên màn hình, từng lớp được bóc ra. Đó là cách đi dần vào bên trong như bóc vỏ hành[1].

Nhóm nghiên cứu đặt tên phương pháp này là mở cuộn ảo. Trong tiếng Anh, nó được gọi là virtual unwrapping. Năm 2016, phương pháp này thành công lần đầu. Đối tượng là cuộn En Gedi đã thấy ở phần trước. Di vật này được tìm thấy tại nền hội đường cổ vào năm 1970. Bề ngoài là một khối cháy đen. Khi nhóm nghiên cứu mở nó trên màn hình, chữ Hebrew hiện ra. Đó là phần mở đầu của sách Lêvi trong Cựu Ước[1].

Mở cuộn ảo gồm năm bước. Đó là thu ảnh 3 chiều, tách lớp, gán giá trị mực, làm phẳng và ghép nối. Với bản chép En Gedi, trình tự này hoạt động đúng như vậy. Vì mực đã hiện sẵn trên ảnh[1].

Cuộn giấy Herculaneum có điều kiện xấu hơn bản chép En Gedi. Nó bị ép bởi sức nóng núi lửa nên mặt cắt biến dạng. Hình tròn bị sụp xuống và các lớp dính vào nhau. Hơn nữa, mực không hiện trên ảnh. Trong năm bước, bước thứ hai và thứ ba cùng trở thành trở ngại[2].

Vì vậy, với Herculaneum, trình tự được thay đổi. Trước hết là tách lớp và làm phẳng. Việc tìm mực được thực hiện riêng sau đó. Mục 1.3 là câu chuyện về việc tìm mực ấy[2].

Ảnh 3 chiều do máy gia tốc hạt chụp

Bước đầu tiên là chụp toàn bộ bên trong cuộn giấy[1]. CT bệnh viện thường phân biệt được đến khoảng 1 milimét. Lớp giấy cói mỏng hơn nhiều. Độ dày của một lớp còn chưa bằng sợi tóc. Với thiết bị bệnh viện, các lớp bị nhòe vào nhau[3].

Nhóm nghiên cứu tìm đến máy gia tốc hạt. Ban đầu, họ đến Diamond Light Source ở Anh[3]. Sau đó, họ cũng dùng Cơ sở Bức xạ Synchrotron châu Âu ở Grenoble, Pháp[4]. Những cơ sở này tạo ra tia X rất sáng và đi rất thẳng. Ánh sáng càng sáng thì càng thu được ảnh rõ trong thời gian ngắn[3].

Nhóm nghiên cứu dựng cuộn giấy trong ống và cố định chắc chắn. Vì chỉ cần rung nhẹ trong khi chụp, ảnh cũng sẽ bị mờ. Dữ liệu thu được theo cách này có độ phân giải khoảng 8 micromét cho mỗi voxel. Voxel là điểm ảnh không gian có cả giá trị chiều sâu. Có thể hiểu nó như việc biến một điểm trong chương trình vẽ thành một khối lập phương nhỏ[3].

Lượng dữ liệu rất lớn. Khi chụp một cuộn giấy, khoảng 300 terabyte dữ liệu được tạo ra. Đó là dung lượng đủ lấp đầy bộ nhớ của hàng nghìn điện thoại thông minh. Chính việc xử lý khối dữ liệu này đã trở thành một bài toán mới[4].

Cơ sở ở Grenoble, Pháp được gọi là nguồn sáng thế hệ thứ tư. Năm 2020, thiết bị mới bắt đầu vận hành. 21 nước châu Âu đã chi 150 triệu euro để nâng cao hiệu năng. Cơ sở này vốn được xây để quan sát vật chất và hiện tượng sống. Ban đầu, không ai biết nó sẽ được dùng để đọc các cuộn giấy cổ đại[4].

Phân đoạn thể tích để tách các lớp

Sau khi thu ảnh 3 chiều, bước tiếp theo là tách lớp. Công việc này được gọi là phân đoạn thể tích. Đó là việc tách từng lớp trong ảnh không gian. Máy tính lần theo đường đi của bề mặt giấy trên từng ảnh mặt cắt. Khi nối các đường ấy lại, mô hình 3 chiều của một tờ giấy được tạo ra. Mô hình này được gọi là mesh[1].

Ban đầu, con người dùng chuột để đánh dấu từng ranh giới[2]. Chiều dài giấy chứa trong một cuộn vượt quá 1 mét[4]. Nó quá dài và quá dày đặc để lần theo bằng tay. Hơn nữa, các lớp bị ép dính vào nhau nên ranh giới mờ[2].

Tốc độ tăng lên khi những người tham gia cuộc thi đưa ra công cụ tự động hóa. Tiêu biểu là chương trình do Julian Schilliger của Thụy Sĩ tạo ra. Chương trình này lần theo hoa văn của các ảnh mặt cắt lân cận để nối các lớp. Công việc từng mất nhiều ngày với con người được rút xuống còn vài giờ[2].

Công việc này không phải quy mô mà một phòng thí nghiệm có thể làm một mình. Ban tổ chức mở nguyên dữ liệu chụp trên Internet. Những người tham gia công khai mã nguồn chương trình mình tạo ra. Người sau thêm chức năng lên trên công cụ của người trước. Bản đọc và mã nguồn hiện nay vẫn có thể được tải xuống bởi bất kỳ ai[5].

Làm phẳng rồi ghép nối

Khi tách lớp xong, mô hình giấy vẫn cuộn tròn. Nếu để nguyên như vậy, chữ sẽ trông bị cong. Vì vậy, máy tính thực hiện phép tính để làm phẳng mô hình. Bước này được gọi là làm phẳng[1].

Máy tính biến các đỉnh của mô hình thành những hạt nhỏ. Nó giả định rằng giữa hạt này và hạt kia được nối bằng lò xo. Sau đó, nó lặp lại phép tính đặt tấm lưới này xuống mặt phẳng. Mức độ lò xo giãn ra hoặc co lại được giảm đến mức thấp nhất. Khi đó, hình dạng chữ không bị méo nhiều[1].

Bước cuối cùng là ghép nối. Cuộn giấy không thể được mở toàn bộ trong một lần. Máy tính chia nó thành nhiều khu vực và làm phẳng riêng. Các khu vực đã làm phẳng phải được ghép lại thành một tờ. Khi đó, người ta dùng hoa văn sợi dạng lưới đặc trưng của giấy cói. Nếu hoa văn sợi của hai khu vực ăn khớp, điều đó có nghĩa chúng vốn nằm cạnh nhau. Khi cả bước đối chiếu này kết thúc, mở cuộn ảo hoàn tất[1].

Năm bước này không kết thúc chỉ trong một lần. Nếu có sai sót ở một bước, tất cả các bước sau sẽ lệch. Nếu tách lớp sai, chữ của tờ khác sẽ lẫn vào. Vì vậy, nhóm nghiên cứu xử lý lặp lại cùng một cuộn giấy nhiều lần. Họ kiểm tra kết quả bằng mắt, đổi thiết lập rồi chạy lại[5].

Tài liệu tham khảo

[1] Seales, W. B., Parker, C. S., Segal, M., Tov, E., Shor, P., & Porath, Y. (2016). From damage to discovery via virtual unwrapping: Reading the scroll from En-Gedi. Science Advances, 2(9), e1601247. https://doi.org/10.1126/sciadv.1601247

[2] Vesuvius Challenge. (2024, 2월 5일). Vesuvius Challenge 2023 Grand Prize awarded: we can read the scrolls! scrollprize.org. https://scrollprize.org/grandprize (truy cập ngày 3 tháng 9 năm 2026)

[3] Vesuvius Challenge. FAQ. scrollprize.org. https://scrollprize.org/faq (truy cập ngày 3 tháng 9 năm 2026)

[4] University of Kentucky. (2026, 6월 25일). The day the Herculaneum scrolls began speaking again. UKNow. https://uknow.uky.edu/research/day-herculaneum-scrolls-began-speaking-again (truy cập ngày 3 tháng 9 năm 2026)

[5] Vesuvius Challenge. (2026, 6월 25일). An entire Herculaneum scroll has been read for the first time. scrollprize.org. https://scrollprize.org/firstscroll (truy cập ngày 3 tháng 9 năm 2026)

1.3. Truy vết của đội điều tra khoa học 2: tìm dấu vân tay vi mô của mực

Carbon viết trên carbon

Ngay cả sau khi mở cuộn ảo kết thúc, màn hình vẫn còn trống. Giấy đã được làm phẳng, nhưng chữ không hiện ra[1]. Nguyên nhân là vấn đề carbon đã thấy ở trên. Giấy cói cũng là carbon và mực đen cũng là carbon. Tia X không phân biệt được hai thứ[2].

Nhóm nghiên cứu của Giáo sư Seales đưa ra một giả thuyết. Đó là mực không thể hoàn toàn không để lại dấu vết[1]. Mực chạm vào giấy ở trạng thái lỏng. Chất lỏng thấm vào giữa các sợi. Và khi khô, nó kéo nhẹ bề mặt giấy[3].

Khi đó, hình dạng bề mặt giấy thay đổi rất nhỏ. Dù mật độ giống nhau, độ gồ ghề vẫn khác. Nhóm nghiên cứu quyết định tìm độ gồ ghề này[3]. Vì kích thước quá nhỏ để mắt nhìn thấy, họ phải giao việc đó cho máy[1].

Công cụ được dùng ở đây là học máy. Đây không phải cách con người viết từng quy tắc một. Người ta cho máy xem rất nhiều ví dụ có gắn đáp án để nó tự tìm quy tắc. Nguyên lý giống như việc cho xem hàng nghìn ảnh mèo rồi máy nhận ra mèo. Trong việc tìm mực, người ta cho máy xem mẫu đã đánh dấu nơi có mực và nơi không có mực[1].

Manh mối gọi là crackle

Tháng 3 năm 2023, một cuộc thi mang tên Vesuvius Challenge được tổ chức[1]. Giáo sư Seales, Nat Friedman và Daniel Gross cùng khởi xướng cuộc thi này[4]. Ban tổ chức phát miễn phí dữ liệu chụp trên Internet. Họ treo giải thưởng và thu hút lập trình viên trên khắp thế giới[1].

Trong vài tháng đầu, không ai tìm được mực. Mùa hè năm đó, một người tham gia tên Casey Handmer đã nắm được manh mối. Anh nhìn thật lâu bằng mắt thường vào dữ liệu bề mặt đã được trải ra. Rồi anh phát hiện một hoa văn khác thường. Đó là những vết nứt nhỏ li ti[1].

Có thể nghĩ đến mặt ruộng khô hạn. Bùn khô nứt ra như một tấm lưới. Ở nơi có mực bám cũng xuất hiện những vết nứt nhỏ tương tự. Những người tham gia gọi hoa văn này là crackle. Nhờ phát hiện này, Handmer nhận giải mực đầu tiên[1].

Handmer thông báo phát hiện này ngay trong phòng trò chuyện trực tuyến. Việc chia sẻ đó đã đổi dòng chảy của cuộc thi. Những người tham gia khác lập tức dùng hoa văn này làm tài liệu học cho máy[1].

Chỉ cho nhìn qua một cửa sổ nhỏ

Sinh viên Mỹ Luke Farritor là người hành động trước. Anh dạy mô hình học máy nhận biết hoa văn crackle. Mô hình trả lời bằng xác suất về nơi có mực. Càng học, câu trả lời càng chính xác[1].

Vấn đề là dữ liệu học quá ít. Khi dữ liệu ít, máy sẽ học thuộc câu trả lời. Điều này được gọi là quá khớp. Nó giống một học sinh chỉ học thuộc đáp án mà không hiểu đề thi[1].

Những người tham gia xoay và lật dữ liệu để tăng số lượng. Dù xoay hình, tính chất của hoa văn vết nứt vẫn giữ nguyên. Nhờ vậy, máy chỉ học đặc điểm của chính hoa văn[1].

Kích thước cửa sổ đầu vào cũng được chỉnh lại. Ban đầu, họ cho máy thấy một vùng rộng trong mỗi lần. Khi nhìn rộng, hình dạng chữ hiện ra trọn vẹn. Khi đó, máy bắt đầu bịa ra chữ. Nó vẽ thêm những nét có vẻ hợp lý ở nơi không có mực. Điều này được gọi là ảo giác của AI[1].

Các nhà nghiên cứu thu cửa sổ xuống còn 64 nhân 64 pixel. Với độ rộng đó, hình dạng chữ không hiện ra. Máy phải phán đoán chỉ bằng kết cấu của các vết nứt nhỏ. Việc bịa ra chữ giảm mạnh. Thay vào đó, dấu bút thật hiện lên đúng trên màn hình[1].

Mô hình đảm nhiệm phép tính này được tạo bằng cách ghép hai loại mô hình. Một loại là mạng nơ-ron tích chập 3 chiều. Đó là cấu trúc quét hoa văn trong một khối lập phương nhỏ để rút ra đặc điểm. Loại còn lại là transformer. Đó là cấu trúc xét xem các tín hiệu ở xa nhau liên quan với nhau đến mức nào. Khi dùng hai cấu trúc cùng nhau, máy không bỏ sót hoa văn mờ trong những lớp mỏng[1].

Màn hình do máy đưa ra không được nhận ngay là đáp án đúng. Các chuyên gia cổ văn tự học đọc riêng màn hình đó. Họ xem cách sắp xếp từ Hy Lạp có đúng ngữ pháp không. Họ cũng kiểm tra xem câu tiếp theo có nghĩa không. Nếu cách đọc của máy và con người không khớp, đối tượng đó được tính lại[1].

Chỉ bằng độ gồ ghề của bề mặt

Sau đó, phương pháp tìm mực tiếp tục tiến bộ. Người tham gia Youssef Nader dùng phương pháp học trước. Đây là cách cho máy xem trước một lượng lớn dữ liệu cuộn giấy không có bảng đáp án. Máy tự học hình dạng của sợi giấy cói. Sau đó, người ta chỉ dạy việc tìm mực bằng một bảng đáp án nhỏ[1].

Tháng 3 năm 2026, một nghiên cứu xử lý chính độ gồ ghề đã xuất hiện. Giorgio Angelotti, Federica Nicolardi, Paul Henderson và Brent Seales cùng viết nghiên cứu này. Nhóm nghiên cứu đo bề mặt của mẫu giấy cói đã được trải ra bằng thiết bị quang học. Họ cho máy học dữ liệu chỉ chứa độ cao thấp. Họ không đưa vào thông tin về độ sáng hay mật độ[3].

Dù vậy, máy vẫn phân biệt được nơi có mực. Điều đó có nghĩa là có thể tìm mực chỉ bằng độ cao thấp của bề mặt. Nhóm nghiên cứu cũng tính luôn cần đo dày đến mức nào. Con số này trở thành chuẩn độ phân giải cho lần chụp tiếp theo. Bài báo này được đăng trên tạp chí Scientific Reports trong cùng năm[3].

Tuy vậy, việc tìm mực vẫn chưa phải là công nghệ hoàn chỉnh. Những đoạn còn lại rất ít mực vẫn hiện ra mờ. Ở nơi các lớp dính ép vào nhau, bản thân dữ liệu chụp cũng đục. Hai tờ giấy khác nhau đôi khi dính thành một tờ. Mô hình hoạt động tốt với một cuộn giấy cũng có khi chạy sai trên cuộn giấy khác. Vì vậy, con người vẫn kiểm tra bằng mắt và chỉnh sửa ở từng bước giữa. Nghiên cứu hiện vẫn tiến lên từng bước[5].

Tài liệu tham khảo

[1] Vesuvius Challenge. (2024, 2월 5일). Vesuvius Challenge 2023 Grand Prize awarded: we can read the scrolls! scrollprize.org. https://scrollprize.org/grandprize (truy cập ngày 3 tháng 9 năm 2026)

[2] Mocella, V., Brun, E., Ferrero, C., & Delattre, D. (2015). Revealing letters in rolled Herculaneum papyri by X-ray phase-contrast imaging. Nature Communications, 6, 5895. https://doi.org/10.1038/ncomms6895

[3] Angelotti, G., Nicolardi, F., Henderson, P., & Seales, W. B. (2026). Ink detection from surface topography of the Herculaneum papyri. Scientific Reports. https://doi.org/10.1038/s41598-026-58467-1

[4] University of Kentucky. (2026, 6월 25일). The day the Herculaneum scrolls began speaking again. UKNow. https://uknow.uky.edu/research/day-herculaneum-scrolls-began-speaking-again (truy cập ngày 3 tháng 9 năm 2026)

[5] Vesuvius Challenge. (2026, 7월 10일). Open problems: why reading every Herculaneum scroll is still a challenge. scrollprize.org. https://scrollprize.org/2026_open_problems (truy cập ngày 3 tháng 9 năm 2026)

1.4. Sự thật được hé lộ

Từ đầu tiên được lấy lại trong năm 2023

Thành quả đầu tiên là một từ. Thông báo được đưa ra vào ngày 12 tháng 10 năm 2023. Nhân vật chính là sinh viên Mỹ Luke Farritor. Khi đó anh 21 tuổi. Từ anh đọc được là porphyras. Đó là một từ Hy Lạp có nghĩa là màu tím. Đây không phải là từ thường dùng trong văn bản cổ. Farritor nhận 40.000 đô la cho giải chữ cái đầu tiên. Người đứng thứ hai, Youssef Nader, nhận 10.000 đô la. Đây là trường hợp đầu tiên đọc được một từ từ cuộn giấy chưa mở[7].

Kết quả cuộc thi được công bố vào ngày 5 tháng 2 năm 2024. Giải lớn được trao cho ba người. Đó là Youssef Nader của Ai Cập, Luke Farritor của Mỹ và Julian Schilliger của Thụy Sĩ. Tiền thưởng là 700.000 đô la[1].

Hiện vật mà họ đọc được là cuộn giấy do Viện Hàn lâm Paris sở hữu. Tên của nó là PHerc. Paris. 4. Ba người đã khôi phục hơn 2.000 ký tự tiếng Hy Lạp từ cuộn giấy này. Đó là văn bản của 15 cột. Cột là ô chữ viết chia theo chiều dọc trên cuộn giấy. Phần này chiếm khoảng 5 phần trăm toàn bộ cuộn giấy[1].

Tiêu chuẩn chấm rất nghiêm ngặt. Người dự thi phải nộp bốn đoạn, mỗi đoạn 140 ký tự. Trong mỗi đoạn, phải đọc được hơn 85 phần trăm số chữ mới qua. Các chuyên gia cổ văn tự học thẩm định trong trạng thái đã che tên người dự thi[1].

Nội dung là một cuộc bàn luận về khoái lạc. Nó xét xem thứ hiếm có có nhất thiết vui hơn thứ thường gặp hay không. Trong văn bản có một đoạn lấy thức ăn làm ví dụ. Ý chính là không tin rằng một thứ chỉ vì hiếm mà lập tức vui hơn. Câu văn cho thấy rõ lối nghĩ của trường phái Epicurus[1].

Cuộn giấy lấy lại được nhan đề

Thành quả tiếp theo là nhan đề. Giải nhan đề đầu tiên được trao vào ngày 5 tháng 5 năm 2025. Người đoạt giải là Marcel Roth và Micha Novak. Tiền thưởng là 60.000 đô la[2].

Hiện vật được xét là PHerc. 172 của Thư viện Bodleian ở Oxford. Những chữ hiện trên màn hình là 'Về thói xấu' của Philodemus. Nó được đọc là quyển 1. Nhóm nghiên cứu của Federica Nicolardi thuộc Đại học Naples Federico II đã xác nhận điều này. Đây là trường hợp đầu tiên đọc được nhan đề của một cuộn giấy đang cuộn mà không mở nó ra[2].

Sau đó, hơn 70 cột đã được đọc từ cuộn giấy này. Đây là văn bản bàn về những khiếm khuyết trong tính cách như tham lam, kiêu ngạo và nịnh bợ. Giờ đây có thể theo dõi dòng lập luận từ đầu[3].

Một cuốn sách lần đầu được đọc đến hết

Ngày 25 tháng 6 năm 2026, có một tin lớn hơn. Đó là thông báo rằng một cuộn giấy đã được đọc từ đầu đến cuối. Hiện vật là PHerc. 1667 của Thư viện Quốc gia Naples. Những người tham gia cuộc thi gọi nó là cuộn giấy số 4[4].

Việc chụp được thực hiện tại Cơ sở Bức xạ Synchrotron châu Âu ở Pháp. Sau đó, nó lần lượt trải qua mở cuộn ảo và tìm mực. Độ dài phần chữ được khôi phục khoảng 1,4 mét. Tính theo cột thì khoảng 22 cột. Đây là lần đầu tiên người ta đọc trọn một cuộn giấy Herculaneum[4].

Nội dung là một luận văn tiếng Hy Lạp bàn về đạo đức và phẩm chất con người[3]. Trong văn bản có nhắc đến cháu của triết gia Khắc kỷ Chrysippus. Tên người đó là Aristocreon[4]. Thời điểm văn bản được viết được cho là từ cuối thế kỷ 3 trước Công nguyên đến khoảng thế kỷ 2 trước Công nguyên. Tác giả là ai thì vẫn chưa được xác định[3].

Kết quả nghiên cứu cũng được công bố dưới dạng bài báo. Tổng cộng 27 người đứng tên, trong đó có Giorgio Angelotti, Stephen Parsons và Federica Nicolardi[5]. Dữ liệu, mã nguồn và bản đọc cũng được công bố cùng lúc[4].

Nhóm nghiên cứu cũng công bố quy trình kiểm chứng. Họ chụp lại PHerc. Paris. 4, cuộn giấy đã được đọc năm 2023, ở độ phân giải cao hơn. Bản đọc thu được từ dữ liệu mới khớp nguyên với kết quả trước đó. Đây là bằng chứng cho thấy đó không phải là chữ do máy bịa ra. Nó cũng có nghĩa là có thể áp dụng cùng phương pháp cho các cuộn giấy khác[4].

600 cuốn vẫn còn lại

Trong cùng thông báo còn có tin khác. Nhan đề của cuộn giấy PHerc. 139 đã được đọc. Đó là quyển 8 của 'Về các vị thần' của Philodemus. Trước đó, người ta chỉ biết quyển 1 của tác phẩm này. Lần đầu tiên sự thật rằng tác phẩm này kéo dài hơn tám quyển được hé lộ[3].

Hiện vẫn còn hơn 600 cuộn giấy chưa đọc được[3]. Cuộc thi đã đặt ra Grand Prix năm 2027. Tổng tiền thưởng là 1 triệu đô la. Người đứng thứ nhất nhận 800.000 đô la. Đối tượng là 13 cuộn giấy được chỉ định. Ở mỗi cột, phải đọc được hơn 70 phần trăm số chữ còn lại. Hạn chót là ngày 25 tháng 6 năm 2027[6].

Câu chuyện này vẫn còn nhiều chỗ trống. Dù đọc được chữ, việc hiểu nghĩa của nó vẫn là việc riêng. Chỉ riêng việc chỉnh lý học thuật một cột bản đọc cũng mất nhiều thời gian. Vẫn còn những cuộn giấy chưa rõ tác giả[3].

Dù vậy, hướng đi đã rõ. Những cuốn sách mà không ai mở được suốt 2.000 năm đã được trải ra trên màn hình. Thư viện từng tưởng đã biến mất trong lửa đang dần trở lại[4].

Người làm nên việc này không phải chỉ một người. Có các nhà vật lý phụ trách chụp ảnh và các lập trình viên chia tách từng lớp. Cũng có các học giả cổ văn tự học đọc tiếng Hy Lạp cổ. Sinh viên và nhân viên công ty tham gia cuộc thi cũng góp sức. Việc cuốn sách được mở sau 2.000 năm là kết quả do những người này cùng tạo ra. Những cuộn giấy còn lại cũng có thể được mở theo cách ấy[4].

Tài liệu tham khảo

[1] Vesuvius Challenge. (2024, 2월 5일). Vesuvius Challenge 2023 Grand Prize awarded: we can read the scrolls! scrollprize.org. https://scrollprize.org/grandprize (truy cập ngày 3 tháng 9 năm 2026)

[2] Vesuvius Challenge. (2025, 5월 5일). $60,000 First Title Prize awarded! scrollprize.substack.com. https://scrollprize.substack.com/p/60000-first-title-prize-awarded (truy cập ngày 3 tháng 9 năm 2026)

[3] University of Kentucky. (2026, 6월 25일). The day the Herculaneum scrolls began speaking again. UKNow. https://uknow.uky.edu/research/day-herculaneum-scrolls-began-speaking-again (truy cập ngày 3 tháng 9 năm 2026)

[4] Vesuvius Challenge. (2026, 6월 25일). An entire Herculaneum scroll has been read for the first time. scrollprize.org. https://scrollprize.org/firstscroll (truy cập ngày 3 tháng 9 năm 2026)

[5] Angelotti, G., Parsons, S., Nicolardi, F., Nader, Y., Johnson, S., Josey, D., et al. (2026). Complete virtual unwrapping and reading of a rolled Herculaneum papyrus. arXiv:2606.29085. https://arxiv.org/abs/2606.29085 (truy cập ngày 3 tháng 9 năm 2026)

[6] Vesuvius Challenge. Open Prizes. scrollprize.org. https://scrollprize.org/prizes (truy cập ngày 3 tháng 9 năm 2026)

[7] Vesuvius Challenge. (2023, 10월 12일). First word discovered in unopened Herculaneum scroll by 21yo computer science student. scrollprize.org. https://scrollprize.org/firstletters (truy cập ngày 3 tháng 9 năm 2026)

Chương 2. Lời kinh nở ra từ hòm thánh đã cháy: sách Lê-vi bị niêm phong của En-Gedi

2.1. Hồ sơ vụ việc

Ốc đảo phía tây Biển Chết

Ein Gedi là một ốc đảo ở bờ phía tây Biển Chết. Tên này cũng được viết là En-Gedi. Xung quanh là sa mạc khô cằn, nhưng nơi đây có nước suối trào lên. Có nước nên con người tụ lại. Trong hàng nghìn năm, con người đã lập làng và sống bên nguồn nước này. Tên vùng đất này cũng xuất hiện trong sách 1 Samuel của Cựu Ước. Đó là câu chuyện David trốn vua Saul và ẩn náu trong hoang mạc Ein Gedi.

Theo khảo sát di tích, con người đã sống trên ngọn đồi này từ cuối thế kỷ 8 trước Công nguyên. Rồi ngôi làng sụp đổ trong một trận hỏa hoạn lớn vào khoảng năm 600 sau Công nguyên[1]. Đây là vùng đất mà hơn 1.300 năm đã chất thành từng lớp trong lòng đất.

Khu vực quanh Biển Chết ít mưa và không khí khô. Khí hậu này giúp đồ vật cũ được giữ lại lâu. Gỗ, da và vải ít bị mục hơn những nơi khác. Đây cũng là lý do Cuộn Biển Chết được tìm thấy trong vùng này.

Năm 1970, một cuộc khai quật lớn bắt đầu tại đây. Cuộc khai quật do Dan Barag và Ehud Netzer dẫn dắt. Hai người thuộc Viện Khảo cổ học của Đại học Hebrew Jerusalem. Yosef Porath đã đào được cuộn giấy này vào ngày 5 tháng 5 năm 1970[1]. Thứ đoàn khai quật tìm thấy là nền của một giáo đường Do Thái cổ. Giáo đường là tòa nhà nơi cộng đồng Do Thái tụ họp để cầu nguyện và đọc Kinh Thánh. Trên sàn có lớp khảm làm bằng những viên đá nhỏ có màu khác nhau.

Bên trong giáo đường có một không gian đặt hòm thánh. Hòm thánh là chiếc hộp dùng để cất các cuộn Kinh Thánh linh thiêng. Cộng đồng Do Thái xem chiếc hộp này là nơi linh thiêng trong giáo đường. Đoàn khai quật dọn phần tàn tích cháy khỏi hòm thánh. Trong đống tàn tích, một khối màu đen xuất hiện[1].

Việc nó xuất phát từ hòm thánh tự nó đã là một manh mối quan trọng. Người ta không đặt bất kỳ văn bản nào vào hòm thánh của giáo đường. Họ đặt vào đó các cuộn Kinh Thánh mà cộng đồng đọc và tuân giữ. Đây là căn cứ để đoán bản chất của khối màu đen.

Bản chép Kinh Thánh cổ tự nó là tư liệu quý. Vào thời chưa có máy in, con người chép lại bằng tay. Mỗi lần chép lại, chữ có thể thay đổi. Khi một bản chép cổ xuất hiện, ta có thể kiểm tra văn bản đã được truyền lại ra sao. Đó là lý do các học giả bám sát khối màu đen lấy ra từ hòm thánh.

Khối đen do lửa để lại

Lửa có sức xóa bỏ ghi chép. Giấy và da đều thành tro trong lửa. Nhưng ở Ein Gedi, tình hình hơi khác. Cuộn giấy bị tàn tích đè lên đã được nung chậm trong tình trạng thiếu oxy. Nó không tan thành tro mà cứng lại như than. Sự biến đổi cứng lại như than này được gọi là carbon hóa. Khi cháy trong lửa thiếu không khí, hiện tượng này xảy ra.

Nguyên lý giống quá trình gỗ trở thành than trong lò than. Nếu chỉ đốt gỗ, chỉ còn lại tro. Nếu phủ đất để chặn không khí rồi nung, than đen còn lại. Cuộn giấy bị carbon hóa cũng được tạo ra như vậy.

Khối bị carbon hóa bị chôn dưới tro và đất của tòa nhà đổ nát. Vì không tiếp xúc với không khí nên sự mục rữa cũng dừng lại. Da vốn là vật liệu dễ mục. Nếu nó nằm nguyên trên mặt đất, nó đã biến mất từ lâu. Ngọn lửa phá hủy di vật cũng là sức mạnh giữ lại di vật.

Khối mà đoàn khai quật lấy ra thoạt nhìn giống một cục than. Nhưng ở mặt bên bị đứt vẫn còn những lớp cuộn tròn. Các nhà khảo cổ xác định đây là một cuộn giấy bằng da đã được cuộn lại[1]. Di vật được đặt tên theo vùng đất nơi nó được tìm thấy. Tên gọi cuộn giấy Ein Gedi ra đời như vậy.

Cuộn giấy này là giấy da dê cừu được làm từ da động vật đã xử lý[1]. Cuộn giấy tìm thấy ở Herculaneum của Ý được làm bằng giấy cói. Giấy cói là loại giấy thực vật làm bằng cách ép thân cây lại với nhau. Hai di vật khác nhau ngay từ vật liệu nền. Khác biệt này về sau trở thành điểm rẽ trong cách đọc.

Giấy da dê cừu là vật liệu tốn nhiều công để làm. Người ta lấy lông và mỡ khỏi da động vật rồi rửa sạch. Da được căng trên khung và phơi khô. Chữ được viết lên tấm mỏng đã xử lý như vậy. Nó dai và bền hơn giấy, nhưng đắt tiền.

Tình trạng khi phát hiện cũng hiếm gặp. Các cuộn giấy cổ thường được tìm thấy dưới dạng mảnh vỡ nhỏ. Cuộn Biển Chết tìm thấy trong các hang Qumran từ năm 1947 cũng có nhiều mảnh. Di vật ở Ein Gedi được tìm thấy trong khi vẫn giữ nguyên hình dạng cuộn. Điều đó có nghĩa là bên trong còn thêm vài lớp nữa.

Di vật không thể chạm vào

Muốn đọc chữ thì phải mở cuộn giấy ra. Nhưng di vật này không thể mở được. Vì sức nóng làm da co lại và các lớp dính vào nhau. Bên ngoài cứng như đá, bên trong dễ vỡ vụn. Nó ở tình trạng có thể thành bột nếu chạm bằng đầu ngón tay[1].

Không phải là không có cách. Có thể dùng dao tách các lớp ra. Cũng có thể bôi hóa chất để làm mềm da. Cả hai cách đều làm hỏng di vật. Một cuộn giấy đã vỡ thì không thể phục hồi.

Chỉ nhìn bên ngoài thì không thể biết có chữ hay không. Bề mặt bị carbon hóa hoàn toàn đen và không có hoa văn. Mực cũng đen, da cũng đen, nên mắt thường không phân biệt được. Việc bên trong có chữ cũng chỉ là suy đoán từ hình dạng các lớp.

Những người xử lý di sản văn hóa luôn đặt điểm này lên trước. Di vật là thứ duy nhất trên đời và không thể chịu nổi một lần thất bại. Vì vậy, khoa học bảo tồn có nguyên tắc khảo sát mà không chạm vào. Phương pháp khảo sát này được gọi là khảo sát không xâm lấn. Xâm lấn nghĩa là đi sâu vào bên trong, còn không xâm lấn là ngược lại. Cuộn giấy Ein Gedi là di vật chỉ được phép khảo sát không xâm lấn.

45 năm im lặng

Cuộn giấy được đưa vào kho lưu giữ di vật của Israel. Về sau, Cơ quan Cổ vật Israel tiếp nhận và bảo quản di vật này[1]. Chờ đợi trong hộp là cách bảo tồn duy nhất lúc đó. Kho lưu giữ là không gian giữ nhiệt độ và độ ẩm ổn định. Nó chặn không khí, ánh sáng và côn trùng để di vật không hư hại thêm. Dù là di vật chưa đọc được, việc giữ tình trạng của nó vẫn tiếp tục.

Pnina Shor, người phụ trách bảo tồn Cuộn Biển Chết, tiếp nhận việc quản lý. Về sau, bà cũng có tên trong nghiên cứu đọc được cuộn giấy này[1].

Sự chờ đợi kéo dài từ năm 1970 đến năm 2015[1][2]. Trong 45 năm, không ai nhìn thấy chữ bên trong. Vì mặt ngoài chỉ đen kịt nên thậm chí khó biết có chữ hay không. Trong giới học giả, cũng có người nói đây là di vật sẽ không bao giờ đọc được.

Sự chờ đợi có lý do. Với kỹ thuật của năm 1970, người ta không thể xử lý khối này. Thiết bị chụp chia nhỏ bên trong vật thể cũng chưa phổ biến. Dù chụp được, cũng không có máy tính để tính toán lượng dữ liệu lớn như vậy. Di vật được để nguyên và chờ kỹ thuật phát triển.

Trong thời gian đó, những băn khoăn tương tự tiếp diễn ở nhiều nơi trên thế giới. Giấy cói Herculaneum của Ý cũng bị chặn bởi cùng một bức tường. Những nỗ lực cũ nhằm mở nó bằng phương pháp vật lý chỉ làm vỡ di vật. Hai di vật khác nơi tồn tại và khác vật liệu, nhưng vấn đề chỉ có một. Cần một cách đọc bên trong mà không chạm vào.

Điều phá vỡ sự im lặng không phải là một kỹ thuật dùng tay mới. Đó là chụp ảnh nhìn vào bên trong mà không cắt vật thể và tính toán bằng máy tính. Nhóm nghiên cứu Seales đã nghiên cứu lâu dài phương pháp đọc chữ cổ mà không chạm vào. Nghiên cứu này được Quỹ Khoa học Quốc gia Hoa Kỳ và Google hỗ trợ[2]. Họ cần một di vật để thử nghiệm phương pháp đã tích lũy lâu năm, và cuộn giấy Ein Gedi trở thành đối tượng đó.

Năm 2015, nhóm nghiên cứu của Brent Seales thuộc Đại học Kentucky ở Hoa Kỳ đảm nhận việc này. Kết quả nghiên cứu được đăng trên tạp chí Science Advances vào ngày 21 tháng 9 năm 2016[1][2].

Bây giờ chỉ còn một câu hỏi. Làm thế nào họ đọc được cả chữ bên trong mà không chạm vào? Câu trả lời nằm ở trình tự của thiết bị chụp và tính toán.

Tài liệu tham khảo

[1] Seales, W. B., Parker, C. S., Segal, M., Tov, E., Shor, P., & Porath, Y. (2016). From damage to discovery via virtual unwrapping: Reading the scroll from En-Gedi. Science Advances, 2(9), e1601247. https://doi.org/10.1126/sciadv.1601247

[2] University of Kentucky Research. (2016, 9월 22일). Virtually Unwrapping the En-Gedi Scroll. research.uky.edu. https://research.uky.edu/news/virtually-unwrapping-en-gedi-scroll (truy cập ngày 3 tháng 9 năm 2026)

2.2. Truy vết của đội điều tra khoa học: làm sạch kỹ thuật số bằng X-ray

Chụp bên trong mà không cắt

Công cụ đầu tiên mà nhóm nghiên cứu chọn là micro CT. CT là viết tắt của chụp cắt lớp vi tính. Nguyên lý giống kỹ thuật chụp dùng trong bệnh viện để nhìn vào bên trong cơ thể. Tia X được chiếu từ nhiều góc để đo mật độ bên trong vật thể. Máy tính gom các giá trị đó lại và tạo bản đồ ba chiều. Micro CT là thiết bị thực hiện việc này ở thang đo rất nhỏ.

Việc chụp cuộn giấy Ein Gedi dùng thiết bị Bruker SkyScan 1176. Đơn vị phụ trách chụp là Merkel Technologies của Israel. Công ty này tặng dữ liệu chụp cho Cơ quan Cổ vật Israel. Độ chia khi chụp là 18 micromet mỗi ô[1]. 1 micromet bằng 1 phần 1000 của 1 milimét. Đó là kích thước tương đương một phần nhỏ của độ dày sợi tóc.

Kích thước độ chia này có lý do. Muốn tách từng lớp để xem, phải đo nhỏ hơn nhiều so với một lớp da. Nếu độ chia lớn hơn độ dày của lớp, hai lớp sẽ bị nhập thành một khối. Độ chia khi chụp là điều kiện quyết định giới hạn của lớp có thể đọc được.

Dữ liệu do việc chụp tạo ra không phải là một bức ảnh. Đó là một khối ba chiều gồm các hạt lập phương rất nhỏ xếp dày đặc. Hạt này được gọi là voxel. Nếu pixel của ảnh là hạt phẳng, thì voxel là hạt ba chiều. Trong mỗi hạt có một con số biểu thị mật độ tại vị trí đó.

Ở bước này, di vật đã được chuyển vào máy tính mà không cần chạm vào. Từ đây, đối tượng được xử lý không còn là da mà là con số. Con số thì dù chạm bao nhiêu cũng không vỡ. Nếu tính sai, có thể làm lại từ đầu. Sức mạnh của khảo sát không xâm lấn nằm ở đây.

Những điểm sáng do mực pha kim loại để lại

Có dữ liệu ba chiều không có nghĩa là chữ hiện ra. Muốn thấy chữ, mật độ của mực và nền phải khác nhau. Điều kiện này không có ở giấy cói Herculaneum. Mực ở đó được làm từ bột than, còn giấy cói cháy cũng là khối carbon. Mật độ của hai thứ giống nhau nên khó phân biệt bằng tia X.

Cuộn giấy Ein Gedi có điều kiện tốt. Nhóm nghiên cứu xác nhận trong dữ liệu quét rằng mật độ ở nơi có chữ cao hơn vùng xung quanh. Vì vậy, họ cho rằng trong mực này có thể lẫn kim loại như sắt hoặc chì.

Tuy nhiên, vì không có chỗ nào lộ ra bề mặt nên họ không thể phân tích trực tiếp thành phần. Nhóm nghiên cứu ghi rõ giới hạn này trong bài báo[1].

Kim loại hấp thụ nhiều tia X. Nơi hấp thụ nhiều hiện sáng trên màn hình. Da cháy hấp thụ tia X ít hơn nên hiện tối. Độ chênh sáng giữa chữ và nền hình thành như vậy. Chênh lệch này trở thành manh mối đầu tiên để máy tính tìm chữ. Đây cũng là lý do cuộn giấy Ein Gedi được đọc trước Herculaneum.

Tuy nhiên, thành phần của mực này vẫn chưa được xác định. Những gì nhóm nghiên cứu đưa ra chỉ là quan sát về mật độ cao và suy đoán từ quan sát đó[1]. Muốn biết người chép cổ đã trộn gì, phải có mẫu vật. Không xử lý điều chưa chắc chắn như sự thật đã chắc chắn là nguyên tắc của lĩnh vực này.

Bốn bước mở cuộn ảo và loại bỏ vết loang

Tìm được điểm sáng chưa có nghĩa là đọc được chữ. Chữ nằm rải trên các lớp cuộn tròn. Cần mở các lớp này ra trong máy tính. Công việc này được gọi là mở cuộn ảo. Nhóm nghiên cứu Seales chia mở cuộn ảo thành bốn bước[1].

Bước đầu tiên là tìm bề mặt của từng lớp. Máy tính cắt dữ liệu ba chiều thành lát mỏng và xác định tọa độ của lớp da. Các điểm thu được được nối lại để tạo thành bề mặt dạng tấm mỏng. Nếu các lớp dính vào nhau, bước này sẽ mất hướng. Nếu đọc nhầm hai lớp thành một lớp, chữ cũng sẽ bị trộn lẫn.

Bước thứ hai là phủ giá trị độ sáng lên bề mặt đó. Các giá trị chụp tại tọa độ mà bề mặt đi qua được đọc và chuyển lên mặt phẳng. Lúc này, mực có mật độ cao hiện thành các nét sáng.

Bước thứ ba là làm phẳng bề mặt bị nhăn. Khi ép hình khối ba chiều thành mặt phẳng, hình dạng sẽ bị kéo giãn hoặc méo mó. Đây là vấn đề giống khi trải quả địa cầu thành bản đồ giấy thì vùng cực bị phóng rộng. Phép tính trải bề mặt theo hướng giảm tối đa sự méo này.

Bước thứ tư là ghép nhiều bề mặt đã được trải riêng lại với nhau. Ranh giới giữa các bề mặt được căn khớp để hợp thành một hình lớn[1].

Trong bốn bước, bước đầu tiên khó nhất. Nếu xác định sai lớp một lần, cả ba bước sau đều lệch. Trong hình được trải sai, thậm chí có thể xuất hiện chữ vốn không có. Đó là lý do các nhà nghiên cứu kiểm tra bước này bằng tay.

Sau bốn bước này, ống trụ màu đen trở thành một mặt phẳng. Nhóm nghiên cứu đã trải năm lớp da bằng cách này. Diện tích bề mặt được trải ước tính khoảng 94 centimét vuông[1]. Diện tích này nhỏ hơn một chút so với một trang vở.

Hình đã trải vẫn còn vấn đề. Độ dày của da khác nhau ở nhiều chỗ, và nếp nhăn vẫn còn. Nhiễu sinh ra trong quá trình chụp cũng lẫn vào. Vì vậy, có chỗ quá sáng và có chỗ tối. Khi chữ và vết loang trộn vào nhau, việc đọc trở nên khó khăn.

Nhóm nghiên cứu loại bỏ các vết loang này bằng tính toán. Trước hết, họ đo độ sáng trung bình của các pixel xung quanh để tạo bản đồ độ sáng của nền. Bản đồ này chỉ chứa dòng sáng lan rộng. Sau đó, họ chia hình gốc cho bản đồ nền này. Sau phép chia, chênh lệch độ sáng lan rộng biến mất. Chỉ tín hiệu mực tập trung trong bề rộng hẹp còn lại rõ nét. Trong hình đã qua quá trình này, chữ Hebrew hiện ra[1].

Bài toán còn tiếp diễn trong năm 2026

Mở cuộn ảo sau đó cũng được dùng cho các di vật khác. Ngày 25 tháng 6 năm 2026, Vesuvius Challenge công bố một tin mới. Nội dung là họ đã đọc một cuộn giấy Herculaneum từ đầu đến cuối. Số di vật là PHerc. 1667. Việc chụp được thực hiện tại Cơ sở Bức xạ Synchrotron châu Âu ở Grenoble, Pháp. Phần văn bản Hy Lạp được mở ra gồm hai mươi hai cột, dài khoảng 1,4 mét. Nó được xác định là một tác phẩm về đạo đức học từ thế kỷ 2 trước Công nguyên[2].

Synchrotron là thiết bị làm các electron đang chạy nhanh bị bẻ cong bằng từ trường. Khi electron bị bẻ cong, tia X rất mạnh phát ra. Nếu chụp bằng ánh sáng này, cả khác biệt rất nhỏ về mật độ cũng hiện ra. Đó là độ rõ khó đạt được bằng thiết bị chụp trong phòng thí nghiệm. Đây là công cụ cao hơn một bậc so với thiết bị đã dùng ở Ein Gedi.

Thành quả này có được nhờ cải tiến phương pháp chụp. Trong phương pháp chụp mới, mực carbon hiện trực tiếp bên trong tư liệu. Nó bắt được phần mực trước đây không thấy vì khác biệt mật độ quá nhỏ. Khác biệt độ sáng vốn tự nhiên có ở Ein Gedi đã được tạo ra bằng kỹ thuật chụp. Người ta cho biết có thể dùng cùng phương pháp cho hơn 600 cuộn giấy còn lại[3].

Vẫn còn bài toán chưa giải. Bản tổng kết công bố ngày 10 tháng 7 năm 2026 nêu sáu vấn đề còn lại. Ở những đoạn các lớp bị ép chồng lên nhau, tia X tán xạ và làm hình ảnh mờ đi. Phép tính tìm lớp tạo ra lỗ hổng hoặc dán hai lớp vào nhau. Dữ liệu voxel không chứa thông tin lớp nào nối tiếp lớp nào. Dấu gán huấn luyện do con người tạo ra không chính xác đến mức từng hạt. Mô hình học từ một cuộn giấy không hoạt động tốt ở cuộn giấy khác. Vị trí mực nằm ở độ sâu nào trong lớp cũng chưa rõ[4].

Các cách giải quyết cũng được nêu cùng lúc. Nếu thiếu dấu gán đáp án, có thể để mô hình học quy luật từ chính dữ liệu. Mô hình học theo cách tự học được xem là một ứng viên. Các cấu trúc mạng nơ-ron mới cũng đang được dùng trong phép tính tìm lớp[4].

Một manh mối mới cũng xuất hiện. Nghiên cứu công bố ngày 29 tháng 3 năm 2026 dựa vào điểm rằng mực làm thay đổi hình dạng bề mặt. Nhóm nghiên cứu đo rất chính xác độ gồ ghề trên bề mặt giấy cói đã được mở sẵn rồi dạy cho máy học. Kết quả là chỉ bằng độ gồ ghề cũng có thể phân biệt nơi có chữ. Nghiên cứu này cũng tính xem cần thang đo chụp ở mức nào để đọc được độ gồ ghề[5]. Đây là nghiên cứu tìm lại manh mối từng có được từ mật độ ở Ein Gedi qua một tính chất khác.

Tài liệu tham khảo

[1] Seales, W. B., Parker, C. S., Segal, M., Tov, E., Shor, P., & Porath, Y. (2016). From damage to discovery via virtual unwrapping: Reading the scroll from En-Gedi. Science Advances, 2(9), e1601247. https://doi.org/10.1126/sciadv.1601247

[2] Vesuvius Challenge. (2026, 6월 25일). An entire Herculaneum scroll has been read for the first time. scrollprize.org. https://scrollprize.org/firstscroll (truy cập ngày 3 tháng 9 năm 2026)

[3] ASNT. (2026, 8월 13일). X-Ray Imaging and AI Recover First Complete Text from Sealed Ancient Scroll. Materials Evaluation. https://www.asnt.org/me/26/8/x-ray-ai-recover-text-sealed-ancient-scroll (truy cập ngày 3 tháng 9 năm 2026)

[4] Vesuvius Challenge. (2026, 7월 10일). Open Problems: Why Reading Every Herculaneum Scroll Is Still a Challenge. scrollprize.org. https://scrollprize.org/2026_open_problems (truy cập ngày 3 tháng 9 năm 2026)

[5] Angelotti, G., Nicolardi, F., Henderson, P., & Seales, W. B. (2026, 3월 29일). Ink Detection from Surface Topography of the Herculaneum Papyri. arXiv:2603.27698. https://arxiv.org/abs/2603.27698 (truy cập ngày 3 tháng 9 năm 2026)

2.3. Sự thật được làm sáng tỏ

Sách Lê-vi hiện lên trên màn hình

Những chữ hiện lên trên màn hình máy tính là chữ Hebrew. Nhóm nghiên cứu đã mở được năm lớp da, và bên trong xuất hiện hai cột chữ. Văn bản là sách Lê-vi, quyển thứ ba của Cựu Ước. Phần đọc được là chương 1 và chương 2 của sách Lê-vi[1].

Hình ảnh ghép lại chứa ba mươi lăm dòng chữ của một tờ trong cuộn giấy. Trong đó, mười tám dòng còn chữ. Mười bảy dòng còn lại là phần các học giả phục dựng. Phục dựng không phải là việc máy tính tự viết ra văn bản. Đó là việc đối chiếu nét còn lại với văn bản đã biết để điền vào chỗ thiếu. Số chữ và khoảng trắng trong một dòng là từ ba mươi ba đến ba mươi bốn[1].

Điểm quan trọng là bài báo đã tách rõ dòng còn lại và dòng phục dựng. Nhờ vậy, người đọc có thể phân biệt đâu là sự thật đã được xác nhận. Nếu công bố lẫn lộn phần còn lại và phần được điền thêm, người sau không thể kiểm chứng. Nhờ đó, các học giả khác có thể xem xét lại bằng cùng tư liệu.

Chương 1 sách Lê-vi nói về quy tắc của lễ thiêu. Lễ thiêu là nghi lễ dâng vật tế bằng cách đốt hoàn toàn trên bàn thờ. Chương 2 nói về quy tắc của lễ dâng bằng ngũ cốc. Cách trộn dầu vào bột mịn và đặt hương lên trên được ghi ở đây. Từ hòm thánh bị cháy đã xuất hiện quy tắc của nghi lễ dâng bằng lửa.

Văn bản được phục dựng cũng có câu đầu tiên của sách Lê-vi[1]. Đó là câu nói Đức Jehovah gọi Moses từ Lều Hội Kiến và phán với ông. Việc phần mở đầu của sách còn lại có nghĩa là cuộn giấy này là phần bắt đầu của một bản sách Lê-vi.

Cụm từ hai cột cũng có ý nghĩa. Cuộn giấy cổ thường chia văn bản thành các ô dài theo chiều dọc để viết. Khi viết đầy một ô, người viết chuyển sang ô bên cạnh. Việc xuất hiện hai cột có nghĩa là hai ô như vậy nối tiếp nhau đã hiện ra.

Thứ được đọc ra không chỉ là chữ. Trên màn hình còn lưu cả độ dày của nét do người chép ấn bút viết. Đó là chứng cứ cho thấy đây là bản sao được con người chép tay.

Lấp đầy khoảng trống một nghìn năm

Nhóm nghiên cứu cũng đo tuổi của cuộn giấy. Phương pháp định tuổi bằng carbon phóng xạ cho ra kết quả từ thế kỷ 3 đến thế kỷ 4 sau Công nguyên[1]. Phương pháp này đo sự biến đổi của carbon còn lại trong di vật để biết tuổi. Kết quả này lấp một khoảng trống lâu đời trong nghiên cứu bản chép Kinh Thánh cổ.

Trước hết hãy xem khoảng trống đó là gì. Các bản chép Kinh Thánh thuộc Cuộn Biển Chết tìm thấy trong hang Qumran có niên đại từ thế kỷ 3 trước Công nguyên đến thế kỷ 1 sau Công nguyên. Ngược lại, bản chép Kinh Thánh Hebrew chuẩn còn lại ở dạng hoàn chỉnh có niên đại khoảng năm 1000 sau Công nguyên. Giữa hai nhóm này có một khoảng trống gần một nghìn năm. Trong giai đoạn này, rất hiếm bản chép cho biết văn bản đã được truyền lại như thế nào.

Khi khoảng trống lớn, nghi ngờ cũng lớn. Lý do là không có cách nào xác nhận bản chép thời trung cổ có chép nguyên văn bản cổ hay không. Một số học giả từng cho rằng người chép thời trung cổ đã chỉnh sửa văn bản.

Cuộn giấy Ein Gedi xuất hiện ngay giữa khoảng trống này. Và văn bản đọc được hoàn toàn giống phần văn bản phụ âm của Kinh Thánh Hebrew thời trung cổ[1]. Tiếng Hebrew chỉ ghi chữ phụ âm, không ghi nguyên âm. Văn bản chỉ ghi các phụ âm đó được gọi là văn bản phụ âm. Văn bản trung cổ này được gọi là văn bản Masoretic. Đây là văn bản chuẩn mà các người chép Do Thái giáo đã đặt quy tắc và gìn giữ. Ngay cả vị trí chia đoạn cũng giống nhau[2]. Kết quả này cho thấy văn bản đã được truyền lại nguyên vẹn qua thời gian dài.

Người chép Do Thái giáo được biết là tuân thủ nghiêm ngặt các quy tắc khi chép Kinh Thánh. Cách đếm chữ để kiểm tra cũng là một trong các quy tắc đó. Cuộn giấy Ein Gedi trở thành vật chứng cho thấy quy tắc ấy đã thật sự được giữ. Một bản chép đã bổ sung thêm một tư liệu vào cuộc tranh luận lâu dài.

Bản chép này còn có thêm một ghi nhận. Trong số các bản chép Ngũ Thư Moses từng được tìm thấy trong hòm thánh, đây là bản có niên đại sớm nhất cho đến nay[1]. Ngũ Thư Moses là tên gọi chung năm quyển đầu của Cựu Ước.

Điều đã xác định và điều còn lại

Cần tách riêng điều đã được làm sáng tỏ và điều vẫn đang được bàn luận.

Điều đã được làm sáng tỏ là như sau. Văn bản đọc được là chương 1 và chương 2 của sách Lê-vi. Kết quả định tuổi bằng carbon chỉ về thế kỷ 3 đến thế kỷ 4 sau Công nguyên. Văn bản phụ âm giống văn bản chuẩn thời trung cổ. Mật độ của mực cao hơn da, nên có vẻ mực đã được pha thành phần kim loại[1].

Những vấn đề còn lại cũng rõ ràng. Thành phần của mực chưa được phân tích trực tiếp[1]. Lý do là không có bề mặt lộ ra ngoài để lấy mẫu. Cũng có ý kiến định niên đại khác dựa trên kiểu chữ. Nhà cổ văn tự học Ada Yardeni cho rằng bản này thuộc nửa sau thế kỷ 1 hoặc đầu thế kỷ 2 sau Công nguyên. Nhà cổ văn tự học là học giả so sánh kiểu chữ cổ để đoán niên đại. Về điểm này, Drew Longacre phản bác rằng định tuổi bằng carbon đáng tin hơn[2]. Hai quan điểm này vẫn chưa được giải quyết.

Thời gian bảo quản cũng là vấn đề chưa được giải. Cuộn giấy được tạo ra trong thế kỷ 3 đến thế kỷ 4 sau Công nguyên. Hội đường bị cháy vào khoảng năm 600 sau Công nguyên[1]. Như vậy, bản chép này đã nằm trong hòm thánh hơn hai trăm năm. Vì sao cộng đồng tiếp tục giữ một bản chép cũ như vậy vẫn chưa được làm rõ. Chỉ có phỏng đoán rằng họ đã trân trọng bản chép được truyền lại.

Phạm vi đọc được cũng chỉ là một mảnh của cuộn giấy. Phần được mở ra là năm lớp bên ngoài[1]. Ở sâu bên trong, các lớp bị ép chặt hơn nhiều. Vấn đề không thể tách lớp ở những đoạn bị ép vẫn còn tồn tại vào năm 2026. Đây là bài toán mà các nhà nghiên cứu xử lý cuộn giấy Herculaneum đến nay vẫn đang bám sát[3].

Muốn đọc phần bên trong, hai điều phải cùng được cải thiện. Cần thang đo chụp có thể chia lớp nhỏ hơn. Phép tính tách các lớp đã dính vào nhau cũng phải tốt hơn. Khi hai điều kiện này được đáp ứng, cuộn giấy Ein Gedi có thể được đưa lên bàn chụp một lần nữa. Điều đó có thể làm được vì di vật vẫn còn nguyên.

Các phương pháp hỗ trợ chính việc định tuổi cũng đang tăng lên. Năm 2025, một mô hình học máy học kiểu chữ của các bản chép Hebrew cổ để ước tính niên đại đã được công bố. Mô hình này học cả kết quả định tuổi bằng carbon và dữ liệu kiểu chữ. Đó là cách đo cả những độ cong của nét mà mắt người khó nhận ra[4]. Khi các công cụ như vậy tích lũy thêm, tranh luận về niên đại của cuộn giấy Ein Gedi cũng có thể được xem xét lại.

Con đường mà mở cuộn ảo mở ra không chỉ dừng ở bản chép Kinh Thánh. Cùng phương pháp này đã được dùng cho những lá thư cổ bị gấp và niêm phong. Chụp và tính toán cũng đi vào việc xử lý sách và văn bản dễ vỡ. Nguyên tắc đọc mà không chạm đang trở thành nền tảng trong xử lý tư liệu.

Hình ảnh do máy tính đưa ra tự nó không phải là đáp án cuối cùng. Trong nghiên cứu năm 2016, việc máy tính làm chỉ dừng ở mở các lớp và tạo hình ảnh. Người đọc chữ và gắn nghĩa cho chữ trong hình ảnh đó là con người. Vì vậy, nhà cổ văn tự học phải xem lại từng chữ một. Điều sự kiện này để lại không chỉ là một văn bản. Một phương pháp xử lý di vật sẽ vỡ nếu chạm vào đã ra đời. Đó là trình tự lấy số liệu bằng chụp, mở bằng tính toán, rồi để con người kiểm tra. Trình tự này cũng được dùng nguyên như vậy cho nhiều di vật sẽ xuất hiện về sau.

Sách Lê-vi đọc được ở Ein Gedi cũng đã trải qua quá trình kiểm tra đó và trở thành ghi nhận của giới học thuật. Thứ phá vỡ 45 năm im lặng là công việc chung của máy móc và con người. Thiết bị chụp đo mật độ, còn phép tính mở các lớp. Và con người đọc chữ rồi gắn nghĩa. Chương sau sẽ tiếp tục với câu chuyện quan sát chính hình dạng của chữ.

Tài liệu tham khảo

[1] Seales, W. B., Parker, C. S., Segal, M., Tov, E., Shor, P., & Porath, Y. (2016). From damage to discovery via virtual unwrapping: Reading the scroll from En-Gedi. Science Advances, 2(9), e1601247. https://doi.org/10.1126/sciadv.1601247

[2] En-Gedi Scroll. (2026, 9월 3일). Wikipedia. https://en.wikipedia.org/wiki/En-Gedi_Scroll (truy cập ngày 3 tháng 9 năm 2026)

[3] Vesuvius Challenge. (2026, 7월 10일). Open Problems: Why Reading Every Herculaneum Scroll Is Still a Challenge. scrollprize.org. https://scrollprize.org/2026_open_problems (truy cập ngày 3 tháng 9 năm 2026)

[4] Popović, M., Dhali, M. A., Schomaker, L., van der Plicht, J., Rasmussen, K. L., La Nasa, J., Degano, I., Colombini, M. P., & Tigchelaar, E. (2025). Dating ancient manuscripts using radiocarbon and AI-based writing style analysis. PLOS ONE, 20(6), e0323185. https://doi.org/10.1371/journal.pone.0323185

Chương 3. Hai bóng dáng ẩn trong Cuộn Isaiah: giám định người chép

3.1. Hồ sơ vụ việc

Cuộn da đi ra từ hang động

Năm 1947, một cuộn da cổ được phát hiện trong hang động ở vùng Qumran, phía tây Biển Chết. Những người Bedouin chăn dê đã tìm thấy chum và cuộn giấy trong hang trên vách đá. Những di vật này được gọi là Cuộn Biển Chết. Sau đó, việc khai quật tiếp tục ở nhiều hang động. Đến nay, Cuộn Biển Chết được sắp xếp thành khoảng 930 bản chép. Sách được chép bằng tay gọi là bản chép. Nếu tính từng tờ, số mảnh da và giấy cói vượt quá vài nghìn tờ[1]. Niên đại ghi chép được xem là từ thế kỷ 3 trước Công nguyên đến thế kỷ 1 sau Công nguyên[2].

Biển Chết là một hồ có độ mặn cao. Xung quanh là sa mạc khô cằn. Mưa hầu như không rơi và không khí khô. Nhiệt độ trong hang cũng không dao động lớn. Đây là lý do da và giấy cói không mục nát mà còn tồn tại. Rất hiếm nơi còn giữ được một tập văn bản cổ có quy mô như vậy.

Trong số các di vật này, có một cuộn giấy thu hút ánh mắt của các học giả. Đó là Cuộn Isaiah Lớn. Ký hiệu được dùng trong giới học thuật là 1QIsaa. Điều này có nghĩa là bản chép Isaiah thứ nhất lấy từ hang số 1 ở Qumran. Đây là bản chép bằng tiếng Hebrew trên da thú vào thế kỷ 2 trước Công nguyên. Chiều dài khoảng 7,3 mét. Nó chứa gần như toàn bộ nội dung sách Isaiah trong Cựu Ước. Văn bản được chia thành 54 cột dọc dài[2]. Cột này được gọi là column trong thuật ngữ học thuật. Đây là bản chép Kinh Thánh được bảo tồn tốt nhất trong số các bản lấy từ các hang ở Biển Chết.

Cuộn giấy được làm bằng cách nối nhiều tấm da với nhau bằng chỉ. Người chép viết vài cột trên một tấm rồi chuyển sang tấm tiếp theo. Cuộn giấy viết xong được cuộn tròn để cất giữ. Khi đọc, người ta mở một bên ra và cuộn bên kia lại. Tên gọi cuộn giấy bắt nguồn từ cách đó.

Giá trị của bản chép này đến từ tuổi đời của nó. Cuộn Biển Chết chứa những bản chép Kinh Thánh Hebrew cổ nhất còn lại đến nay[3]. Cuộn Isaiah Lớn cũng là một trong số đó. Nhờ vậy, người ta có tư liệu để đối chiếu cách văn bản Kinh Thánh được truyền lại qua thời gian dài. Nhưng về người tạo ra cuộn giấy này thì không có thông tin rõ ràng.

Giới học thuật chia rẽ về câu hỏi ai đã viết

Đọc cuộn giấy và xác định ai đã viết là hai việc khác nhau. Chữ thì đọc được, nhưng bản chép này không có tên người chép. Người phụ trách chép lại văn bản được gọi là người chép. Cuộn giấy này không có chữ ký của người chép và cũng không có ngày tháng[2]. Vì vậy, các học giả phải nhìn hình dạng chữ để ước lượng số người.

Về vấn đề này, giới học thuật đứng trước hai hướng. Một phía cho rằng một người chép đã viết một mình từ đầu đến cuối. Cơ sở là hình dạng chữ trên toàn bộ cuộn giấy khá đều. Phía kia cho rằng có từ hai người trở lên chia nhau viết. Cơ sở là khi đi qua phần giữa, hình dạng nét chữ dần thay đổi đôi chút.

Cả hai lập luận đều có một vấn đề khó. Chữ do một người viết cũng không giống hệt từ đầu đến cuối. Khi tay mỏi, nét chữ lỏng hơn. Khi thay bút, độ dày nét thay đổi. Nếu không biết độ lớn của dao động này, không thể phán đoán người viết có thay đổi hay không. Cả hai lập luận đều dựa vào ấn tượng bằng mắt. Không có quả cân để phân định bên nào đúng.

Người ta cho rằng ở Qumran từng có một cộng đồng làm và bảo vệ các cuộn giấy. Họ đặt cuộn giấy vào chum rồi giấu trong hang. Có vẻ họ muốn bảo vệ văn bản khi chiến tranh đến gần. Tên của những người đã giấu chúng cũng không còn lại. Tìm bàn tay đã làm ra cuộn giấy cũng là con đường đến gần họ.

Cuộc tranh luận này không dừng ở chuyện chữ viết. Nếu nhiều người cùng chia nhau làm một bản chép, điều đó có nghĩa là từng có một tổ chức nơi các người chép làm việc cùng nhau. Nó liên quan đến việc làm rõ văn bản Kinh Thánh đã đi qua bàn tay những ai để đến ngày nay[2]. Đây là lý do các nhà nghiên cứu Cuộn Biển Chết giữ câu hỏi này trong thời gian dài.

Giới hạn của giám định bằng mắt

Ngành nghiên cứu chữ viết cổ được gọi là cổ văn tự học. Đây là cách học giả cầm kính lúp và quan sát hình dạng chữ. Họ đo chiều cao, độ nghiêng và độ dày của nét chữ rồi chép lại thành hình. Họ cũng xếp chữ viết của nhiều bản chép theo trình tự thời gian để nắm dòng biến đổi. Trong thời gian dài, phương pháp này là công cụ giám định duy nhất.

Phương pháp này có điểm yếu. Nó dựa nhiều vào mắt và kinh nghiệm của học giả. Cùng nhìn một chữ, các học giả thường đưa ra đánh giá khác nhau[2]. Mắt người không đo được khác biệt trị số rất nhỏ. Lực nhấn bút sậy hay độ dày của vết mực loang rất khó ước lượng bằng mắt. Trí nhớ cũng có giới hạn. Không dễ đặt hai chữ cách nhau vài mét cạnh nhau trong đầu để so sánh. Cuối cùng, tranh cãi chỉ lặp lại.

Một bức tường khác là cách đào tạo của thời đó. Người muốn trở thành người chép được rèn luyện để chép y nguyên chữ của thầy. Họ học cách không bộc lộ cá tính riêng. Kết quả là chữ của những người chép xuất thân từ cùng một trường nhìn bên ngoài gần như giống hệt nhau. Với học giả con người, đây là một lớp sương dày. Họ không thể phân biệt khác biệt rất nhỏ ở đầu và cuối cuộn giấy là thay đổi tình trạng của một người hay là do đổi người[2].

Nếu giám định lệch, phán đoán đi theo cũng lệch. Nếu người chép chỉ có một, một bản chép chứa thời gian của một người. Nếu có hai người, điều đó có nghĩa là từng có một tổ chức chia việc. Quy mô và tính chất của nơi làm bản chép sẽ thay đổi. Nhiều kết luận treo trên một lần giám định chữ viết.

Có thể hình dung những học sinh trong lớp cùng chép chính tả theo một mẫu chữ. Tình hình cũng tương tự. Những cuốn vở mà tất cả cùng cố viết ngay ngắn nhìn bên ngoài rất khó phân biệt. Dù vậy, thói quen bẻ nét vẫn khác nhau đôi chút ở mỗi học sinh. Vấn đề là khi đó không có thước để đo khác biệt ấy.

Tình hình thay đổi khi ảnh bản chép được công bố dưới dạng số. Cuộn giấy thật yếu trước ánh sáng và độ ẩm nên không thể mở ra thường xuyên. Nếu có ảnh, có thể quan sát bao nhiêu lần cũng được mà không chạm vào di vật. Cơ quan Cổ vật Israel cùng Google đã tạo bộ tư liệu số về Cuộn Biển Chết. Họ dùng thiết bị gọi là MegaVision để chiếu ánh sáng ở nhiều bước sóng và chụp di vật. Ảnh ghi cả cận hồng ngoại, loại ánh sáng mắt người không thấy, để làm sống lại chữ đã mờ. Mục tiêu là đạt chất lượng ảnh gần với việc nhìn trực tiếp hiện vật[1].

Khi ảnh độ phân giải cao mà máy tính có thể đọc được tích lũy, con đường cho công cụ mới mở ra. Ảnh là tập hợp các con số. Một điểm nhỏ tạo nên màn hình gọi là điểm ảnh. Mỗi điểm ảnh có giá trị độ sáng. Điều đó có nghĩa là có thể đo độ dày và góc của nét bằng con số. Việc chuẩn bị để đổi giám định bằng mắt thành tính toán đã hoàn tất. Từ đây, thám tử AI bước vào.

Tài liệu tham khảo

[1] Israel Antiquities Authority. (n.d.). The Digital Library. The Leon Levy Dead Sea Scrolls Digital Library. https://www.deadseascrolls.org.il/about-the-project/the-digital-library (truy cập ngày 3 tháng 9 năm 2026)

[2] Popović, M., Dhali, M. A., & Schomaker, L. (2021). Artificial intelligence based writer identification generates new evidence for the unknown scribes of the Dead Sea Scrolls exemplified by the Great Isaiah Scroll (1QIsaa). PLOS ONE, 16(4), e0249769. https://doi.org/10.1371/journal.pone.0249769

[3] University of Groningen. (2021, 4월 21일). Cracking the code of the Dead Sea Scrolls. rug.nl. https://www.rug.nl/news/2021/04/cracking-the-code-of-the-dead-sea-scrolls?lang=en (truy cập ngày 3 tháng 9 năm 2026)

3.2. Truy vết của đội điều tra khoa học: dấu vân tay sinh học còn lại trong nét chữ

Đội điều tra được lập ra

Nhóm nghiên cứu của Đại học Groningen ở Hà Lan đã nhận vụ việc này. Nhà cổ văn tự học Mladen Popović dẫn dắt nhóm. Các nhà nghiên cứu AI Lambert Schomaker và Maruf Dhali cùng tham gia[1]. Năm 2017, họ trước hết công bố phương pháp đo nét chữ của Cuộn Biển Chết bằng máy tính[2]. Kết quả họ tinh chỉnh phương pháp đó rồi nhắm vào Cuộn Isaiah Lớn được công bố năm 2021[1, 5].

Mục tiêu của nhóm nghiên cứu rất rõ. Đó là biến đánh giá ấn tượng của con người thành con số. Trước hết, họ đo độ lớn của dao động sinh ra trong chữ do cùng một người viết. Sau đó, họ xét xem khác biệt giữa phần trước và phần sau cuộn giấy có lớn hơn dao động ấy hay không. Nếu khác biệt lớn hơn, điều đó có nghĩa là người khác đã viết.

Phương pháp tính đặc trưng được tạo năm 2017 đã được dùng để so sánh giữa các bản chép[2]. Nhóm nghiên cứu áp dụng lại phương pháp đó, lần này tập trung vào một bản là Cuộn Isaiah Lớn. Tìm vị trí đổi người chép trong cùng một bản khó hơn so sánh giữa các bản chép. Vì khác biệt chữ viết nhỏ hơn nhiều.

Cách này có một điểm tốt. Nếu ghi lại quy trình tính toán, ai cũng nhận được cùng một đáp án từ cùng một ảnh. Không cần dựa vào con mắt của học giả. Nếu kết quả sai, có thể lần lại xem sai ở bước nào. Ngay cả khi tiếp tục tranh luận về kết quả giám định, căn cứ tranh luận vẫn còn lại dưới dạng con số.

Bước đầu tiên xóa vết bẩn

Công việc đầu tiên là chỉ để lại mực trong ảnh. Việc này được gọi là nhị phân hóa. Đó là xử lý chia mọi điểm ảnh thành một trong hai loại, mực hoặc nền. Nói cách khác, ảnh được tách thành chữ đen và nền sáng.

Có lý do cần bước này. Cuộn giấy đã ở trong hang khoảng 2.000 năm. Trên bề mặt da có vết bẩn, dấu mốc và vết nứt nhỏ. Với máy tính, các dấu này và nét mực đều là điểm ảnh tối. Nếu giữ nguyên nền, máy sẽ học thớ da thay vì chữ viết. Khi đó, nó sẽ phân biệt loại da thay cho người chép, làm giám định lệch[1].

Nhóm nghiên cứu đã tạo riêng một mạng nơ-ron phụ trách việc này. Tên của nó là BiNet. Mạng nơ-ron là cấu trúc tính toán mô phỏng kết nối tế bào thần kinh trong não người. BiNet nhìn các mẫu được đánh dấu thủ công thành mực và nền rồi học tiêu chuẩn phán đoán đó. Sau khi học xong, BiNet gạt vết bẩn khỏi ảnh mới và chỉ để lại nét chữ.

Ở đây có một quy tắc phải giữ. Không được chạm vào độ dày và hình dạng rìa của nét gốc. Vì chính rìa nét là nơi chứa thói quen của người chép. Nếu trong lúc xóa tạp chất mà làm mịn cả nét chữ, chứng cứ sẽ biến mất. BiNet được thiết kế theo hướng giữ nguyên dấu mực[1]. Công nghệ này xuất hiện trước dưới dạng bản thảo công khai năm 2019[3]. Bài báo sắp xếp lại thiết kế và hiệu năng được đăng trong tuyển tập hội nghị quốc tế về nhận dạng tài liệu năm 2026[4].

Thước đo thói quen của toàn bộ chữ viết

Từ chữ đã được làm sạch, nhóm rút ra hai loại đặc trưng. Loại thứ nhất là đặc trưng về kết cấu. Loại thứ hai là đặc trưng về hình dạng.

Đặc trưng về kết cấu được gọi là hinge. Hinge là từ chỉ bản lề cửa. Ở mỗi chỗ nét bị gập, máy đo góc của hai hướng. Sau đó, nó lập bảng xem từng cặp góc xuất hiện thường xuyên đến mức nào. Bảng này chứa thói quen xoay cổ tay khi viết dưới dạng con số. Nó có thể tính toán mà không cần nhận biết từng chữ[1].

Đặc trưng về hình dạng được gọi là fraglet. Từ này chỉ một mảnh ngắn cắt ra từ nét chữ. Máy cắt nét chữ và gom các mảnh thành danh mục. Nhóm nghiên cứu tạo một danh mục chuẩn bằng các mảnh lấy từ nhiều bản chép. Khi chữ mới đi vào, máy đếm xem mảnh đó giống mục nào trong danh mục. Dạng nét cơ bản mà con người học khi học chữ hiện ra ở đây[1].

Hai đặc trưng này có điểm chung. Chúng có thể tính toán dù không biết nghĩa của chữ. Máy tính không đọc được tiếng Hebrew vẫn có thể đo góc và hình cong của nét. Dù có lẫn chữ bị hư hại, nó vẫn rút giá trị từ phần nét còn lại. Đây là tính chất rất hữu ích khi xử lý văn bản cổ.

Hai đặc trưng nhìn vào những thứ khác nhau. Hinge nhìn chuyển động của bàn tay. Fraglet nhìn kiểu chữ đã học. Nhóm nghiên cứu dùng từng đặc trưng riêng và cũng dùng kết hợp cả hai. Mục đích là kiểm tra xem các thước đo khác nhau có cho cùng đáp án hay không[1].

Dấu vết còn lại trong một chữ aleph

Hai đặc trưng phía trước đo gộp toàn bộ chữ viết. Lần này là lúc phóng to một chữ để nhìn kỹ.

Nhóm nghiên cứu chọn riêng một chữ cái Hebrew để kiểm tra chính xác. Đó là aleph. Aleph là chữ cái đầu tiên của tiếng Hebrew và xuất hiện thường xuyên nhất trong cuộn giấy. Hình dạng của nó gồm một nét chéo và hai nét ngắn gắn vào.

Có lý do chọn aleph. Mẫu càng nhiều thì thống kê càng vững. Chỉ chữ xuất hiện thường xuyên mới có thể thu đều ở vùng trước và vùng sau. Trên toàn bộ cuộn giấy có 5.011 chữ aleph. Nhóm nghiên cứu dùng máy tính chọn ra 758 chữ trong số đó để phân tích[1]. Họ chỉ chọn những chữ còn nguyên vẹn, không bị đứt nét.

Các chữ aleph được chọn được chia thành hai nhóm. Một nhóm đến từ phần trước cuộn giấy. Nhóm kia đến từ phần sau. Họ chồng ảnh của mỗi nhóm vào cùng một vị trí để tạo hình dạng trung bình. Cách này giống như đặt nhiều tờ giấy mờ chồng lên nhau rồi nhìn từ trên xuống. Nét thường chồng lên nhau sẽ đậm hơn. Nét khác nhau theo từng người sẽ mờ hơn. Nếu lấy hai hình trung bình trừ cho nhau, phần còn lại chỉ là khác biệt. Có thể nhìn bằng màu xem khác biệt ấy tập trung ở nét nào[1].

Các điểm tách thành hai nhóm

Đặc trưng đã chuyển thành con số là một danh sách dài gồm hàng trăm giá trị. Con người không thể so sánh các danh sách như vậy bằng mắt. Vì vậy, nhóm nghiên cứu rút danh sách xuống còn hai giá trị rồi chấm lên mặt phẳng. Chữ có giá trị giống nhau được rút lại để tụ gần nhau. Chữ khác nhau được đặt xa hơn. Đây là hình được tạo bằng cách lặp lại phép tính dời vị trí các điểm từng chút một.

Trong 54 cột, tất cả các cột có ảnh nguyên vẹn được đặt lên mặt phẳng này. Cột 16 và cột 46 bị loại vì không có ảnh dùng được. Khi các điểm được chấm lên, hai nhóm xuất hiện. Các cột phía trước tạo thành một nhóm. Các cột phía sau tạo thành nhóm khác[1]. Giữa hai nhóm có một khoảng cách nhìn thấy được.

Nhóm nghiên cứu kiểm chứng xem sự tách đôi này có phải ngẫu nhiên hay không. Họ dùng các phương pháp thống kê như kiểm định chi bình phương và kiểm định t. Kiểm định chi bình phương là phép tính xem một khác biệt có phải do ngẫu nhiên hay không. Kiểm định t là phép tính xem chênh lệch trung bình giữa hai nhóm. Cả hai phương pháp đều tính xác suất để một hình dạng như vậy xuất hiện một cách ngẫu nhiên. Nếu xác suất đó rất nhỏ thì không xem là ngẫu nhiên. Chỉ với dao động trong chữ viết của một người thì không tạo ra khác biệt lớn như vậy[1].

Điều quan trọng là nhiều thước đo cùng cho một đáp án. Đo bằng hinge hay bằng fraglet thì ranh giới vẫn xuất hiện ở cùng một chỗ. Trong hình trung bình của aleph, sự chia tách cũng nằm ở cùng nơi. Ranh

giới nằm giữa cột số 27 và cột số 28[1]. Nếu các phương pháp khác nhau cùng chỉ vào một điểm, ranh giới đó không phải do vết nhiễu trong ảnh. Máy đã kẻ ra một đường mà mắt người bỏ sót.

Nhóm nghiên cứu cũng xem xét khả năng phản bác. Kết quả có thể xuất hiện vì độ sáng của ảnh hoặc tình trạng da khác nhau giữa phần trước và phần sau. Nhưng sau khi dùng nhị phân hóa để loại bỏ nền, ranh giới vẫn còn nguyên[1]. Điều đó có nghĩa nguyên nhân nằm ở chính chữ viết, chứ không phải ở da.

Tài liệu tham khảo

[1] Popović, M., Dhali, M. A., & Schomaker, L. (2021). Artificial intelligence based writer identification generates new evidence for the unknown scribes of the Dead Sea Scrolls exemplified by the Great Isaiah Scroll (1QIsaa). PLOS ONE, 16(4), e0249769. https://doi.org/10.1371/journal.pone.0249769

[2] Dhali, M. A., He, S., Popović, M., Tigchelaar, E., & Schomaker, L. (2017). A digital palaeographic approach towards writer identification in the Dead Sea Scrolls. Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, 693-702. https://doi.org/10.5220/0006249706930702

[3] Dhali, M. A., de Wit, J. W., & Schomaker, L. (2019). BiNet: Degraded-manuscript binarization in diverse document textures and layouts using deep encoder-decoder networks. arXiv:1911.07930. https://arxiv.org/abs/1911.07930 (truy cập ngày 3 tháng 9 năm 2026)

[4] Dhali, M. A., de Wit, J. W., & Schomaker, L. (2026). BiNet: A deep encoder-decoder network for binarizing degraded ancient manuscripts. In Document Analysis and Recognition - ICDAR 2025 Workshops (Lecture Notes in Computer Science, pp. 166-193). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-09371-4_11

[5] University of Groningen. (2021, 4월 21일). Cracking the code of the Dead Sea Scrolls. rug.nl. https://www.rug.nl/news/2021/04/cracking-the-code-of-the-dead-sea-scrolls?lang=en (truy cập ngày 3 tháng 9 năm 2026)

3.3. Sự thật được hé lộ

Bàn tay thay đổi ở cột số 27

Kết quả phân tích ủng hộ giả thuyết có nhiều thư lại. Cuộn Isaiah lớn do hai người chia nhau viết. Người thứ nhất viết từ cột số 1 đến cột số 27. Người thứ hai viết tiếp từ cột số 28 đến cột số 54[1].

Ranh giới này cũng khớp với chứng cứ khác. Phần dưới của cột số 27 để trống một khoảng cao bằng ba dòng. Cột số 28 bắt đầu ở một tấm da mới được nối vào. Đó cũng là nơi một chương mới của Sách Isaiah bắt đầu[1]. Đây là một điểm nối thích hợp để bàn giao công việc. Ranh giới thống kê mà máy tính tìm ra nằm đúng nơi có ranh giới vật lý của cuộn giấy.

Trước đây đã có phỏng đoán rằng thư lại đã thay đổi sau cột số 27. Đó là ý kiến dựa trên nơi các tấm da được nối và nơi văn bản được chia. Nhưng ý kiến này chưa trở thành đồng thuận trong giới học thuật. Phân tích AI đã gắn con số vào phỏng đoán ấy[2].

Ý nghĩa của phát hiện này rất lớn. Đã có căn cứ bằng số liệu cho thấy một bản chép Kinh Thánh được nhiều người nối tiếp nhau tạo ra[1]. Hai thư lại không để lại tên. Họ cũng không để lại chữ ký. Thứ họ để lại chỉ là thói quen của bàn tay khi chuyển động. Hai nghìn năm sau, máy đã đọc được thói quen đó. Popović nói rằng có lẽ ta sẽ không bao giờ biết tên họ, nhưng qua nét chữ, ông có cảm giác như đang bắt tay họ[2].

Hai người học cùng một trường

Chữ viết của hai thư lại vẫn rất giống nhau. Ngay cả sau khi máy đo, khác biệt vẫn rất nhỏ. Nhóm nghiên cứu đọc chính sự giống nhau này như một chứng cứ khác. Nó có nghĩa cách viết của hai người gần nhau[1]. Nhóm nghiên cứu cho rằng hai thư lại được đào tạo ở cùng một nơi hoặc xuất thân từ cùng một hệ thống[2].

Từ đây có thể hình dung diện mạo của trường đào tạo thư lại thời cổ. Người học luyện kiểu chữ được công nhận theo từng nét. Muốn trở thành thư lại giỏi, họ phải nén cá tính riêng. Nhờ kỷ luật đó, chữ viết trong một cuộn giấy được giữ khá đều.

Tuy vậy, không thể xóa hết thói quen của cơ thể. Góc cầm bút sậy mỗi người mỗi khác. Tốc độ di chuyển tay và lực ấn cũng khác nhau. Những khác biệt này còn lại trong góc gãy của nét và hình dạng mép nét. Mắt người không thể gom các khác biệt rất nhỏ này để phán đoán. Máy tính gom hàng nghìn giá trị rồi tính trung bình và độ phân tán. Dấu vết cá nhân bị kỷ luật che phủ đã hiện ra như vậy[1].

Kết quả này cho thấy việc tạo bản chép là một công việc chung. Để làm một cuộn giấy dài cần nhiều tháng. Không có lý do bắt buộc một người phải giữ việc đó đến cuối. Từ đây xuất hiện khả năng công việc được chia ra và bàn giao cho nhau[1]. Văn bản Kinh Thánh đã đi qua những bàn tay như thế để truyền đến hôm nay.

Điều máy vẫn chưa làm được

Thành quả này không có nghĩa AI là vạn năng. Công cụ tiếp theo của cùng nhóm nghiên cứu cho thấy giới hạn đó.

Năm 2025, nhóm nghiên cứu công bố một mô hình ước tính niên đại tên là Enoch. Enoch là mô hình nhìn vào hình dạng chữ viết để đoán thời điểm một bản chép được viết. Định tuổi bằng carbon phóng xạ là phương pháp đo tuổi bằng lượng carbon còn lại trong đồ vật cổ. Enoch học từ 24 mẫu đã được định tuổi bằng phương pháp đó. Nhóm cũng công bố kết quả áp dụng với 135 bản chép chưa biết niên đại. Sai số dự đoán trung bình nằm trong khoảng 27,9 năm đến 30,7 năm. Sau khi các nhà cổ văn tự học xem xét dự đoán này, 79 phần trăm được đánh giá là hợp lý[3].

Tranh luận nổ ra ở 21 phần trăm còn lại. Nhà cổ văn tự học Christopher Rollston nêu một trường hợp để phản bác. Enoch xem bản chép 4Q114 chứa Sách Daniel thuộc khoảng từ năm 230 trước Công nguyên đến năm 160 trước Công nguyên. Rollston cho rằng các chương 7 đến 12 của Sách Daniel chỉ về một sự kiện. Đó là việc Antiochos IV làm ô uế Đền thờ Jerusalem vào năm 167 trước Công nguyên. Văn bản ấy không thể được viết trước sự kiện. Nhưng khoảng niên đại mà Enoch đưa ra phần lớn lại sớm hơn năm 167. Rollston chỉ ra rằng niên đại của Enoch quá sớm. Ông nói Enoch chỉ nên là một trong nhiều công cụ trong hộp công cụ[4].

Cuộc tranh luận này cho ta biết một điều. Máy chỉ nhìn hình dạng của chữ. Máy không đọc được sự kiện lịch sử được ghi trong văn bản hay ý nghĩa của câu. Vì vậy, đáp án của máy không phải kết luận cuối cùng mà là một ứng viên. Sau khi thu hẹp các ứng viên, nhà sử học phải đối chiếu văn cảnh để phán định. Người quyết định phải đo chữ nào cũng là nhà cổ văn tự học. Phán đoán chọn aleph đến từ kinh nghiệm lâu năm quan sát chữ viết cổ.

Số lượng dữ liệu học cũng là trở ngại. Enoch học từ 24 mẫu đã biết niên đại[3]. Với di vật cổ, rất khó tăng số mẫu. Vì để định tuổi, phải lấy một phần nhỏ của di vật rồi đốt nó. Mô hình học từ ít mẫu sẽ dao động khi gặp thứ vượt ra ngoài phạm vi đã học. Việc giám định thư lại cũng có tình hình tương tự. Những chữ xuất hiện thường xuyên có đủ mẫu, nhưng chữ hiếm chỉ có vài trường hợp. Ở nơi có ít mẫu, phán đoán của máy mất sức mạnh.

Vì vậy, nhóm nghiên cứu cũng không đặt AI làm quan tòa duy nhất. Nơi tấm da được nối, nơi văn bản được chia, và ý kiến mà các học giả đã tích lũy đều được đặt cạnh nhau. Khi nhiều chứng cứ cùng chỉ về một nơi, kết luận mới đứng vững. Đáp án trong Cuộn Isaiah lớn cũng xuất hiện theo cách đó. Máy chỉ trả lời bằng con số cho câu hỏi do con người đặt ra.

Mục tiêu điều tra tiếp theo

Nghiên cứu về Cuộn Isaiah lớn dẫn đến một kế hoạch lớn hơn. Hội đồng Nghiên cứu châu Âu quyết định trao cho Popović khoản tài trợ Advanced Grant vào ngày 23 tháng 6 năm 2026. Kinh phí là 2,5 triệu euro và thời hạn là 5 năm[5]. Kế hoạch mới có tên là Tracing Scribes and Scrolls[6]. Tên này có nghĩa là lần theo dấu vết của thư lại và cuộn giấy.

Cơ quan Cổ vật Israel cho biết vào ngày 30 tháng 6 năm 2026 rằng họ tham gia kế hoạch này. Đại học Pisa, Đại học Napoli Federico II, Đại học Nam Đan Mạch và Đại học Leuven cũng cùng tham gia. Các bảo tàng Ai Cập ở Berlin và Torino cũng có tên trong danh sách[6].

Câu hỏi lần này chuyển từ ai đã viết sang được làm ở đâu. Nhóm nghiên cứu khảo sát khoảng 250 mẫu Cuộn Biển Chết. Họ đo thành phần hóa học của da, giấy cói và mực. Sau đó họ đối chiếu các thành phần đó với thành phần của giấy cói Ai Cập cổ đại. Nếu biết vật liệu đến từ đâu, ta có thể thu hẹp nơi tạo ra bản chép. AI đảm nhận việc sắp xếp dữ liệu hóa học này và tìm quy luật ẩn trong chữ viết[7]. Popović nói rằng nếu kết hợp phân tích trong phòng thí nghiệm với nghiên cứu chữ viết cổ, ta có thể tiến gần những câu hỏi trước đây chưa chạm tới[6].

Phương pháp này không chỉ dùng cho Cuộn Isaiah lớn. Trong Cuộn Biển Chết vẫn còn nhiều bản chép chưa biết người chép. Nếu gom các bản chép do cùng một bàn tay viết, ta sẽ có danh mục công việc của một thư lại. Khi những danh mục như vậy tích lũy, ta có thể hình dung các thư lại đã liên hệ với nhau như thế nào[1]. Nghiên cứu tìm ra hai người này đã trở thành điểm khởi đầu.

Một thành quả khác mà nghiên cứu này để lại là chính phương pháp. Trình tự đã được sắp xếp rõ. Trước hết loại bỏ vết nhiễu khỏi ảnh. Sau đó biến nét chữ thành con số. Cuối cùng kiểm định bằng thống kê. Có thể áp dụng cùng trình tự này cho văn bản cổ bằng ngôn ngữ khác. Các nghiên cứu về bảng đất sét và thẻ gỗ ở các chương sau của cuốn sách này cũng đi theo con đường tương tự.

Cuộc điều tra bắt đầu từ Cuộn Isaiah lớn vẫn chưa kết thúc. Công cụ đã tìm ra bóng dáng của hai người nay lên đường tìm nơi đã làm ra cuộn giấy.

Việc đọc văn bản cổ không kết thúc ở nhận ra chữ. Ta còn phải hỏi ai đã viết chữ đó và viết ở đâu. Cuộn Isaiah lớn là trường hợp đầu tiên trả lời bằng con số cho câu hỏi có mấy người đã viết. Cuộn giấy đã im lặng suốt 2 nghìn năm. Nay con người đọc câu trả lời ấy trong độ run của nét chữ.

Tài liệu tham khảo

[1] Popović, M., Dhali, M. A., & Schomaker, L. (2021). Artificial intelligence based writer identification generates new evidence for the unknown scribes of the Dead Sea Scrolls exemplified by the Great Isaiah Scroll (1QIsaa). PLOS ONE, 16(4), e0249769. https://doi.org/10.1371/journal.pone.0249769

[2] University of Groningen. (2021, 4월 21일). Cracking the code of the Dead Sea Scrolls. rug.nl. https://www.rug.nl/news/2021/04/cracking-the-code-of-the-dead-sea-scrolls?lang=en (truy cập ngày 3 tháng 9 năm 2026)

[3] Popović, M., Dhali, M. A., Schomaker, L., van der Plicht, J., Rasmussen, K. L., La Nasa, J., Degano, I., Colombini, M. P., & Tigchelaar, E. (2025). Dating ancient manuscripts using radiocarbon and AI-based writing style analysis. PLOS ONE, 20(6), e0323185. https://doi.org/10.1371/journal.pone.0323185

[4] Steinmeyer, N. (2025, 6월 20일). Can AI date the Dead Sea Scrolls? Bible History Daily, Biblical Archaeology Society. https://www.biblicalarchaeology.org/daily/biblical-artifacts/dead-sea-scrolls/can-ai-date-the-dead-sea-scrolls/ (truy cập ngày 3 tháng 9 năm 2026)

[5] University of Groningen. (2026, 6월 23일). Mladen Popović awarded ERC Advanced Grant to trace the origins of the Dead Sea Scrolls. rug.nl. https://www.rug.nl/about-ug/latest-news/news/archief2026/nieuwsberichten/mladen-popovi-awarded-erc-advanced-grant-to-trace-the-origins-of-the-dead-sea-scrolls?lang=en (truy cập ngày 3 tháng 9 năm 2026)

[6] The Jerusalem Post. (2026, 6월 30일). New AI-powered research project aims to uncover the origins of the Dead Sea Scrolls. jpost.com. https://www.jpost.com/archaeology/article-900938 (truy cập ngày 3 tháng 9 năm 2026)

[7] Archaeology News Online Magazine. (2026, 6월). Where the Dead Sea Scrolls come from: AI and chemical analysis trace origins. archaeologymag.com. https://archaeologymag.com/2026/06/dead-sea-scrolls-origins-studied-with-ai-and-chemical-analysis/ (truy cập ngày 3 tháng 9 năm 2026)

Chương 4. Chất keo số ghép lại câu đố bằng đất sét và da thuộc gồm hàng chục nghìn mảnh

4.1. Hồ sơ vụ việc

Chiếc chum được phát hiện khi đi tìm dê

Vào mùa xuân năm 1947, một sự việc xảy ra trên vách đá phía tây Biển Chết. Các cậu bé Bedouin đang tìm một con dê bỏ chạy. Các cậu bé thấy một cửa hang hẹp giữa các tảng đá. Trong hang có những chiếc chum đất sét cao đứng bên trong. Trong chum có các cuộn viết chữ trên da[1].

Những cuộn này là Cuộn Biển Chết. Cuộn Biển Chết là các văn bản Do Thái được viết từ khoảng thế kỷ 3 trước Công nguyên đến thế kỷ 1 sau Công nguyên. Trong chín năm sau đó, người ta tìm thêm mười hang gần đó. Tổng cộng có mười một hang chứa các cuộn. Các văn bản tìm thấy ở đây vượt quá 900 tác phẩm. Chúng được viết bằng tiếng Hebrew, tiếng Aram và tiếng Hy Lạp[1]. Phần lớn được viết trên da thuộc làm từ da cừu hoặc da dê được cán mỏng. Cũng có bản chép trên giấy làm từ thân cây gọi là giấy cói[7].

Trong các cuộn có những bản chép Kinh Thánh Hebrew thuộc loại cổ nhất. Cũng có văn bản ghi quy tắc của cộng đồng và văn bản ghi lịch[1]. Những hang động trong hoang mạc khô cằn đã giữ gìn các chữ này trong 2.000 năm.

Gần các hang có di tích một ngôi làng cổ tên là Qumran. Nhiều học giả cho rằng cộng đồng Do Thái sống ở đây đã để lại các cuộn. Họ chép Kinh Thánh và ghi lại quy tắc của mình. Người ta đoán rằng khi quân đội La Mã tiến đến gần, họ đã giấu các cuộn trong hang.

Đống mảnh vỡ ở Hang 4

Tình trạng di vật khác nhau rất lớn tùy từng hang. Các cuộn ở Hang 1 còn lại khá nguyên vẹn trong chum. Hang 4 thì khác. Các cuộn trong hang này không được đặt trong chum mà chất trên nền hang. Chỉ riêng nơi này đã cho ra hàng nghìn mảnh của khoảng 500 cuộn[1].

Độ ẩm trong hang làm da hư hại dần. Côn trùng và thú gặm ăn da. Qua thời gian dài, các cuộn vỡ vụn. Những gì còn lại là một đống mảnh nhỏ bằng móng tay.

Các mảnh do Cơ quan Cổ vật Israel bảo quản vượt quá 25.000 mảnh[2]. Phần lớn các mảnh không ghi chúng rơi ra từ cuộn nào. Việc một mảnh chỉ còn hai hoặc ba chữ cũng rất thường gặp. Các học giả xếp chúng trên những tấm kính lớn và ghép bằng mắt.

Manh mối để ghép chỉ có vài loại. Họ xem hình dạng chữ có giống nhau không. Họ xem màu da có giống nhau không. Họ thử xem mép gãy có ăn khớp không. Với cách này, mỗi lần chỉ có thể so sánh hai hoặc ba mảnh. Khi số mảnh tăng lên đến hàng chục nghìn, số cặp cần so sánh tăng đến mức không kiểm soát được.

Ghép câu đố thủ công mất rất lâu. Mắt người không thể so sánh cùng lúc mép của hàng chục nghìn mảnh. Hơn nữa, mép các mảnh đã mòn và co lại. Dù các học giả bám vào công việc này trong vài chục năm, vẫn còn nhiều mảnh chưa tìm được vị trí đúng.

Việc quyền đọc văn bản ban đầu chỉ được trao cho một số ít học giả cũng là vấn đề. Các nhà nghiên cứu bên ngoài thậm chí khó xem ảnh[1]. Vào tháng 9 năm 1991, Thư viện Huntington ở Hoa Kỳ đã mở khóa cánh cửa này. Thư viện mở các cuộn phim ảnh mà mình lưu giữ mà không đặt điều kiện[3]. Khi ảnh được mở, các nỗ lực ghép mảnh bằng máy tính bắt đầu.

Việc di vật bị phân tán sang nhiều nước cũng là trở ngại. Trong giai đoạn đầu sau phát hiện, một số cuộn được bán qua các nhà buôn đồ cổ ở Bethlehem[1]. Các cơ quan của những nước tài trợ chi phí khai quật

cũng được chia một số mảnh. Vì vậy, các mảnh rơi ra từ cùng một cuộn lại nằm riêng ở nhiều thành phố. Muốn xem hai mảnh có khớp nhau không, con người phải lên máy bay.

Những mảnh giả xuất hiện trên thị trường

Khi giá các mảnh Cuộn Biển Chết tăng lên, một vấn đề khác xuất hiện. Từ khoảng năm 2002, trên thị trường đồ cổ xuất hiện những vật được gọi là mảnh Cuộn Biển Chết. Khoảng 70 mảnh có ghi câu Kinh Thánh đã xuất hiện[4]. Các học giả gọi những vật này là "mảnh hậu 2002". Những vật này không có hồ sơ cho biết chúng đến từ hang nào và xuất hiện khi nào[6].

Bảo tàng Kinh Thánh ở Washington, Hoa Kỳ, cũng mua 16 mảnh như vậy và trưng bày. Nhiều học giả nghi ngờ vì hình dạng chữ trông gượng gạo. Bảo tàng giao việc giám định kỹ cho một công ty điều tra giả mạo tác phẩm nghệ thuật. Người dẫn dắt cuộc điều tra là điều tra viên Colette Loll[4].

Kết quả được công bố vào tháng 3 năm 2020. Cả 16 mảnh đều là đồ giả được làm trong thời hiện đại[4][5]. Nhóm điều tra quan sát bề mặt bằng kính hiển vi 3D. Vật liệu không phải là da thuộc mỏng của Cuộn Biển Chết thật. Ít nhất 15 mảnh là da dày và có thớ thô. Mực đọng trong các khe nứt đã có sẵn[4]. Đó là bằng chứng rằng chữ được viết sau khi da đã nứt.

Vụ việc này làm rung chuyển toàn bộ giới khảo cổ Kinh Thánh. Giáo sư Christopher Rollston của Đại học George Washington nói rằng đồ giả vẫn tiếp tục xuất hiện trên thị trường. Ông thường gặp những người tin rằng vật của mình là thật, nhưng ông nói gần như tất cả đều là giả. Sự khác biệt giữa di vật có hồ sơ khai quật và di vật mua trên thị trường lộ rõ ở đây[6].

Việc ghép các mảnh vỡ và việc lọc đồ giả có cùng gốc. Cả hai đều phải đọc những dấu vết vật lý rất nhỏ mà mắt người không bắt được. AI đã bước vào công việc này. Vào tháng 6 năm 2026, Hội đồng Nghiên cứu châu Âu cấp một khoản tài trợ nghiên cứu mới. Hội đồng cấp 2,5 triệu euro cho một nghiên cứu kéo dài 5 năm nhằm xác định xuất xứ của Cuộn Biển Chết. Tên nghiên cứu là "Theo dấu người chép và cuộn sách". Nghiên cứu này kết hợp phân tích thành phần vật liệu, AI và nghiên cứu chữ viết cổ[7].

Có hai vụ việc sẽ được theo dõi trong chương này. Một là nối lại các mảnh đã phân tán. Hai là lọc đồ giả. Hai công việc đặt những câu hỏi khác nhau đối với cùng một tư liệu. Câu hỏi trước là hai mảnh này vốn có phải là một không. Câu hỏi sau là vật này vốn có phải là cổ vật không.

Tài liệu tham khảo

[1] Israel Antiquities Authority. (n.d.). Discovery and publication. The Leon Levy Dead Sea Scrolls Digital Library. https://www.deadseascrolls.org.il/learn-about-the-scrolls/discovery-and-publication (truy cập ngày 3 tháng 9 năm 2026)

[2] Israel Antiquities Authority. (n.d.). The digital library. The Leon Levy Dead Sea Scrolls Digital Library. https://www.deadseascrolls.org.il/about-the-project/the-digital-library (truy cập ngày 3 tháng 9 năm 2026)

[3] Deseret News. (1991, 9월 22일). Research library will give scholars access to Dead Sea Scroll photos. Deseret News. https://www.deseret.com/1991/9/22/18942605/research-library-will-give-scholars-access-to-dead-sea-scroll-photos/ (truy cập ngày 3 tháng 9 năm 2026)

[4] Greshko, M. (2020, 3월 13일). "Dead Sea Scrolls" at the Museum of the Bible are all forgeries. National Geographic. https://www.nationalgeographic.com/history/article/museum-of-the-bible-dead-sea-scrolls-forgeries (truy cập ngày 3 tháng 9 năm 2026)

[5] Solly, M. (2020, 3월 16일). All of the Museum of the Bible's Dead Sea Scrolls are fake, report finds. Smithsonian Magazine. https://www.smithsonianmag.com/smart-news/all-museum-bibles-dead-sea-scrolls-are-fake-report-finds-180974425/ (truy cập ngày 3 tháng 9 năm 2026)

[6] Govier, G. (2026, 4월 8일). New Dead Sea Scrolls exhibit is the real deal. Christianity Today. https://www.christianitytoday.com/2026/04/new-dead-sea-scrolls-exhibit-is-the-real-deal/ (truy cập ngày 3 tháng 9 năm 2026)

[7] The Jerusalem Post. (2026, 6월 30일). New AI-powered research project aims to uncover the origins of the Dead Sea Scrolls. The Jerusalem Post. https://www.jpost.com/archaeology/article-900938 (truy cập ngày 3 tháng 9 năm 2026)

4.2. Truy vết của đội điều tra khoa học 1: Ái lực mép hình học và dấu vân tay vật lý

Biến mép rách thành con số

Manh mối đầu tiên để ghép các mảnh vỡ là hình dạng mép. Khi xé giấy bằng tay, một đường lởm chởm như răng cưa xuất hiện. Đường của hai mảnh kề nhau ăn khớp với nhau. Con người so sánh đường này bằng mắt. Máy tính biến đường này thành con số.

Máy tính lần theo mép và đo mức độ cong. Khi xếp mức độ cong theo thứ tự, ta có một chuỗi số giống hình sóng. Nếu chuỗi số của hai mảnh khớp tốt khi lật ngược và chồng lên nhau, chúng trở thành ứng viên ghép. Cách này được gọi là tính ái lực mép. Có thể hiểu là điểm ăn khớp mép.

Có thể gọi điểm này là điểm ăn khớp mép. Đây là con số cho thấy mép của hai mảnh khớp nhau đến mức nào.

Phương pháp chỉ nhìn mép có giới hạn rõ ràng. Mép của mảnh cổ đã mòn. Nhiều chỗ co lại khi khô và mất hình dạng ban đầu. Vì vậy, các nhà nghiên cứu xem cả những manh mối ngoài mép.

Dấu vân tay do chính vật liệu để lại

Chỉ dựa vào mép là không đủ. Rìa của mảnh thường đã mòn. Vì vậy, các nhà nghiên cứu nhìn vào chính vật liệu.

Giấy cói được làm bằng cách chẻ mảnh thân cây thành sợi nhỏ rồi xếp chồng ngang dọc. Vì vậy, hoa văn thân cây còn lại trên bề mặt như thớ. Hai mảnh vốn thuộc cùng một tờ sẽ có thớ nối tiếp nhau[1].

Nhóm nghiên cứu của Roy Avitbol tại Đại học Haifa, Israel, đã đọc các thớ này. Nhóm cắt vùng gần mép mảnh thành các ô vuông nhỏ. Mạng nơ-ron nhìn hai ô vuông và chấm điểm xem thớ có nối tiếp nhau không. Nhóm gom các điểm đó để phán định cặp của toàn bộ mảnh[1].

Đối tượng thử nghiệm là một tập hợp mảnh Cuộn Biển Chết trong tình trạng xấu. Vì đây là tập hợp không biết đáp án, nhiều ứng viên sai cũng xuất hiện. Dù vậy, họ đã xóa 98 phần trăm các cặp không thể khớp[1]. Khối lượng con người phải xem giảm xuống còn 1/50.

Ý nghĩa của kết quả này rất rõ. Máy tính không xác định chắc chắn cặp đúng. Máy tính dọn đi các cặp không cần xem. Trong ghép mảnh, phần AI thường đảm nhận là việc lọc như vậy.

Nhóm nghiên cứu của Antoine Pirrone ở Bordeaux, Pháp, dùng mạng nơ-ron song sinh cho việc này. Mạng nơ-ron song sinh đưa hai hình vào cạnh nhau và phán định chúng có cùng nguồn gốc không. Nhóm đặt tên mô hình là Papy-S-Net. Mô hình học từ khoảng 500 mảnh giấy cói. Nó chọn cặp chỉ bằng thớ giấy, không nhìn chữ. Trên tư liệu thật, nó ghép đúng 79 phần trăm các cặp[2].

Da thuộc cũng có dấu vân tay tương tự. Trên da còn lại dấu lỗ chân lông, là nơi từng mọc lông. Các mảnh lấy từ da của cùng một con vật sẽ có cách sắp xếp các dấu này nối tiếp nhau. Lỗ do côn trùng đục xuyên qua cuộn cũng là manh mối. Khi mở một cuộn từng được cuộn lại, các lỗ côn trùng lặp lại thành hàng. Lớp bên trong của cuộn ngắn hơn và lớp bên ngoài dài hơn. Vì vậy, khoảng cách giữa các lỗ rộng dần ra phía ngoài. Nếu đo khoảng cách này, có thể đoán mảnh đó từng nằm ở lớp thứ mấy.

Muốn xử lý những dấu vân tay vật lý này, trước hết phải chỉnh lý ảnh. Trong ảnh Cuộn Biển Chết của Cơ quan Cổ vật Israel có cả thước, bảng màu và bảng số. Nền màu đen nên dễ nhầm với mực. Nhóm nghiên cứu của Bronson Brown-deVost đã tạo một quy trình bốn bước để xóa các vật gây nhiễu này trước. Nhóm cũng công bố một tập tư liệu có đáp án đi kèm[3].

Việc công bố tập tư liệu có đáp án có ý nghĩa lớn. Nhà nghiên cứu khác có thể thử phương pháp của mình trên cùng một tư liệu. Cũng có thể so sánh bằng con số xem phương pháp nào tốt hơn. Nghiên cứu cổ văn tự đang chuyển thành một lĩnh vực trong đó các bên kiểm chứng lẫn nhau như vậy.

Công nghệ lan sang bảng đất sét, xương và giấy

Cùng một ý tưởng đã được chuyển sang các di vật khác. Ở Đại học Ludwig Maximilian München, Đức, có Giáo sư Enrique Jiménez. Ông dẫn dắt dự án Thư viện Babylon điện tử (eBL). Công cụ ghép mảnh của dự án này là Fragmentarium[4].

Người Mesopotamia dùng bút sậy ấn chữ hình nêm lên đất sét ướt. Những bảng đất sét này dễ vỡ. Nhóm nghiên cứu chụp ảnh các bảng đất sét tại Bảo tàng Anh và Bảo tàng Iraq. Từ năm 2018, họ xử lý khoảng 22.000 mảnh làm dữ liệu[5]. Máy tính so sánh các chuỗi ký tự được phiên chép từ chữ còn lại trên mảnh. Khi đó, họ mượn phương pháp dùng trong sinh học để so sánh trình tự gene. Lý do là các thư lại cổ viết chính tả khác nhau tùy từng vùng[4].

Thành quả hiện ra bằng con số. Trong 5 năm, dự án này tìm được khoảng 1.500 kết hợp mới. Trong 150 năm trước đó, các học giả đã tìm được khoảng 5.000 kết hợp[4].

Bảng đất sét có thêm một manh mối mà da thuộc không có. Đó là hình dạng ba chiều của mặt gãy. Bảng đất sét thường được phơi nắng, chỉ một phần được nung lửa. Trong quá trình cứng lại, phần bên trong trở nên cứng không đều. Vì vậy, mỗi mặt vỡ có độ lồi lõm khác nhau. Nếu đo mặt này dưới dạng 3D, có thể tìm cặp ăn khớp.

Ở Trung Quốc, công nghệ tương tự được dùng cho các mảnh giáp cốt. Giáp cốt là ghi chép mà người nhà Thương khắc trên yếm rùa và xương bò khi bói toán. Nhóm nghiên cứu của Zhang Chongsheng tại Đại học Hà Nam đã tạo một tập dữ liệu ghép giáp cốt. Thuật toán ghép của nhóm đạt độ chính xác trong top 10 là 98,39 phần trăm[6]. Nhóm của Zhang Zhan cũng huấn luyện một mô hình nối lại sâu bằng ảnh mảnh giáp cốt. Nhóm này cũng công bố tập dữ liệu đi kèm[7].

Cụm từ độ chính xác trong top 10 cần được diễn giải. Máy không đưa ra một đáp án duy nhất. Nó đưa ra mười ứng viên có khả năng cao theo thứ tự. Xác suất đáp án đúng nằm trong mười ứng viên đó là 98,39 phần trăm[6]. Phán đoán cuối cùng do con người thực hiện.

Việc gom dữ liệu vào một nơi cũng được tiến hành. Đại học Sư phạm An Dương vận hành nền tảng dữ liệu giáp cốt tên là Yinqiwenyuan. Tính đến tháng 4 năm 2023, nền tảng này có hơn 230.000 ảnh giáp cốt. Nền tảng cũng gom 152 tác phẩm và hơn 3.000 bài luận liên quan đến giáp cốt văn[8].

Các mảnh kinh Phật từ hang Mogao Đôn Hoàng cũng bị phân tán ở nhiều nước. Nhóm nghiên cứu của Chen Mingkun đưa ra một phương pháp mới vào năm 2026. Phương pháp này tự học ngay cả khi con người không gắn đáp án. Cách học này được gọi là học tự giám sát. Nhóm nghiên cứu gom cả bản chép nguyên vẹn và các mảnh đã biết cặp. Họ cũng đưa vào dữ liệu các mảnh lẻ chưa biết cặp. Máy tính tự tạo bài tập luyện từ dữ liệu này. Nhóm thiết kế một cách mã hóa mới bao gồm cả hướng của mép. Sau đó, họ dùng mạng nơ-ron song sinh để chấm điểm mức độ giống nhau của hai mảnh[9].

Về phía Cuộn Biển Chết, các nhà nghiên cứu châu Âu và Israel cùng tạo một bàn làm việc trực tuyến. Tên của nó là Scripta Qumranica Electronica. Trên bàn làm việc này, nhà nghiên cứu di chuyển, xoay và sắp xếp ảnh mảnh trên màn hình. Kết quả sắp xếp có thể được chia sẻ với nhà nghiên cứu khác[10]. Việc ghép mảnh đã chuyển từ công việc trong đầu một người sang công việc được nhiều người cùng xem xét.

Tài liệu tham khảo

[1] Abitbol, R., Shimshoni, I., & Ben-Dov, J. (2021). Machine learning based assembly of fragments of ancient papyrus. Journal on Computing and Cultural Heritage, 14(3), 1-21. https://doi.org/10.1145/3460961

[2] Pirrone, A., Beurton-Aimar, M., & Journet, N. (2019). Papy-S-Net: A Siamese network to match papyrus fragments. Proceedings of the 5th International Workshop on Historical Document Imaging and Processing, 78-83. https://doi.org/10.1145/3352631.3352646

[3] Brown-deVost, B., Kurar-Barakat, B., & Dershowitz, N. (2024). Segmenting Dead Sea Scroll fragments for a scientific image set. arXiv. https://arxiv.org/abs/2406.15692

[4] Jiménez, E. (2023, 5월 25일). AI is helping us read ancient Mesopotamian literature. The Conversation. https://theconversation.com/ai-is-helping-us-read-ancient-mesopotamian-literature-205091 (truy cập ngày 3 tháng 9 năm 2026)

[5] Phys.org. (2023, 2월 2일). Researcher uses AI to make texts that are thousands of years old readable. Phys.org. https://phys.org/news/2023-02-ai-texts-thousands-years-readable.html (truy cập ngày 3 tháng 9 năm 2026)

[6] Zhang, C., Zong, R., Cao, S., Men, Y., & Mo, B. (2020). AI-powered oracle bone inscriptions recognition and fragments rejoining. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 5309-5311. https://doi.org/10.24963/ijcai.2020/779

[7] Zhang, Z., Zhang, H., Guo, A., Jiao, Q., Liu, Y., Li, B., & Gao, F. (2025). Deep rejoining model and dataset of oracle bone fragment images. npj Heritage Science, 13(1), 66. https://doi.org/10.1038/s40494-025-01651-9

[8] 王者. (2023, 4월 8일). "殷契文渊"数据平台为甲骨文研究提供助力. 人民日报. https://news.hangzhou.com.cn/gnxw/content/2023-04/08/content_8509027.htm (truy cập ngày 3 tháng 9 năm 2026)

[9] Chen, M., Xiang, Y., & Zheng, Y. (2026). A self-supervised learning approach for pairwise matching of ancient Dunhuang fragments. Journal on Computing and Cultural Heritage, 19(1), 1-19. https://doi.org/10.1145/3776642

[10] Scripta Qumranica Electronica. (n.d.). Scripta Qumranica Electronica. University of Haifa. https://megillot.haifa.ac.il/index.php/en/research/scripta-qumranica-electronica (truy cập ngày 3 tháng 9 năm 2026)

4.3. Truy vết của đội điều tra khoa học 2: AI thám tử phát hiện đồ giả

Cách bắt ra 16 mảnh giả

16 mảnh Cuộn Biển Chết mà Museum of the Bible mua đều là giả[1][2]. Nhóm điều tra không chỉ nhìn chữ bằng mắt. Họ mổ xẻ vật liệu và mực bằng các phương pháp vật lý và hóa học. AI không được dùng trong việc giám định này. Con người đã phát hiện bằng kính hiển vi và phân tích thành phần[1].

Việc giám định của Museum of the Bible không kết thúc trong một lần. Tháng 10 năm 2018, Viện Nghiên cứu và Thử nghiệm Vật liệu Liên bang Đức trước tiên đã kiểm tra 5 mảnh. 5 mảnh này không khớp với đặc điểm của di vật cổ đại, nên bị rút khỏi trưng bày[5]. Sau đó, kết quả kiểm tra cả những mảnh còn lại được công bố vào năm 2020[1][2].

Bằng chứng thứ nhất là vật liệu nền. Cuộn Biển Chết thật được viết trên da thuộc mỏng đã được căng phẳng. Trong các mảnh của Museum of the Bible, ít nhất 15 mảnh là da. Loại da này dày hơn da thuộc và có thớ thô hơn[1].

Bằng chứng thứ hai là vị trí của mực. Nhóm điều tra quan sát bề mặt bằng kính hiển vi 3 chiều. Mực đọng trong các khe nứt. Ở mép rách có dấu vết mực chảy xuống dưới[1].

Có thể hiểu vì sao dấu vết này mang tính quyết định nếu nghĩ theo thứ tự. Bản chép thật trước hết được viết trên da thuộc phẳng. Sau đó, qua nhiều năm dài, da khô đi và nứt ra. Vì vậy, ở bản thật, nét chữ bị đứt tại các khe nứt. Bản giả có thứ tự ngược lại. Kẻ làm giả kiếm được miếng da cũ vốn đã nứt, rồi đặt mực lên đó. Mực chảy vào trong các khe đã há miệng[1].

Bằng chứng thứ ba là chất được bôi lên bề mặt. Nhóm điều tra tìm thấy dấu vết cho thấy các mảnh từng được ngâm trong chất lỏng màu hổ phách. Chất lỏng này có vẻ là keo làm từ da động vật. Kẻ làm giả đã cố bắt chước cảm giác bóng láng trên bề mặt Cuộn Biển Chết thật[1].

Ba bằng chứng chỉ về cùng một đáp án từ những hướng khác nhau. Một bằng chứng riêng lẻ có thể là ngẫu nhiên. Khi ba bằng chứng chồng lên nhau, đó không còn là ngẫu nhiên. Đây là lý do việc giám định di vật phải gom nhiều loại bằng chứng khác nhau.

Dấu vết do bàn tay để lại

Người làm giả nhìn ảnh thật rồi đồ lại chữ. Việc đồ lại này để lại dấu vết.

Khi con người viết kiểu chữ quen thuộc, bàn tay chảy đi không dừng. Khi sao chép hình dạng chữ lạ thì khác. Người viết dừng lại khi đang vẽ nét, rồi kiểm tra vị trí của nét tiếp theo. Tốc độ đầu bút giảm xuống, và đường nét rung nhẹ. Khi phóng to lên, khác biệt này hiện ra trong hình dạng đường viền ngoài của nét chữ.

Máy tính bắt được những khác biệt này bằng con số. Tháng 6 năm 2026, nhóm nghiên cứu của Hassan Ugail tại University of Bradford ở Anh công bố một nghiên cứu liên quan. Nhóm nghiên cứu tạo ra AI để phán định một bản ký họa cũ có phải là thật hay không. Mô hình này chỉ nhìn 5 con số rút ra từ tranh. Nó đo mức độ lặp lại của hoa văn và mức độ phân tán của giá trị độ sáng. Nó cũng đo chênh lệch giữa vùng sáng và vùng tối. Nó cộng thêm mức độ đồng đều của chất liệu bề mặt và mức độ đường nét tách nhánh như cành nhỏ[3].

Nhóm nghiên cứu dạy mô hình bằng các bản ký họa thật của 10 họa sĩ. Niên đại tác phẩm trải từ khoảng năm 1510 đến năm 1917. Dữ liệu được thu thập từ 6 bảo tàng công bố tư liệu mở. Metropolitan Museum of Art và Ashmolean Museum nằm trong số đó[3].

Kết quả thử nghiệm bằng 900 phán định được đăng trong bài báo. Độ chính xác cân bằng là 87,6 phần trăm. Tỷ lệ nhận tác phẩm thật là thật là 77,8 phần trăm. Tỷ lệ nhận nhầm tác phẩm của người khác là tác phẩm của họa sĩ đó là 2,6 phần trăm[3].

Nghiên cứu này còn có một kết quả đáng chú ý. Các mô hình học sâu lớn được dùng rộng rãi lại cho kết quả kém hơn trong việc này. Mô hình lớn thường đẩy tác phẩm thật thành giả. Khi dữ liệu ít, mô hình nhỏ được thiết kế phù hợp với bài toán cho kết quả tốt hơn[3].

Kết quả này cũng có giá trị tham khảo cho nghiên cứu cổ văn tự học. Số bản chép cổ còn lại rất ít. Không có đủ dữ liệu để nuôi các mô hình lớn thật no. Điều đó có nghĩa là thiết kế định sẵn việc cần nhìn vào đâu lại tốt hơn.

Cỗ máy đo kiểu chữ

Công nghệ tương tự cũng được dùng khi nghiên cứu di vật thật. Nhóm nghiên cứu của Mladen Popović tại University of Groningen ở Hà Lan đã phân tích kiểu chữ của Cuộn Biển Chết bằng AI. Nhóm nghiên cứu dùng mạng thần kinh để chỉ tách ra dấu vết nơi mực đã đi qua. Sau đó, họ chuyển độ cong và độ nghiêng của chữ thành con số[4].

Nhóm nghiên cứu huấn luyện mô hình bằng 24 bản chép đã được định tuổi bằng carbon. Tên của mô hình là Enoch. Enoch ước tính tuổi của 135 bản chép chưa rõ niên đại. Sai số trung bình khoảng 28 đến 31 năm. Nhiều bản chép được cho là cổ hơn mức đã biết cho đến nay[4].

Công nghệ này cũng giúp ích cho việc phán định thật giả. Chữ của người thư lại thật có những đặc điểm thống kê ổn định theo từng thời đại. Bản giả lệch khỏi phân bố này. Mắt người cảm thấy hai kiểu chữ giống nhau, nhưng máy đưa khác biệt ra thành con số.

Tuy nhiên, sự trợ giúp này chỉ ở mức có thêm một tư liệu tham khảo. Nghiên cứu Enoch có điểm cần thận trọng. Mô hình này được huấn luyện để đoán niên đại. Nó không được huấn luyện để phân biệt thật và giả. Không thể nói ngay là giả chỉ vì kết quả ước tính niên đại trông lạ. Kết quả của máy chỉ có ý nghĩa trong phạm vi việc mà máy đã học.

Việc máy không thể thay thế

Phán định của máy cũng có giới hạn. Thứ đưa ra kết luận cuối cùng về tính thật giả của các mảnh tại Museum of the Bible không phải là một AI duy nhất. Quan sát bằng kính hiển vi, thử nghiệm hóa học và tri thức văn hiến học đã được dùng cùng nhau[1][2].

Sức nặng của hồ sơ xuất xứ vẫn còn nguyên. Hồ sơ xuất xứ là ghi chép cho biết di vật đến từ đâu. Giáo sư Rollston nói rằng hãy phân biệt di vật có hồ sơ khai quật với di vật mua trên thị trường. Với cuộn giấy được khai quật, người ta còn biết nó ra từ hang nào và vào lúc nào. Các mảnh xuất hiện trên thị trường không có hồ sơ đó[5].

Máy đọc giúp những tín hiệu rất nhỏ mà con người bỏ lỡ. Con người phán đoán tín hiệu đó có nghĩa gì. Việc nhận ra đồ giả chỉ kết thúc khi có cả hai việc này.

Tài liệu tham khảo

[1] Greshko, M. (2020, 3월 13일). "Dead Sea Scrolls" at the Museum of the Bible are all forgeries. National Geographic. https://www.nationalgeographic.com/history/article/museum-of-the-bible-dead-sea-scrolls-forgeries (truy cập ngày 3 tháng 9 năm 2026)

[2] Solly, M. (2020, 3월 16일). All of the Museum of the Bible's Dead Sea Scrolls are fake, report finds. Smithsonian Magazine. https://www.smithsonianmag.com/smart-news/all-museum-bibles-dead-sea-scrolls-are-fake-report-finds-180974425/ (truy cập ngày 3 tháng 9 năm 2026)

[3] Ugail, H., Ritch-Frel, J., Matuzava, I., & Stork, D. G. (2026). Verification of historical sketches via one-class learning on compact feature representations. PLOS One, 21(6), e0344796. https://doi.org/10.1371/journal.pone.0344796

[4] Popović, M., Dhali, M. A., Schomaker, L., van der Plicht, J., Lund Rasmussen, K., La Nasa, J., Degano, I., Perla Colombini, M., & Tigchelaar, E. (2025). Dating ancient manuscripts using radiocarbon and AI-based writing style analysis. PLOS One, 20(6), e0323185. https://doi.org/10.1371/journal.pone.0323185

[5] Museum of the Bible. (2018, 10월 22일). Museum of the Bible releases research findings on fragments in its Dead Sea Scrolls collection. Museum of the Bible. https://www.museumofthebible.org/newsroom/museum-of-the-bible-releases-research-findings-on-fragments-in-its-dead-sea-scrolls-collection (truy cập ngày 3 tháng 9 năm 2026)

4.4. Sự thật được làm sáng tỏ

Ranh giới giữa mối ghép đã xác định và mối ghép ứng viên

Ngay cả khi máy tính nói hai mảnh ghép với nhau, việc đó vẫn chưa kết thúc. Máy chỉ đưa ra xác suất. Các nhà nghiên cứu phân biệt nghiêm ngặt giữa mối ghép đã xác định và mối ghép ứng viên.

Mối ghép đã xác định chỉ được công nhận khi nhiều loại bằng chứng cùng khớp. Trước hết, hình dạng mặt gãy phải ăn khít không có kẽ hở. Tiếp theo, thớ vật liệu phải nối tiếp nhau. Với giấy cói, hoa văn thân cây phải nối tiếp[1]. Với da thuộc, dấu lỗ chân lông phải nối tiếp. Sau đó, nét chữ đi qua ranh giới phải nối tiếp tự nhiên. Cuối cùng, phần chữ được ghép lại phải khớp với ngữ pháp của ngôn ngữ đó.

Mối ghép ứng viên là trường hợp chỉ khớp một phần trong các điều kiện này. Kết quả suy đoán vị trí mảnh bằng khoảng cách giữa các lỗ sâu thuộc loại này. Kết quả gom thành một nhóm cuộn giấy vì thành phần mực giống nhau cũng thuộc ứng viên. Những kết quả này trở thành điểm khởi đầu của nghiên cứu, nhưng không phải kết luận. Trình tự là con người kiểm tra các ứng viên mà máy tính đã lọc ra[1].

Nếu không giữ sự phân biệt này thì nguy hiểm. Nếu ghép hai mảnh chỉ vì xác suất cao, một câu vốn không tồn tại sẽ xuất hiện. Nếu câu đó được trích dẫn, sai lầm sẽ lan sang nghiên cứu tiếp theo. Vì vậy, trong hồ sơ dữ liệu, người ta ghi riêng phần đã xác định và phần ứng viên.

Những con số do máy đưa ra cũng dựa trên sự phân biệt này. Tỷ lệ lọc cao không có nghĩa là mối ghép đã được xác định[1]. Xác suất đáp án đúng nằm trong nhóm ứng viên hàng đầu cũng vậy[2]. Dù theo hướng nào, phán định cuối cùng vẫn do con người làm.

Văn bản sống lại

Bằng cách này, văn hiến thật đã sống lại. Fragmentarium đã xử lý khoảng 22.000 mảnh bảng đất sét từ năm 2018. Chương trình này đã tìm ra nhiều mối ghép mới và bản chép mới. Tháng 2 năm 2023, nhóm nghiên cứu mở Fragmentarium cho công chúng. Họ cũng công bố bản số hóa của Sử thi Gilgamesh[3].

Trong số các văn bản sống lại có Sử thi Gilgamesh. Sử thi Gilgamesh là câu chuyện anh hùng của Mesopotamia. Tháng 11 năm 2022, chương trình nhận ra một mảnh. Mảnh này đến từ một bản được sao chép vào khoảng năm 130 trước Công nguyên. Trong các bản chép Gilgamesh đã được xác nhận đến nay, nó thuộc giai đoạn muộn. Điều đó có nghĩa là người ta vẫn chép lại câu chuyện này cho đến thời kỳ đó[3].

Chữ hình nêm có cách viết khác nhau theo từng vùng. Vì các thư lại đổi chữ theo phương ngữ của mình. Khi con người tìm bằng mắt, khác biệt này là một bức tường lớn. Nhóm nghiên cứu vượt qua bức tường này bằng cách mượn phương pháp so sánh chuỗi gene. Nhờ vậy, các chuỗi ký tự hơi khác nhau cũng có thể được so sánh[4].

Ở phía Cuộn Biển Chết, nền dữ liệu cũng mở rộng. Israel Antiquities Authority đã chụp ảnh độ phân giải cao và công bố hơn 25.000 mảnh di vật[5]. Các nhà nghiên cứu di chuyển và sắp xếp các mảnh trên những bức ảnh này[6]. Một phương pháp xử lý tự động xóa vật gây nhiễu như thước và bảng màu cũng đã được công bố[7].

Cánh cửa tiếp theo mở ra vào năm 2026

Nghiên cứu “Truy tìm thư lại và cuộn giấy” bắt đầu vào tháng 6 năm 2026 mở ra cánh cửa tiếp theo. Câu hỏi mà nghiên cứu này đặt ra đi trước việc ghép mảnh một bước. Nó hỏi Cuộn Biển Chết được viết ở đâu và bởi ai[8].

Nhóm nghiên cứu kiểm tra khoảng 250 mẫu da thuộc, giấy cói và mực. Mẫu là vật nhỏ dùng để kiểm tra. Họ chồng thêm phân tích kiểu chữ bằng AI và nghiên cứu chữ viết cổ lên đó. Israel Antiquities Authority, University of Pisa và University of Naples cùng tham gia. University of Southern Denmark và KU Leuven của Bỉ cũng tham gia. Họ còn so sánh với giấy cói Ai Cập do các bảo tàng ở Berlin và Turin lưu giữ[8].

Nếu thành phần của mẫu giống nhau, khả năng chúng đến từ cùng một xưởng tăng lên. Thông tin này cũng được dùng để ghép mảnh. Nếu trước hết gom những mảnh có cùng nguồn gốc vật liệu, số cặp cần so sánh sẽ giảm.

Những mảnh còn lại và việc còn lại

Dù thành quả đã tích lũy, vẫn còn nhiều mảnh chưa tìm được vị trí của mình. Trong các hộp lưu trữ của cơ quan di vật vẫn còn những mảnh chưa biết đến từ cuộn nào[8]. Ở phía bảng đất sét, vẫn còn nhiều mối ghép chưa tìm thấy[4].

Khác biệt về thiết bị cũng vẫn là vấn đề. Để đo di vật bằng máy quét lập thể, phải đưa di vật đến trước thiết bị. Có nhiều bảo tàng khó đưa di vật ra ngoài. Vì vậy, một số di vật chỉ có dữ liệu ảnh, không có dữ liệu lập thể. Với những di vật này, không thể dùng phương pháp so sánh mặt gãy.

Công nghệ cũng để lại bài toán phải giải. Ảnh mảnh được chụp bằng thiết bị khác nhau ở từng quốc gia. Khi ánh sáng và nền khác nhau, cùng một thuật toán cũng đưa ra đáp án khác nhau. Vì vậy, các nhà nghiên cứu đã tạo một bộ dữ liệu công khai có gắn đáp án đúng. Mục đích là so sánh kết quả của nhau bằng cùng một thước đo[7]. Nghiên cứu các mảnh Đôn Hoàng cũng chuyển hướng sang cách học mà không cần con người gắn đáp án đúng[9].

Công nghệ này không đẩy con người ra ngoài. Máy giảm số cặp cần so sánh, còn con người phán đoán ý nghĩa. Muốn biết câu được khôi phục có khớp với lối nói của thời đại đó hay không, con người phải đọc. Vì hai vai trò không chồng lên nhau nên chúng có thể làm việc cùng nhau.

Cách công bố cũng đã thay đổi. Trước đây, chỉ một số ít học giả được xem ảnh gốc. Giờ đây, dữ liệu bảng đất sét và ảnh Cuộn Biển Chết đã mở trên internet[4][5]. Sinh viên cũng có thể di chuyển các mảnh trên màn hình[6].

Keo dán kỹ thuật số không dán các mảnh lại về mặt vật lý. Di vật vẫn nằm nguyên trong hộp. Cuộn giấy chỉ được nối lại trên màn hình máy tính. Vì vậy, nếu phán định sai, người ta có thể quay lại bất cứ lúc nào. Văn bản từng tản mát suốt 2.000 năm đang từng chút trở lại thành câu như vậy.

Tài liệu tham khảo

[1] Abitbol, R., Shimshoni, I., & Ben-Dov, J. (2021). Machine learning based assembly of fragments of ancient papyrus. Journal on Computing and Cultural Heritage, 14(3), 1-21. https://doi.org/10.1145/3460961

[2] Zhang, C., Zong, R., Cao, S., Men, Y., & Mo, B. (2020). AI-powered oracle bone inscriptions recognition and fragments rejoining. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 5309-5311. https://doi.org/10.24963/ijcai.2020/779

[3] Phys.org. (2023, 2월 2일). Researcher uses AI to make texts that are thousands of years old readable. Phys.org. https://phys.org/news/2023-02-ai-texts-thousands-years-readable.html (truy cập ngày 3 tháng 9 năm 2026)

[4] Jiménez, E. (2023, 5월 25일). AI is helping us read ancient Mesopotamian literature. The Conversation. https://theconversation.com/ai-is-helping-us-read-ancient-mesopotamian-literature-205091 (truy cập ngày 3 tháng 9 năm 2026)

[5] Israel Antiquities Authority. (n.d.). The digital library. The Leon Levy Dead Sea Scrolls Digital Library. https://www.deadseascrolls.org.il/about-the-project/the-digital-library (truy cập ngày 3 tháng 9 năm 2026)

[6] Scripta Qumranica Electronica. (n.d.). Scripta Qumranica Electronica. University of Haifa. https://megillot.haifa.ac.il/index.php/en/research/scripta-qumranica-electronica (truy cập ngày 3 tháng 9 năm 2026)

[7] Brown-deVost, B., Kurar-Barakat, B., & Dershowitz, N. (2024). Segmenting Dead Sea Scroll fragments for a scientific image set. arXiv. https://arxiv.org/abs/2406.15692

[8] The Jerusalem Post. (2026, 6월 30일). New AI-powered research project aims to uncover the origins of the Dead Sea Scrolls. The Jerusalem Post. https://www.jpost.com/archaeology/article-900938 (truy cập ngày 3 tháng 9 năm 2026)

[9] Chen, M., Xiang, Y., & Zheng, Y. (2026). A self-supervised learning approach for pairwise matching of ancient Dunhuang fragments. Journal on Computing and Cultural Heritage, 19(1), 1-19. https://doi.org/10.1145/3776642

Chương 5. Ghép mặt gãy của bảng đất sét: tái dựng sử thi Gilgamesh bằng khoa học máy tính

5.1. Hồ sơ vụ việc

Câu chuyện vượt qua 4.000 năm

Sử thi Gilgamesh thuộc nhóm tác phẩm văn học kể chuyện cổ xưa còn lại đến nay. Những bài ca Sumer sớm hơn cũng còn được truyền lại. Nhân vật chính là Gilgamesh, vị vua cai trị thành Uruk ở Lưỡng Hà[1]. Ông cùng người bạn Enkidu đi đến một rừng tuyết tùng xa xôi. Hai người chiến đấu với Humbaba, quái vật canh giữ khu rừng ấy[2]. Sau đó, khi Enkidu chết, Gilgamesh lên đường tìm con đường không phải chết[1].

Nền để ghi câu chuyện này không phải giấy, mà là bảng đất sét. Các thư lại thời cổ vát xiên cây sậy để làm một que nhọn. Que này được gọi là bút trâm. Khi dùng bút trâm ấn vào đất sét ướt, dấu vết hình nêm sẽ còn lại. Đây là chữ được tạo bằng cách ấn xuống, chứ không phải kẻ đường. Khi đất sét khô, dấu vết ấy cứng lại nguyên dạng. Vì vậy loại chữ này được gọi là chữ hình nêm. Những học trò muốn trở thành thư lại đã chép sử thi này để luyện chữ. Nhờ vậy, các mảnh bản sao được tìm thấy tại nhiều di tích ở Trung Đông. Tại di tích Amarna của Ai Cập cũng có mảnh của sử thi này. Điều đó có nghĩa là câu chuyện đã lan tới nơi rất xa Lưỡng Hà[1].

Các bản truyền lại được chia lớn thành hai loại. Một là bản Cổ Babylon, được ghi vào đầu thiên niên kỷ 2 TCN. Loại còn lại là bản Babylon tiêu chuẩn, được chỉnh lý về sau. Bản tiêu chuẩn được cấu thành từ mười hai bảng đất sét. Trong đó, mười một bảng là phần chính. Bảng thứ mười hai được xem là phụ lục thêm vào sau. Các bản dịch chúng ta đọc hôm nay lấy bản tiêu chuẩn này làm nền. Những chỗ trống được lấp bằng bản Cổ Babylon và các mảnh khác. Người chỉnh lý bản tiêu chuẩn được biết là một thư lại tên Sin-leqi-unninni. Ông đã gom những bài ca cổ rời rạc thành một tác phẩm dài. Ông quyết định nội dung đưa vào từng bảng và chỉnh số dòng. Nhưng ngay trong bản tiêu chuẩn này cũng còn rất nhiều chỗ trống. Không có bản sao nào còn nguyên vẹn từ đầu đến cuối[1].

Nội dung câu chuyện đến nay đọc vẫn không xa lạ. Gilgamesh là một vị vua có sức mạnh lớn nhưng hành hạ dân chúng. Các thần tạo ra Enkidu và gửi xuống làm người đối đầu với ông. Hai người đánh nhau rồi trở thành bạn. Họ cùng đánh bại Humbaba và hạ cả Bò Trời. Cái giá phải trả là Enkidu lâm bệnh rồi chết. Chứng kiến cái chết của bạn, Gilgamesh nhận ra chính mình cũng sẽ chết. Bảng 11 có cảnh ông tìm đến người sống sót sau trận đại hồng thủy và nghe câu chuyện của người ấy[1].

Câu chuyện này đã bị quên lãng hơn 2.000 năm. Lý do là không còn ai đọc được chữ hình nêm. Đến thế kỷ 19, khi các học giả giải mã lại chữ hình nêm, sử thi này trở về với thế giới. Khởi đầu là khi một nhà nghiên cứu lục trong các thùng bảng đất sét chở từ Nineveh về và nhận ra câu chuyện hồng thủy. Sau đó, trong 150 năm, các học giả đã gom các bản sao rời rạc và từng chút lấp đầy văn bản[1].

Nineveh và ngọn lửa năm 612 TCN

Phần lớn bản sao của bản tiêu chuẩn đến từ một thư viện cổ. Đó là kho sách do vua Ashurbanipal của Assyria thế kỷ 7 TCN lập trong cung điện Nineveh. Thư viện này thu thập đủ loại sách bói toán, sách y học, từ điển và tác phẩm văn học. Nhà vua sai thu gom bảng đất sét từ khắp đế quốc. Ông gửi thư cho các quan, lệnh họ tìm mang về những cuốn sách còn thiếu. Ashurbanipal là một vị vua có thể tự đọc chữ hình nêm[3].

Nineveh bị một trận hỏa hoạn lớn bao trùm vào khoảng năm 612 TCN. Cung điện sụp đổ và thư viện cũng cháy theo. Ngọn lửa này cùng lúc làm hai việc với di vật. Những bảng đất sét vốn chỉ được phơi khô

trong bóng râm đã được nung cứng trong lửa[4]. Bảng đất sét sau khi nung có được thân thể bền lâu như gạch ra khỏi lò. Nếu giấy cói hay da nằm trong cùng ngọn lửa ấy, chúng hẳn đã hóa tro và biến mất.

Nhưng sức nặng của tòa nhà sụp đổ đã làm những tấm bảng cứng ấy vỡ tan. Hiện nay, các bảng đất sét và mảnh vỡ khai quật từ Nineveh mà Bảo tàng Anh lưu giữ vượt quá 30.000 hiện vật[3][4]. Đó là những di vật mà các đoàn khai quật giữa thế kỷ 19 thu từ đống đất rồi chở lên tàu mang về[3]. Một câu chuyện đã đến với chúng ta trong trạng thái vỡ thành hàng chục mảnh.

Kích thước các mảnh không giống nhau. Có mảnh lớn bằng lòng bàn tay, có mảnh chỉ bằng móng ngón tay cái. Trong nhiều trường hợp, chữ còn lại trên một mảnh chỉ vài dòng. Không ai nói cho ta biết vài dòng ấy nằm ở đâu trong mười hai bảng của sử thi. Việc đọc chữ và việc tìm vị trí là hai công việc khác nhau.

Khi một mảnh tìm được đúng chỗ, ý nghĩa cũng có thể thay đổi. Phải nối được trước sau mới biết ai nói và đó là cảnh nào. Khi còn rời rạc, chữ chỉ là chuỗi từ khó hiểu. Khi vào đúng chỗ, chúng thành câu. Một lần ghép nối có thể làm sống lại trọn một cảnh của câu chuyện.

Không thể ghép 350.000 dòng bằng tay

Trong thời gian dài, việc ghép các mảnh vỡ được giao cho bàn tay và trí nhớ của con người. Học giả nhặt một mảnh trong kho lưu trữ. Rồi họ nhớ lại ảnh và bản vẽ tay của các mảnh khác để tìm cặp khớp. Khi cho rằng hai mảnh đến từ cùng một bảng, họ gọi đó là một lần ghép nối.

Cách này bị chặn bởi dung lượng trí nhớ của con người. Trong 150 năm sau khi Assyriology bắt đầu, số lần ghép nối theo cách này khoảng 5.000[5]. Số mảnh còn lại lớn hơn thế rất nhiều. Riêng Bảo tàng Anh đã có hơn 30.000 bảng đất sét từ Nineveh[3]. Nếu tính cả các bộ sưu tập khác, con số còn tăng nhiều hơn. Trong đó, nhiều hiện vật chưa từng được đọc từ thời cổ đại đến nay[5]. Nếu đếm tất cả bảng đất sét chữ hình nêm còn lại trong các bảo tàng trên thế giới, con số được ước tính khoảng 500.000[6].

Còn một lý do nữa khiến công việc khó khăn. Các mảnh nằm rải rác ở nhiều nước. Bảo tàng Anh, Berlin, Đại học Yale và Bảo tàng Penn ở Philadelphia đều có bộ sưu tập riêng[7]. Muốn trực tiếp thử ghép hai mảnh, người ta phải đi qua hai châu lục.

Bản thân chữ hình nêm cũng đòi hỏi nhiều công sức. Cùng một từ nhưng mỗi thư lại ghi hơi khác nhau. Chính tả dao động tùy vùng và thời đại[5]. Ngay cả khi con người dùng mắt so sánh chữ trên hai mảnh, cũng khó nhận ra chúng có thuộc cùng một tác phẩm hay không.

Hơn nữa, một ký hiệu hình nêm có thể được đọc bằng nhiều âm. Cùng một dấu vết có giá trị khác nhau tùy ngữ cảnh. Tìm cặp khớp trong hàng trăm nghìn dòng chữ như vậy là việc mắt người không thể làm xong.

Vì vậy, Giáo sư Enrique Jiménez tại Đại học Ludwig Maximilian ở Munich, Đức, đã chọn một con đường khác. Năm 2018, ông bắt đầu một dự án nhằm phục hồi văn bản Babylon trọn vẹn nhất có thể. Đó là việc gom các bản đọc bảng đất sét rải rác trên thế giới vào một nơi[5]. Dự án này có tên là Thư viện Điện tử Văn học Babylon, viết tắt là eBL. Phòng làm việc ghép mảnh do eBL tạo ra là Fragmentarium. Đến năm 2024, phòng làm việc này đã tích lũy bản đọc của khoảng 25.000 bảng đất sét. Nếu tính theo số dòng, con số vượt quá 350.000 dòng[7].

Lý do việc này cấp bách nằm ở thời gian. Dù đã được nung trong lửa, bảng đất sét không tồn tại mãi mãi. Muối và độ ẩm làm bề mặt phồng lên và xóa chữ. Ngay lúc này, vẫn còn hàng trăm nghìn bảng đất sét đang chờ được đọc trong kho bảo tàng[5]. Không ai biết trong đó có bao nhiêu câu chuyện.

Số người đọc được chữ hình nêm cũng là một vấn đề. Trên thế giới không có nhiều nhà nghiên cứu Assyriology. Trong số các bảng đất sét còn lại, chỉ một phần đã được dịch[6]. Số bảng đất sét mà một người có thể đọc trong đời là có giới hạn. Đây là lý do người ta tìm đến sự trợ giúp của máy móc.

Tài liệu tham khảo

[1] George, A. R. (2003). The Babylonian Gilgamesh Epic: Introduction, Critical Edition and Cuneiform Texts. Oxford University Press.

[2] Al-Rawi, F. N. H., & George, A. R. (2014). Back to the Cedar Forest: The beginning and end of Tablet V of the Standard Babylonian Epic of Gilgameš. Journal of Cuneiform Studies, 66, 69-90. https://doi.org/10.5615/jcunestud.66.2014.0069

[3] The Ashurbanipal Library Project. (n.d.). What is the Library? Oracc, University of Pennsylvania. https://oracc.museum.upenn.edu/asbp/whatisthelibrary/index.html (truy cập ngày 3 tháng 9 năm 2026)

[4] British Museum. (n.d.). What was Ashurbanipal's Library? britishmuseum.org. https://www.britishmuseum.org/research/projects/what-was-ashurbanipals-library (truy cập ngày 3 tháng 9 năm 2026)

[5] Jiménez, E. (2023, 5월 25일). AI is helping us read ancient Mesopotamian literature. The Conversation. https://theconversation.com/ai-is-helping-us-read-ancient-mesopotamian-literature-205091 (truy cập ngày 3 tháng 9 năm 2026)

[6] Wang, Z. (2026). TabletCraft: Bridging a 4,000-year cultural gap with bidirectional Akkadian NMT and cuneiform rendering. arXiv:2608.02609. https://arxiv.org/abs/2608.02609 (truy cập ngày 3 tháng 9 năm 2026)

[7] Cobanoglu, Y., Laasonen, J., Simonjetz, F., Khait, I., Cohen, S., Földi, Z., Hätinen, A., Heinrich, A., Mitto, T., Rozzi, G., Sáenz, L., & Jiménez, E. (2024). Transliterated cuneiform tablets of the electronic Babylonian Library platform. Journal of Open Humanities Data, 10, 19. https://doi.org/10.5334/johd.148

5.2. Cuộc truy dấu của đội điều tra khoa học: dung hợp số Clay Edge và Fragmentarium

Bước đầu tiên là chuyển chữ sang bảng chữ cái

Điểm xuất phát của Fragmentarium là chữ viết. Nhà nghiên cứu nhìn ảnh bảng đất sét và chuyển chữ hình nêm sang chữ La-tinh[1]. Công việc này gọi là phiên tự. Nếu so với tiếng Hàn, nó giống việc ghi chữ Hán bằng âm Hangul. Đó là việc đổi từng ký hiệu hình nêm thành chữ cái theo giá trị âm. Những chỗ vỡ, không đọc được, được đánh dấu bằng ký hiệu khoảng trống.

Văn bản phiên tự được tạo như vậy trở thành chuỗi ký tự mà máy tính có thể xử lý. eBL đã tích lũy các văn bản phiên tự này từ tháng 5 năm 2018 đến tháng 8 năm 2023. Kết quả là khoảng 25.000 bảng đất sét và hơn 350.000 dòng chữ trở thành dữ liệu có thể tìm kiếm. Danh mục hồ sơ bảng đất sét do eBL quản lý vượt quá 260.000 mục. Các bộ sưu tập của Bảo tàng Anh, Bộ sưu tập Babylon Yale và Bảo tàng Penn đều nằm trong đó[1].

Phiên tự do con người thực hiện. Lý do là máy chưa thể trích xuất chữ từ ảnh một cách hoàn hảo. Nhóm nghiên cứu eBL đã tiếp tục công việc phiên tự này hơn năm năm. Kết quả được công bố dưới dạng dữ liệu mở mà bất cứ ai cũng có thể tải xuống[1]. Muốn giao việc cho máy, trước hết con người phải tạo dữ liệu cho máy.

Ghép chuỗi ký tự giống xét nghiệm gen

Thuật toán cốt lõi đến từ một lĩnh vực bất ngờ. Đó là phương pháp căn chỉnh chuỗi ký tự mà các nhà sinh học dùng khi so sánh trình tự gen[2]. Gen là một chuỗi dài nối bốn chữ A, T, G, C. Khi so sánh gen của hai sinh vật, người ta đặt chúng vào cùng vị trí để xem, dù một vài chữ có khác nhau.

Trong chữ hình nêm cũng nảy sinh vấn đề tương tự. Lý do là mỗi thư lại ghi cùng một từ theo cách khác nhau. Giáo sư Jiménez xem dao động này giống đột biến trong gen. Vì vậy, ông sửa thuật toán căn chỉnh của sinh học cho phù hợp với chữ hình nêm[2].

Thứ máy tính thực sự đếm là những mẩu chữ ngắn. Các mẩu được cắt từ văn bản phiên tự theo từng vài ký tự được gọi là n-gram. Hệ thống đếm số n-gram mà hai bảng đất sét cùng có. Cặp nào cùng có mẩu dài hơn sẽ nhận điểm cao hơn. Những cặp có điểm cao được đưa lên trước con người dưới dạng ứng viên.

Fragmentarium mở trên giao diện web. Nhà nghiên cứu ở bất kỳ nước nào trên thế giới cũng có thể truy cập và xem cùng một dữ liệu. Họ có thể đặt mảnh ở London và mảnh ở Munich cạnh nhau trên cùng một màn hình để so sánh[8]. Không còn phải đóng hành lý và lên máy bay.

Cách này chịu được cả khi chính tả dao động. Dù một từ được ghi khác đi, nếu các mẩu xung quanh trùng nhau thì điểm vẫn còn. eBL đã đưa chương trình ghép cặp này lên kho lưu trữ công khai. Các nhà nghiên cứu khác cũng có thể tải cùng công cụ về dùng[6].

Máy tính còn giúp cả việc đọc chữ. Một ký hiệu hình nêm thường được đọc bằng nhiều âm. Mô hình do eBL huấn luyện đạt độ chính xác 98 phần trăm trong bài toán chọn nghĩa đúng giữa nhiều nghĩa ấy[2].

Có thể hiểu dễ vì sao việc này cần thiết qua một ví dụ. Cùng một dấu hình nêm, trong câu này được đọc như một âm. Trong câu khác, nó được đọc như ký hiệu chỉ cả một từ. Nếu đọc sai, điểm ghép cặp sẽ lệch. Chỉnh đúng cách đọc chính là nâng độ chính xác của việc ghép nối.

Thành quả xuất hiện bằng con số. Trong 150 năm, Assyriology đã ghép bằng tay khoảng 5.000 trường hợp. eBL tìm thêm khoảng 1.500 trường hợp ghép nối chỉ trong năm năm[2]. Chỉ trong quá trình tạo tập văn bản phiên tự, khoảng 1.250 trường hợp ghép nối đã lộ ra[1]. Hàng nghìn mảnh tuy không trực tiếp chạm nhau nhưng thuộc cùng một tác phẩm cũng được xác nhận[2].

Mô hình ngôn ngữ lấp chỗ vỡ

Ngay cả khi ghép các mảnh lại, văn bản vẫn còn trống chỗ này chỗ kia. Những chỗ trống này được gọi là lacuna. Từ đây, mô hình ngôn ngữ bước vào.

Năm 2023, nhóm nghiên cứu tại Đại học Tel Aviv và Đại học Ariel công bố một mô hình dịch tiếng Akkad. Mô hình này dịch sang tiếng Anh dù đầu vào là ký hiệu chữ hình nêm hay văn bản phiên tự bằng bảng chữ cái. Nó xử lý cả hai loại đầu vào. Mô hình không xử lý việc dịch ngược từ tiếng Anh sang tiếng Akkad. Dữ liệu dùng để huấn luyện là các bản đọc bảng đất sét đã được xuất bản[3].

Năm 2026, một công cụ xử lý cả chiều ngược lại xuất hiện. Đó là phần mềm mở có tên TabletCrafter. Công cụ này dịch hai chiều, từ tiếng Akkad sang tiếng Anh và từ tiếng Anh sang tiếng Akkad. Dữ liệu huấn luyện gồm 116.000 cặp câu. Bảng chuyển đổi ký hiệu hình nêm chứa hơn 14.000 cặp tương ứng. Nhà phát triển cho biết cần có công cụ để đọc 500.000 bảng đất sét còn lại[4].

Những mô hình như vậy nhìn vào ngữ cảnh trước và sau chỗ trống rồi đưa ra những từ có khả năng điền vào theo xác suất. Thơ Babylon có quy tắc đặt cạnh nhau hai dòng có cùng cấu trúc. Mô hình cũng lấy hình thức này làm manh mối. Thứ mô hình đưa ra không phải đáp án đã xác định, mà là danh sách ứng viên.

Công cụ năm 2026 đọc được cả nét chữ

Ngày 18 tháng 5 năm 2026, Đại học Würzburg của Đức công bố một công cụ mới. Công cụ này do Viện Hàn lâm Mainz và Đại học Kỹ thuật Dortmund cùng phát triển. Tên của nó là Paleograficum. Công cụ này học hình dạng của từng ký hiệu hình nêm trong ảnh bảng đất sét. Sau đó, nó tìm trong toàn bộ bộ sưu tập những ký hiệu được ấn viết theo cùng một cách[5].

Công cụ này xử lý 70.000 bức ảnh và hơn 5 triệu ký hiệu. Trước đây, việc so sánh nét chữ của 5 mảnh vỡ mất 3 ngày. Dùng công cụ này thì chỉ 5 phút là xong. Giống như mỗi người có nét chữ khác nhau, mỗi thư lại cũng có góc ấn nêm và khoảng cách khác nhau. Có thể lấy khác biệt đó làm manh mối để chọn ra các mảnh vỡ đến từ cùng một tấm bảng[5].

Chữ hình nêm Hittite mà công cụ này xử lý có 375 ký hiệu. Số mảnh bảng đất sét Hittite đã biết đến nay lên tới 30.000 mảnh. Giáo sư Daniel Schwemer, người dẫn dắt nghiên cứu, nói rằng công cụ này có thể tiết kiệm hàng nghìn giờ[5].

Mục tiêu tiếp theo của nhóm nghiên cứu là nhận diện từng thư lại. Nghĩa là tìm ra tất cả các bảng đất sét do cùng một người viết. Nếu làm được việc đó, ta có thể vẽ lại một thư lại đã làm việc ở đâu và đã chuyển đến đâu. Đời sống nghề nghiệp của con người 3.500 năm trước sẽ hiện ra bằng tư liệu[5].

Một con đường khác để ghép mép đất sét bị vỡ

Nghiên cứu ghép mảnh vỡ bằng hình dạng thay vì bằng chữ cũng đã kéo dài từ lâu. Bảng đất sét bị vỡ để lại mặt gãy. Mặt này được gọi là mặt đứt gãy. Nếu mặt đứt gãy của hai mảnh khớp vào nhau, chúng đến từ cùng một tấm bảng.

Dự án tiêu biểu đi theo con đường này là Dự án phục dựng ảo bảng đất sét chữ hình nêm. Sandra Woolley của Đại học Keele ở Anh dẫn dắt dự án này. Tim Collins và Erlend Gehlken cùng tham gia. Nhóm nghiên cứu quét các mảnh vỡ bằng máy quét 3D. Sau đó, họ để máy tính phán đoán xem mặt gãy của hai mảnh có khớp với nhau hay không. Các mảnh vỡ cổ thường có mép bị mòn. Sự mài mòn này cũng được đưa vào phép tính[9].

Một thành quả của phương pháp này xuất hiện trong câu chuyện đại hồng thủy. Đó là mảnh vỡ của bảng thứ ba trong sử thi lũ lụt Atrahasis. Một mảnh ở London, mảnh kia ở Geneva. Hai nơi cách nhau khoảng 1.000 kilômét. Đây là cặp mảnh không thể áp sát hiện vật vào nhau nên không xác nhận được mối ghép. Khi chồng dữ liệu 3D lên nhau, hai mảnh đã nối liền[9]. Phương pháp này được công bố thành bài báo vào năm 2014[10].

Nghiên cứu xử lý dữ liệu 3D vẫn tiếp tục đến nay. Vào tháng 3 năm 2026, một bài báo được công bố. Bài báo nói về mạng nơ-ron phân loại thông tin hiện vật bằng dữ liệu đám mây điểm 3D của bảng đất sét. Đám mây điểm là dữ liệu ghi bề mặt vật thể bằng tọa độ của vô số điểm. Khi quét bảng đất sét bằng máy quét 3D, ta thu được loại dữ liệu này. Độ lồi lõm và độ dày của bề mặt được lưu lại dưới dạng con số. Nghiên cứu này đạt kết quả tốt hơn mô hình dựa trên transformer trước đó[7].

Tuy nhiên, vũ khí chính mà Fragmentarium đang dùng trong thực tế hiện nay không phải là hình dạng 3D. Đó là ảnh và bản phiên âm, tức chữ và hình[1][2]. Việc ghép mặt đứt gãy đòi hỏi phải quét từng mảnh một, nên tốn nhiều công sức. Hai con đường này không thay thế nhau. Chúng đi cùng nhau.

Việc tìm ký hiệu trong ảnh cũng tiến nhanh. Vào tháng 6 năm 2026, nhóm nghiên cứu eBL công bố một bài báo mới. Đó là phương pháp tự động tìm vị trí ký hiệu nêm trong ảnh bảng đất sét. Số ảnh mảnh vỡ được xử lý vượt quá 87.000 tấm. Số ký hiệu tìm được lên tới 2,9 triệu. Kết quả cao hơn nhiều so với các phương pháp trước[11].

Tài liệu tham khảo

[1] Cobanoglu, Y., Laasonen, J., Simonjetz, F., Khait, I., Cohen, S., Földi, Z., Hätinen, A., Heinrich, A., Mitto, T., Rozzi, G., Sáenz, L., & Jiménez, E. (2024). Transliterated cuneiform tablets of the electronic Babylonian Library platform. Journal of Open Humanities Data, 10, 19. https://doi.org/10.5334/johd.148

[2] Jiménez, E. (2023, 5월 25일). AI is helping us read ancient Mesopotamian literature. The Conversation. https://theconversation.com/ai-is-helping-us-read-ancient-mesopotamian-literature-205091 (truy cập ngày 3 tháng 9 năm 2026)

[3] Gutherz, G., Gordin, S., Sáenz, L., Levy, O., & Berant, J. (2023). Translating Akkadian to English with neural machine translation. PNAS Nexus, 2(5), pgad096. https://doi.org/10.1093/pnasnexus/pgad096

[4] Wang, Z. (2026). TabletCraft: Bridging a 4,000-year cultural gap with bidirectional Akkadian NMT and cuneiform rendering. arXiv:2608.02609. https://arxiv.org/abs/2608.02609 (truy cập ngày 3 tháng 9 năm 2026)

[5] Julius-Maximilians-Universität Würzburg. (2026, 5월 18일). AI reads cuneiform: A milestone for Ancient Near Eastern Studies. EurekAlert!. https://www.eurekalert.org/news-releases/1128630 (truy cập ngày 3 tháng 9 năm 2026)

[6] Electronic Babylonian Literature. (n.d.). ngram-matcher [소프트웨어]. GitHub. https://github.com/ElectronicBabylonianLiterature/ngram-matcher (truy cập ngày 3 tháng 9 năm 2026)

[7] Hagelskjær, F. (2026). A novel network for classification of cuneiform tablet metadata. arXiv:2603.03892. https://arxiv.org/abs/2603.03892 (truy cập ngày 3 tháng 9 năm 2026)

[8] electronic Babylonian Library. (n.d.). Fragmentarium. ebl.lmu.de. https://www.ebl.lmu.de/fragmentarium (truy cập ngày 3 tháng 9 năm 2026)

[9] International Association for Assyriology. (2022, 10월 22일). In the spotlight: The "Virtual Cuneiform Tablet Reconstruction Project". iaassyriology.com. https://iaassyriology.com/in-the-spotlight-the-vctr-project/ (truy cập ngày 3 tháng 9 năm 2026)

[10] Collins, T., Woolley, S. I., Hernandez Munoz, L., Lewis, A., Ch'ng, E., & Gehlken, E. (2014). Computer-assisted reconstruction of virtual fragmented cuneiform tablets. In Proceedings of the 2014 International Conference on Virtual Systems and Multimedia (pp. 70-77). IEEE. https://doi.org/10.1109/VSMM.2014.7136691

[11] Che, W., Garcés Arias, E., Niaz, A., Bender, A., & Jiménez, E. (2026). Automated sign detection across the Electronic Babylonian Library: A large-scale dataset and end-to-end cuneiform OCR pipeline. arXiv:2606.22608. https://arxiv.org/abs/2606.22608 (truy cập ngày 3 tháng 9 năm 2026)

5.3. Sự thật được làm sáng tỏ

Những dòng chữ trở lại với Sử thi Gilgamesh

Fragmentarium đã tìm mới 20 mảnh vỡ của Sử thi Gilgamesh. Nhờ đó, hơn 100 dòng văn bản trở về đúng vị trí. Một số cảnh từng biến mất cũng được hồi sinh. Một trong số đó là cảnh Enkidu ngăn Gilgamesh, khuyên ông đừng giết Humbaba. Cảnh hai người giết Humbaba rồi đi đến Nippur cũng được nối lại. Nippur là đô thị trung tâm của tôn giáo thờ thần Enlil[1].

Một trường hợp cho thấy phương pháp này nhanh đến mức nào xuất hiện vào năm 2025. Ammar Fadhil của Đại học Baghdad và Giáo sư Jimenez đã hồi sinh một bài thơ dài ca ngợi Babylon. Bài thơ này đã bị quên lãng hơn 1.000 năm. Nhóm nghiên cứu dùng nền tảng eBL để tìm ra 30 bản sao thuộc cùng bài thơ. Nếu làm bằng tay, việc đó sẽ mất hàng chục năm. Khi gom các bản sao rải rác lại, phần lớn bài thơ dài 250 dòng đã được nối lại[3].

Bài thơ này cho thấy một ngày trong thành phố Babylon. Trong đó có nước sông, cổng thành và công việc của con người. Nó cũng chứa ghi chép mới về công việc do các nữ tư tế đảm nhận. Việc có tới 30 bản sao xuất hiện cho thấy bài thơ này từng được đọc rộng rãi. Khả năng cao đây là văn bản học sinh sao chép ở trường[3].

Thời gian cần cho công việc này cho thấy sức mạnh của phương pháp. Muốn tìm 30 bản sao bằng tay, phải lật từng danh mục của nhiều bảo tàng trên thế giới. Trên nền tảng eBL, cùng việc đó kết thúc bằng một lần tìm kiếm[3].

Phương pháp này không chỉ được dùng cho Sử thi Gilgamesh. Nó cũng được dùng cho sách bói toán, sách y học, từ điển và lời cầu nguyện của Babylonia. Các mối ghép mà eBL tìm được trải rộng trên nhiều loại văn bản[5].

Rừng tuyết tùng được bàn tay con người giải trước

Máy móc không giải được mọi thứ. Phát hiện mới được biết rộng rãi trong Sử thi Gilgamesh đến từ bàn tay con người.

Năm 2011, Bảo tàng Sulaymaniyah ở Iraq mua một bảng đất sét. Farouk Al-Rawi và Andrew George đã đọc bảng đất sét này. Hai người công bố kết quả vào năm 2014. Tấm bảng chứa phần đầu và phần cuối của bảng thứ năm trong bản chuẩn. Từ đây, khoảng 20 dòng văn bản mới xuất hiện[4].

Phần văn bản mới vẽ rừng tuyết tùng bằng âm thanh. Khỉ kêu ồn ào, ve sầu hòa giọng. Đủ loài chim hót líu lo. Những âm thanh này làm Humbaba, kẻ canh giữ khu rừng, vui mỗi ngày. Cũng có cảnh hai người lần đầu nhìn thấy khu rừng thì kinh ngạc đứng lại. Sau khi giết Humbaba, Gilgamesh hối tiếc vì đã chặt cây[4]. Trên đồng bằng Mesopotamia không có khu rừng có khỉ sinh sống. Vì vậy, nguồn gốc của mô tả này vẫn là câu hỏi của giới học giả.

Con đường xuất hiện của bảng đất sét này không trong sạch. Bảo tàng mua nó từ một người buôn cổ vật. Đó là biện pháp nhằm ngăn cổ vật bị đào trộm ở Iraq sau chiến tranh tiếp tục bị phân tán. Không có ghi chép cho biết nó đến từ di chỉ nào[4]. Với loại cổ vật này, manh mối để tìm mảnh ghép còn ít hơn.

Phát hiện này xuất hiện dù không có trí tuệ nhân tạo. Nhưng từ nay, việc có thể tạo ra các phát hiện cùng loại thường xuyên đến đâu sẽ phụ thuộc vào công cụ. Để tấm bảng Sulaymaniyah được thế giới biết đến, cần có ngẫu nhiên và con mắt của con người. eBL muốn biến sự ngẫu nhiên đó thành tìm kiếm[5].

Câu chuyện con thuyền tròn

Câu chuyện đại hồng thủy cũng có tình hình tương tự. Nếu bảng thứ năm của Sử thi Gilgamesh chứa rừng tuyết tùng, thì bảng thứ mười một chứa trận lụt lớn[6]. Bảng đất sét sẽ xem dưới đây không phải là bản sao của Sử thi Gilgamesh. Đó là một tấm bảng cổ hơn, chứa cùng truyền thống kể chuyện lũ lụt[8].

Năm 2009, Tiến sĩ Irving Finkel của Bảo tàng Anh xem xét một bảng đất sét. Đó là tấm bảng một nhà sưu tập tư nhân mang đến bảo tàng. Ông cho rằng nó được viết vào khoảng năm 1750 trước Công nguyên. Kích thước khoảng 11,5 xentimét chiều ngang và 6 xentimét chiều dọc. Trên đó có 60 dòng chữ[8]. Nội dung là câu chuyện một vị thần chỉ thị cho một người đóng thuyền. Cũng có lời bảo đưa thú vật lên thuyền từng đôi một[7].

Điểm nổi bật là hình dạng con thuyền. Con thuyền này không phải chiếc hộp dài, mà là thuyền tròn. Nó có hình dạng giống loại thuyền thúng tròn gọi là coracle, từng dùng trên sông[7]. Trên bảng đất sét có ghi dây thừng và khung gỗ. Cũng có lượng bitum cần trét để nước không rò vào. Bitum là dầu tự nhiên màu đen và dính. Tiến sĩ Finkel xuất bản cách đọc này thành sách vào năm 2014[8].

Bảng đất sét này không phải do AI đọc. Con người đã đọc từng dòng. Phần của AI nằm ở nơi khác. Đó là gom các bản sao câu chuyện lũ lụt rải rác khắp thế giới về một chỗ và tìm những dòng trùng nhau[5].

Việc máy làm và việc người làm

Lĩnh vực này chia kết quả thành ba tầng để xử lý.

Tầng thứ nhất là ghép vật lý. Đó là trường hợp mặt gãy của hai mảnh thực sự khớp vào nhau. Khi đó, có thể xác định hai mảnh đến từ một bảng đất sét duy nhất.

Tầng thứ hai là ghép văn bản. Đó là trường hợp hai mảnh không trực tiếp chạm vào nhau, nhưng được xác định là các phần khác nhau của cùng một tác phẩm. Phần lớn kết quả Fragmentarium tìm được thuộc loại này[5]. Hàng nghìn mảnh thuộc cùng tác phẩm đã được xác nhận theo cách đó[2].

Hai tầng này có tính chất khác nhau. Ghép vật lý, nếu được xác nhận bằng mắt, sẽ trở thành sự thật. Ghép văn bản dựa vào xác suất, nên có thể thay đổi khi mảnh mới xuất hiện. eBL ghi chép hai tầng này riêng biệt[5].

Tầng thứ ba là diễn giải. Đó là việc xem văn bản được hồi sinh có nghĩa gì. Máy không biết tiếng khỉ trong rừng tuyết tùng đến từ ký ức nào. Máy cũng không trả lời được con thuyền tròn đã trở thành chiếc

tàu hình vuông trong Kinh Thánh như thế nào. Những câu hỏi này thuộc về các học giả nghiên cứu văn bản cổ và các nhà sử học.

Ứng viên do máy đưa ra phải qua con người xem xét thì mới trở thành ghi chép chính thức. Học giả mở danh sách ứng viên và nhìn ảnh của hai mảnh đặt cạnh nhau. Sau khi xem hình dạng chữ, khoảng cách dòng, thậm chí cả độ dày của tấm bảng, họ mới xác nhận mối ghép[5]. Việc của máy chỉ dừng ở chỗ chọn ra vài cặp đáng xem trong vô số cặp[2].

Một bài tổng thuật ra vào tháng 6 năm 2026 cũng vạch cùng ranh giới. AI hiện nay không tự giải mã được chữ hình nêm. Nó là công cụ giúp nghiên cứu của con người[9]. Paleographicum của Đại học Würzburg cũng vậy. Nó chỉ rút việc so sánh nét chữ từng mất 3 ngày xuống còn 5 phút. Việc quyết định so sánh cái gì và phán đoán kết quả vẫn là của con người[10].

Điều đã thay đổi là tốc độ. Trong 150 năm có 5.000 mối ghép, nhưng chỉ trong 5 năm đã tăng thêm 1.500 mối ghép[2]. Sự sống vĩnh cửu mà Gilgamesh đi tìm cuối cùng đã thuộc về câu chuyện của ông. Những câu văn vỡ nát nằm trong đất chờ 2.600 năm nay đang lần lượt trở về đúng chỗ.

Việc còn lại vẫn rất nhiều. Trong 500.000 bảng đất sét trên thế giới, chỉ một phần đã được đọc[11]. Số bảng đất sét đã đi vào dữ liệu phiên âm của eBL là 25.000 tấm[5]. 30.000 mảnh Hittite nay mới bắt đầu được sắp xếp bằng ảnh[10]. Công việc 3D ghép mặt gãy mới ở giai đoạn khởi đầu. Sinh viên thế hệ sau sẽ tiếp nhận danh sách này. Khi đó, những chỗ trống trong Sử thi Gilgamesh sẽ ít hơn bây giờ. Những mảnh vỡ mà ngọn lửa Nineveh để lại vẫn chưa được đọc hết. Công cụ để đọc chúng tăng lên qua từng năm.

Tài liệu tham khảo

[1] Ofgang, E. (2024, 8월 12일). Piecing together an ancient epic was slow work. Until AI got involved. The New York Times. https://edition.pagesuite.com/tribune/article_popover.aspx?guid=235bb893-7f2d-41cd-9857-139c7d1e8355 (truy cập ngày 3 tháng 9 năm 2026)

[2] Jiménez, E. (2023, 5월 25일). AI is helping us read ancient Mesopotamian literature. The Conversation. https://theconversation.com/ai-is-helping-us-read-ancient-mesopotamian-literature-205091 (truy cập ngày 3 tháng 9 năm 2026)

[3] Fadhil, A. A., & Jiménez, E. (2025). Literary texts from the Sippar Library V: A hymn in praise of Babylon and the Babylonians. Iraq, 1-58. https://doi.org/10.1017/irq.2024.23

[4] Al-Rawi, F. N. H., & George, A. R. (2014). Back to the Cedar Forest: The beginning and end of Tablet V of the Standard Babylonian Epic of Gilgameš. Journal of Cuneiform Studies, 66, 69-90. https://doi.org/10.5615/jcunestud.66.2014.0069

[5] Cobanoglu, Y., Laasonen, J., Simonjetz, F., Khait, I., Cohen, S., Földi, Z., Hätinen, A., Heinrich, A., Mitto, T., Rozzi, G., Sáenz, L., & Jiménez, E. (2024). Transliterated cuneiform tablets of the electronic Babylonian Library platform. Journal of Open Humanities Data, 10, 19. https://doi.org/10.5334/johd.148

[6] George, A. R. (2003). The Babylonian Gilgamesh Epic: Introduction, Critical Edition and Cuneiform Texts. Oxford University Press.

[7] PBS NewsHour. (2014, 1월 24일). New evidence suggests Noah's ark may have been round. pbs.org. https://www.pbs.org/newshour/world/new-evidence-suggests-noahs-ark-may-have-been-round (truy cập ngày 3 tháng 9 năm 2026)

[8] Finkel, I. (2014). The Ark Before Noah: Decoding the Story of the Flood. Hodder & Stoughton.

[9] Centre for the Study of Manuscript Cultures. (2026, 6월 14일). Can artificial intelligence decipher cuneiform tablets? Universität Hamburg. https://www.csmc.uni-hamburg.de/publications/mesopotamia/2026-06-14.html (truy cập ngày 3 tháng 9 năm 2026)

[10] Julius-Maximilians-Universität Würzburg. (2026, 5월 18일). AI reads cuneiform: A milestone for Ancient Near Eastern Studies. EurekAlert!. https://www.eurekalert.org/news-releases/1128630 (truy cập ngày 3 tháng 9 năm 2026)

[11] Wang, Z. (2026). TabletCraft: Bridging a 4,000-year cultural gap with bidirectional Akkadian NMT and cuneiform rendering. arXiv:2608.02609. https://arxiv.org/abs/2608.02609 (truy cập ngày 3 tháng 9 năm 2026)

Chương 6. Biên lai Lưỡng Hà cổ đại và lời thì thầm của các thương nhân: Dịch bằng AI

6.1. Hồ sơ vụ việc

Biên lai khắc trên đất sét

Người Lưỡng Hà cổ đại viết trên bảng đất sét ướt. Họ vót đầu cây sậy thành góc cạnh để làm bút khắc. Bút khắc là dụng cụ viết chỉ để lại dấu, không dùng mực. Họ ấn bút khắc này xuống đất sét rồi nhấc lên ngay[1]. Dấu bị ấn tạo thành hình tam giác giống cái nêm. Tên của loại chữ này là chữ hình nêm. Giới học thuật cũng dùng tên chữ hình nêm. Hình nêm có nghĩa là dạng cái nêm.

Đất sét khô lại thì cứng như đá. Khác với giấy hay da, nó không mục nát theo thời gian. Dù gặp lửa, nó cũng không thành tro. Trái lại, sức nóng nung đất sét cứng như gạch. Khi chiến tranh đốt cháy thành phố, bảng đất sét lại bền hơn[3]. Nhờ vậy, sổ sách từ 4 nghìn năm trước còn lại đến hôm nay.

Nhiều bảng đất sét chỉ lớn bằng lòng bàn tay. Đó là kích thước có thể bỏ túi mang đi. Người thư lại nhào đất sét ướt rồi nặn thành bảng. Thư lại là người viết và đọc thay cho người khác. Khi viết xong, họ phơi dưới nắng. Việc cố ý nung trong lò hiếm khi xảy ra. Những bảng được nung trong bảo tàng ngày nay là biện pháp để bảo tồn[2]. Khi cần sửa, họ chà bề mặt để xóa.

Số bảng đất sét đã khai quật đến nay ước khoảng 500 nghìn tấm[4]. Dưới đất còn nhiều bảng đất sét chưa được đào lên. Những chữ viết trên đó không chỉ là lời khoe công của vua. Có biên lai ghi đã nhận mấy bao ngũ cốc. Có hợp đồng cho vay bạc và ấn định lãi. Cũng có thư người vợ gửi cho người chồng đi xa[6].

Một tờ biên lai thì ngắn. Tên hàng hóa và số lượng xuất hiện trước. Sau đó là tên người đưa và người nhận. Cuối cùng ghi ngày tháng và tên nhân chứng. Thứ tự này giống biên lai nhận ở cửa hàng ngày nay. Một văn bản như vậy có nội dung nhỏ. Nhưng khi hàng chục nghìn văn bản gom lại, câu chuyện sẽ khác. Ta thấy năm nào giá ngũ cốc tăng. Ta cũng thấy thành phố nào mua bán thứ gì.

Phần lớn chưa được dịch

Vấn đề nằm ở chỗ loại chữ này hầu như chưa được đọc. Trong số các bảng đất sét, chỉ một phần đã được chuyển sang ngôn ngữ hiện đại[5]. Phần còn lại đang ngủ yên trong hộp bảo tàng.

Không phải ai cũng có thể trở thành người viết chữ. Muốn thành thư lại, phải học lâu trong trường. Học sinh viết lặp lại các ký hiệu trên bảng đất sét. Bảng viết sai được hòa với nước rồi nặn lại. Những bảng luyện tập còn lại cũng được khai quật như di vật. Bảng luyện tập cho thấy lớp học của thời đại đó.

Người biết chữ rất ít. Phần lớn nhờ thư lại viết thay. Khi nhận thư, thư lại đọc cho họ nghe. Đây là lý do câu văn trong thư có khuôn mẫu trang trọng. Người nói và người ghi là hai người khác nhau.

Để đọc một bảng đất sét, cũng phải qua nhiều bước. Trước hết nhận ra từng ký hiệu. Tiếp theo xác định ký hiệu đó mang âm nào. Sau đó tách từ và tìm nghĩa. Cuối cùng ghép câu để dịch. Nếu là biểu mẫu quen thuộc thì xong nhanh. Với văn bản lạ, phải giữ nó trên bàn rất lâu.

Lý do là thiếu người đọc. Ngành nghiên cứu chữ hình nêm gọi là Assyriology. Để đào tạo một học giả cần vài chục năm. Vì họ phải học cả ngôn ngữ cổ và lịch sử cổ. Vì vậy, số người đọc được loại chữ này trên thế giới rất ít[5].

Số bảng đất sét mới xuất hiện vượt quá tốc độ đọc của các học giả. Người không tăng, nên hộp cứ chất lên. Tháng 6 năm 2026, Đại học York của Anh khởi động một dự án nghiên cứu mới. Tên dự án là

“Hướng tới khả năng tiếp cận phổ quát di sản chữ hình nêm”. Đây là kế hoạch chuyển tư liệu chữ hình nêm sang cả tiếng Ả Rập để ai cũng có thể xem. Dự án này kéo dài 3 năm, đến năm 2029[7].

Việc gom tư liệu về một nơi cũng đang được tiến hành. Dự án Thư viện Số Chữ Hình Nêm là một công việc như vậy. Dự án gom thông tin về các bảng đất sét rải rác trên thế giới và công bố trên internet. Ảnh, phiên âm và bản dịch được đặt cạnh nhau. Phiên âm là văn bản ghi âm của ký hiệu hình nêm bằng bảng chữ cái. Đến nay, hơn 400 nghìn di vật đã được đưa vào danh mục[4]. Khi tư liệu được gom lại, máy móc cũng có thứ để học.

Bảng đất sét hiện rải rác ở nhiều nước. Nhiều bảng nằm trong bảo tàng ở Iraq và Thổ Nhĩ Kỳ. Các bảo tàng ở London, Paris và Berlin cũng nắm giữ một phần lớn. Chúng cũng được chia ra tại các đại học ở Mỹ và Nhật Bản. Nếu mặt trước và mặt sau của một văn bản nằm ở hai nước khác nhau, rất khó ghép lại. Chỉ khi ảnh được đưa lên internet cùng nhau, người ta mới có thể thử khớp chúng.

Sổ sách bị hồ sơ của vua lấn át

Có một lý do lâu đời khiến văn bản đời thường bị đẩy ra sau. Các đoàn khai quật thế kỷ 19 và đầu thế kỷ 20 đào cung điện và đền thờ trước. Những bia đá lớn và đồ vàng thu hút ánh mắt của mọi người. Các học giả cũng đọc minh văn hoàng gia và bộ luật trước. Họ dành thời gian trước hết để chuyển ngữ thần thoại và sử thi.

Trong lúc đó, biên lai và thư của thương nhân bị đẩy xuống sau. Danh sách cấp phát ngũ cốc bị xem như ghi chép việc vặt. Nếu tách từng tấm ra xem, chúng có vẻ không có nội dung đáng kể. Các học giả hoãn việc chép tay những văn bản buồn tẻ này. Vì thế, tiếng nói của người bình thường bị chôn vùi rất lâu.

Sự lệch hướng này đã làm mất nhiều thứ. Ta khó biết giá cả lên xuống ra sao. Việc phụ nữ đã làm bị che khuất. Tên trẻ em và người hầu chỉ còn trong danh sách. Minh văn của vua ghi lời tự hào. Biên lai không khoe khoang. Nó chỉ ghi sự thật giao dịch. Vì vậy, biên lai đôi khi trở thành tư liệu trung thực hơn.

Một dấu nêm có nhiều nghĩa

Ngay cả khi muốn đọc bằng máy, chữ hình nêm cũng không dễ. Vì một ký hiệu đảm nhiệm chồng lên nhau nhiều vai trò[8]. Ở vị trí này, nó trở thành ký hiệu âm tiết biểu thị âm thanh. Ở vị trí khác, nó biểu thị trọn nghĩa của một từ. Ở vị trí khác nữa, nó không phát âm. Nó chỉ được dùng làm dấu hiệu cho biết từ phía sau là cây hay thần. Dấu hiệu như vậy gọi là định từ.

Chữ hình nêm không phải là chữ viết chỉ ghi một ngôn ngữ. Tiếng Sumer và tiếng Akkad cùng dùng chung loại chữ này. Về sau, tiếng Hittite và tiếng Elam cũng mượn loại chữ này. Cùng một ký hiệu phát âm khác nhau trong từng ngôn ngữ. Người đọc phải xác định ngôn ngữ của văn bản trước.

Ký hiệu Sumer gish, nghĩa là cây, là một ví dụ hay. Khi đứng trước tên đồ nội thất, nó là dấu hiệu cho biết vật liệu là gỗ. Khi đứng một mình, nó chính là từ cây. Ở vị trí khác, nó chỉ biểu thị âm gish. Người đọc phải xem câu trước sau rồi chọn vai trò của nó.

Cách ghi số cũng khác hiện nay. Người Lưỡng Hà lấy 60 làm một đơn vị. 60 phút và 60 giây trên đồng hồ bắt nguồn từ đó. Số lượng trong biên lai cũng được ghi theo cách này. Ngũ cốc và bạc dùng các đơn vị khác nhau. Dù cùng một ký hiệu, giá trị khác nhau khi là ngũ cốc và khi là bạc. Máy phải học cả khác biệt về đơn vị này.

Việc viết liền các từ cũng là trở ngại. Các thư lại xưa không chừa khoảng cách giữa các từ. Máy tính khó biết từ ngắt ở đâu[8]. Bảng đất sét bị vỡ và mòn cũng là vấn đề. Khi dấu nêm nông đi, ngay cả mắt người cũng khó phân biệt. Hình thái từ trong tiếng Akkad thường xuyên thay đổi cũng gây khó. Một gốc từ gắn nhiều đuôi khác nhau. Âm cuối thay đổi tùy theo nó là chủ ngữ hay tân ngữ. Nó lại thay đổi theo thì và ngôi. Trước hết phải tìm ra dạng cơ bản để tra trong từ điển.

Bảng đất sét vỡ vì nhiều lý do. Chúng bị ép dưới đất và bị nước làm ướt. Trong lúc khai quật, góc bảng cũng có thể rơi ra. Do buôn bán cổ vật thời trước, các bảng đất sét bị vỡ đã rải rác ở nhiều nước. Sau khi đưa lên khỏi đất, muối thấm ra ngoài và phủ lên chữ[2]. Các học giả gọi khoảng trống nơi chữ biến mất là phần khuyết.

Vì nhiều lớp tường này, hàng trăm nghìn sổ sách đã im lặng đến nay.

Tài liệu tham khảo

[1] Truman State University. (2026). How it works. Truman State University Cuneiforms. https://exhibits.truman.edu/cuneiform/how-it-works/ (truy cập ngày 3 tháng 9 năm 2026)

[2] Guinan, A., Oller, G., & Ormsby, D. (1976). Nippur rebaked: The conservation of cuneiform tablets. Expedition, 18(3). https://www.penn.museum/sites/expedition/nippur-rebaked/ (truy cập ngày 3 tháng 9 năm 2026)

[3] Bouscaren, D. (2023). Assyrian women of letters. Archaeology, 76(6). https://archaeology.org/issues/november-december-2023/features/assyrian-women-of-letters/ (truy cập ngày 3 tháng 9 năm 2026)

[4] Cuneiform Digital Library Initiative. (2026). About the CDLI. cdli.earth. https://cdli.earth/about (truy cập ngày 3 tháng 9 năm 2026)

[5] Gutherz, G., Gordin, S., Sáenz, L., Levy, O., & Berant, J. (2023). Translating Akkadian to English with neural machine translation. PNAS Nexus, 2(5), pgad096. https://doi.org/10.1093/pnasnexus/pgad096

[6] Michel, C. (2020). Women of Assur and Kanesh: Texts from the Archives of Assyrian Merchants (Writings from the Ancient World 42). SBL Press.

[7] Cuneiform Digital Library Initiative. (2026). Research Programme "Towards Universal Access to Cuneiform Heritage" 2026-2029. cdli.earth. https://cdli.earth/postings/239 (truy cập ngày 3 tháng 9 năm 2026)

[8] Gordin, S., Gutherz, G., Elazary, A., Romach, A., Jiménez, E., Berant, J., & Cohen, Y. (2020). Reading Akkadian cuneiform using natural language processing. PLoS ONE, 15(10), e0240511. https://doi.org/10.1371/journal.pone.0240511

6.2. Cuộc truy dấu của đội điều tra khoa học: Dịch AI tiếng Akkad

Mở hai lối dịch

Năm 2023, nhóm nghiên cứu Đại học Tel Aviv mở đường trước. Gai Gutherz và các đồng nghiệp đã tạo ra một máy dịch tiếng Akkad sang tiếng Anh[1]. Tiếng Akkad là ngôn ngữ được dùng lâu đời ở Lưỡng Hà. Nó được ghi bằng chữ hình nêm, và hiện không còn người sử dụng.

Nhóm nghiên cứu mở hai lối dịch. Lối thứ nhất có tên C2E. Đây là đường đưa vào các ký hiệu hình nêm đã chuyển sang chữ máy tính rồi nhận tiếng Anh. Đây không phải là cách đọc trực tiếp ảnh bảng đất sét. Vì không qua phiên âm của học giả, các bước ngắn hơn. Lối thứ hai có tên T2E. Đây là đường đưa vào văn bản mà học giả đã chuyển âm của ký hiệu hình nêm sang bảng chữ cái. Cách chuyển sang bảng chữ cái này gọi là phiên âm. Ở bước phiên âm, nghĩa của ký hiệu được sắp xếp một lần. Vì vậy, bản dịch phía T2E chính xác hơn[1].

Việc tạo phiên âm cũng không dễ. Học giả nhìn ký hiệu rồi xác định âm. Nếu có nhiều âm, họ xem ngữ cảnh rồi chọn. Việc chọn này chính là phán đoán. Một văn bản đã đi qua phán đoán một lần thì máy dễ xử lý hơn. Đây là lý do điểm của T2E cao hơn.

Bên trong máy dịch được chia làm hai. Phần trước đổi nguyên văn thành các khối số. Phần sau nhìn các con số đó rồi chọn từng từ tiếng Anh. Phần trước gọi là bộ mã hóa, phần sau gọi là bộ giải mã. Bộ mã hóa chứa cả nghĩa và thứ tự của từ. Bộ giải mã nhìn những từ đã chọn trước đó rồi quyết định từ tiếp theo.

Nghiên cứu đọc ký hiệu trực tiếp từ ảnh đang tiếp tục theo hướng riêng. Tháng 6 năm 2026, nhóm nghiên cứu Thư viện Babylonia Điện tử công bố kết quả. Máy đã quét ảnh của hơn 87 nghìn mảnh bảng

đất sét vỡ. Số dấu ký hiệu tìm được gần 2,9 triệu. Kết quả phát hiện tăng tối đa 37 phần trăm so với phương pháp trước. Nó vẫn còn yếu với bảng đất sét vỡ nhiều và các dòng bị xáo trộn[6].

Máy học từ 8 nghìn văn bản

Máy học dịch bằng cách xem câu mẫu. Nhóm nghiên cứu gom văn bản từ bộ tư liệu công khai tên Oracc. Oracc là bộ tư liệu internet sắp xếp và công bố văn hiến chữ hình nêm. Số văn bản gom từ đó là 8.056. Tính theo câu thì vượt quá 50 nghìn câu[1].

Các văn bản gom được có nhiều loại. Khoảng 3 nghìn văn bản là minh văn khắc chiến công của vua. Khoảng 2 nghìn văn bản là thư hành chính. Hồ sơ xét xử và báo cáo chiêm tinh cũng có trong đó[1]. Bên cạnh mỗi văn bản có bản dịch tiếng Anh do học giả tạo ra. Máy nhìn các cặp này hàng chục nghìn lần và tự tìm quy tắc.

Quá trình máy học giống việc ôn thi. Trước hết nó xem câu hỏi và đáp án cùng nhau. Sau đó che đáp án và tự giải. Nếu sai, nó lần lại xem sai ở đâu. Nó lặp lại việc lần lại này hàng triệu lần. Điểm khác con người là tốc độ và sự bền bỉ.

Máy không xử lý cả từ như một khối. Nó chia từ thành những mảnh nhỏ hơn để xử lý. Mảnh này gọi là token. Từ hiếm cũng có thể xử lý nếu chia thành mảnh. Cách này phù hợp với tiếng Akkad, nơi nhiều đuôi từ được gắn vào.

Chất lượng dịch được đo bằng điểm BLEU. Đây là cách đếm bản dịch của máy và bản dịch của người trùng nhau bao nhiêu. Điểm được tính theo tỷ lệ các cụm từ trùng nhau. Nó được biểu thị từ 0 đến 100 điểm. Nhiều khi vượt 30 điểm thì được xem là truyền đạt được nghĩa. Nhưng đó không phải là chuẩn tuyệt đối. Cách diễn giải thay đổi theo loại văn bản và độ dài câu.

Nhóm nghiên cứu chọn cấu trúc gọi là mạng nơ-ron tích chập. Đây là cấu trúc học bằng cách gom các thông tin gần nhau. Lý do là khi dữ liệu ít, việc học nhanh và ổn định. Kết quả kiểm chứng cho thấy C2E đạt 36,52 điểm. T2E cao hơn, đạt 37,47 điểm. Phương pháp tìm kiếm tự động từng dùng trước đây chỉ đạt 27,09 điểm và 23,51 điểm. Kết quả tốt ở những văn bản có câu ngắn và biểu mẫu cố định[1].

Cách chia dữ liệu cũng quan trọng. Phần để học và phần để kiểm tra được tách riêng. Nếu cho xem trước văn bản dùng để kiểm tra, điểm số sẽ vô nghĩa. Nó giống như biết trước đề trong kỳ thi của con người. Nhóm nghiên cứu giữ đúng sự phân chia này rồi tính điểm.

Cùng một nguyên văn cũng được mỗi người dịch khác nhau. Vì vậy, không thể chỉ nhìn điểm để quyết định tốt hay xấu. Ngoài điểm số, họ cũng dùng đánh giá của con người. Họ đưa bản dịch cho học giả xem và cho xếp hạng. Họ hỏi riêng xem nghĩa có thông hay không. Họ cũng hỏi riêng xem có chỗ sai hay không. Khoảng 44 phần trăm bản dịch T2E được học giả xem là có thể dùng. C2E là khoảng 38 phần trăm[1]. Có lúc điểm số và đánh giá của con người lệch nhau. Khi đó, phán đoán của con người được đặt lên trước.

Dịch tiếng Akkad thuộc bài toán ít tài nguyên. Ít tài nguyên có nghĩa là có ít câu mẫu để học. Máy dịch tiếng Anh và tiếng Pháp học hàng chục triệu câu. Các cặp câu mà tiếng Akkad có chưa bằng một phần nghìn con số đó. Khi dữ liệu ít, mô hình lớn lại dễ lúng túng.

Năm 2025, một nhiệm vụ chung để nhiều nhóm nghiên cứu cạnh tranh cũng được mở. Nhiệm vụ là tìm dạng cơ bản của từ và khớp chữ bị thiếu. Dữ liệu được lấy từ Thư viện Babylonia Điện tử và Archibab. Khi đo kết quả bằng cùng một dữ liệu, phương pháp nào tốt hơn lộ ra ngay[2].

Tìm dạng cơ bản là công việc nền cho dịch thuật. Trong học thuật, việc này gọi là tìm lemma. Đó là đưa một từ đã dài ra vì gắn đuôi trở về dạng từ điển. Chỉ khi công việc nền này xong, việc tìm kiếm và thống kê theo từng từ mới có thể thực hiện.

Bắt quả tang lời nói dối có vẻ thật

Dịch máy có một thói quen nguy hiểm. Đó là tạo ra câu nghe có vẻ hợp lý dù không có căn cứ. Thói quen này được gọi là ảo giác. Nhóm nghiên cứu chia ảo giác thành hai loại để xem xét. Một loại là nhìn vào nguyên văn nhưng nối nghĩa sai. Loại kia là bỏ qua nguyên văn và dựng ra chuyện khác[1].

Ảo giác dễ xuất hiện ở các bảng đất sét bị vỡ. Vì máy lấp các chỗ trống nơi chữ đã mất bằng tưởng tượng. Vì vậy bản dịch nhất định phải được học giả xem lại. Câu máy đưa ra là bản nháp, không phải phán quyết cuối cùng.

Có nhiều cách để giảm ảo giác. Đánh dấu rõ chỗ vỡ bằng ô trống. Để máy trả lời rằng nó không biết đoạn đó. Cũng có thể dịch cùng một câu nhiều lần. Nếu câu trả lời lần nào cũng khác, thì không tin. Học giả nhìn các dấu này để quyết định cần xem lại chỗ nào.

Việc chuyển ký hiệu hình nêm sang chữ cái và tách từ đi trước việc dịch. Trong cùng hướng nghiên cứu, người ta cũng thử giao việc này cho máy. Mạng thần kinh gọi là LSTM hai chiều đạt kết quả tốt hơn. Độ chính xác cao nhất đạt 97 phần trăm[3].

Cánh cửa mở theo hai hướng

Các máy dịch cho đến nay chỉ chảy theo một chiều. Chúng chỉ chuyển chữ cổ sang tiếng Anh hiện đại. Vào tháng 5 năm 2026, Wang Zhaohui của Đại học Nam California ở Hoa Kỳ đã mở chiều ngược lại. Công cụ ông tạo ra có tên là TabletCraft[4].

TabletCraft cũng dịch tiếng Anh sang tiếng Akkad. Nó vẽ kết quả dịch thành ký hiệu hình nêm và còn tạo cả hình bảng đất sét. Việc huấn luyện dùng 116.000 cặp câu song ngữ. Bảng đối ứng ký hiệu có hơn 14.000 mục. Kết quả kiểm tra là 49,1 điểm khi dịch từ tiếng Akkad sang tiếng Anh. Khi dịch từ tiếng Anh sang tiếng Akkad, kết quả là 48,5 điểm[4].

Công cụ này đã được dùng trong thiết bị trải nghiệm ở bảo tàng và trong giờ học đại học. Người tạo ra nó cũng nêu rõ một giới hạn. Không được trích dẫn tiếng Akkad do máy tạo ra như thể đó là ghi chép cổ thật. Loại văn bản công cụ xử lý tốt là minh văn hoàng gia thời Tân Assyria[5].

Dịch hai chiều có ích. Sinh viên có thể thử đổi câu của mình thành ký hiệu hình nêm. Khi tự vẽ ký hiệu, cấu trúc của chữ hiện ra trước mắt. Ở bảo tàng, khách tham quan thử khắc tên mình. Chữ cổ chuyển từ vật để ngắm thành công cụ để thao tác.

Các nghiên cứu trên đã công bố cả dữ liệu lẫn chương trình[1][5]. Khi công khai, nhóm khác có thể lặp lại cùng phép thử. Nếu có chỗ sai, nó sẽ lộ ra nhanh. Nghiên cứu ngôn ngữ cổ có ít người tham gia. Việc công khai quyết định lớn đến tốc độ tiến bộ.

Cách dùng máy dịch cũng đang được sắp xếp lại. Học giả hiển thị bản nháp của máy trên màn hình. Họ đặt nguyên văn và bản dịch cạnh nhau để đối chiếu. Họ để lại dấu ở những từ đáng ngờ. Khi các dấu tích lại, chúng trở thành dữ liệu huấn luyện tiếp theo.

Tài liệu tham khảo

[1] Gutherz, G., Gordin, S., Sáenz, L., Levy, O., & Berant, J. (2023). Translating Akkadian to English with neural machine translation. PNAS Nexus, 2(5), pgad096. https://doi.org/10.1093/pnasnexus/pgad096

[2] Gordin, S., Sahala, A., Spencer, S., & Klein, S. (2025). EvaCun 2025 shared task: Lemmatization and token prediction in Akkadian and Sumerian using LLMs. In Proceedings of the Second Workshop on Ancient Language Processing (pp. 242-250). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.alp-1.33

[3] Gordin, S., Gutherz, G., Elazary, A., Romach, A., Jiménez, E., Berant, J., & Cohen, Y. (2020). Reading Akkadian cuneiform using natural language processing. PLoS ONE, 15(10), e0240511. https://doi.org/10.1371/journal.pone.0240511

[4] Wang, Z. (2026, 5월 14일). TabletCraft: Bridging a 4,000-year cultural gap with bidirectional Akkadian NMT and cuneiform rendering. arXiv:2608.02609 (2026년 5월 14일 제출, 2026년 8월 공개). https://arxiv.org/abs/2608.02609 (truy cập ngày 3 tháng 9 năm 2026)

[5] Buyukyildirim, K. (2026, 8월 11일). TabletCraft opens a two-way AI translation channel for archaeologists. Arkeonews. https://arkeonews.net/tabletcraft-opens-a-two-way-ai-translation-channel-for-archaeologists/ (truy cập ngày 3 tháng 9 năm 2026)

[6] Che, W., Garces Arias, E., Niaz, A., Bender, A., & Jimenez, E. (2026, 6월 21일). Automated sign detection across the Electronic Babylonian Library: A large-scale dataset and end-to-end cuneiform OCR pipeline. arXiv:2606.22608. https://arxiv.org/abs/2606.22608 (truy cập ngày 3 tháng 9 năm 2026)

6.3. Sự thật được làm sáng tỏ

Những bức thư của thương nhân Kanesh

Gần Kayseri ở miền trung Thổ Nhĩ Kỳ có một gò đất tên là Kültepe. 4.000 năm trước, nơi này có tên là Kanesh. Các thương nhân Assyria đặt cứ điểm thương mại ở đây. Cộng đồng cứ điểm của thương nhân được gọi là karum. Karum vốn là từ chỉ bến cảng bên sông. Nơi dỡ hàng đã trở thành khu vực của thương nhân.

Karum của Kanesh là tổ chức tự trị của các thương nhân Assyria. Họ tự đặt ra quy tắc và phân xử tranh chấp với nhau. Họ nộp thuế cho vua địa phương và được bảo đảm an toàn. Hàng hóa được chất trên lưng lừa để vận chuyển. Từ Assur đến Kanesh là một chặng đường xa, mất nhiều tuần.

Từ cuộc khai quật bắt đầu năm 1948, các bảng đất sét xuất hiện dồn dập. Số văn bản tìm được đến nay là khoảng 23.500 tấm. Đây là con số do Fikri Kulakoğlu, người chỉ huy khai quật, công bố[1]. Nội dung không phải lời khoe khoang của vua. Có giấy vay nợ và vận đơn hàng hóa. Vận đơn hàng hóa là bảng liệt kê số hàng được gửi đi. Có hồ sơ xét xử và thư gia đình.

Trong thư còn nguyên lời thúc giục và than phiền. Họ hỏi khi nào sẽ gửi tiền. Họ trách vì hàng đến muộn. Họ thêm lời chào nhắc giữ gìn sức khỏe. Có khi họ dặn con trai học hành. Nỗi lo của con người 4.000 năm trước vẫn nằm nguyên trong câu chữ.

Từ các văn bản này, khung của nền kinh tế cổ hiện ra. Các thương nhân chở vải và thiếc từ Assur đến. Ở Anatolia, họ đổi chúng lấy bạc và vàng. Cách nộp thuế và phí qua đường cũng được ghi lại. Cách xét xử khi phát sinh tranh chấp cũng còn lại.

Thiếc là kim loại cần thiết để làm đồng thau. Đồng thau được tạo bằng cách trộn đồng và thiếc. Ở Anatolia, thiếc rất khó kiếm. Thiếc phải được chở đến từ miền đông xa xôi[8]. Sự thiếu hụt này là sức đẩy tạo ra tuyến thương mại dài.

Phần việc của phụ nữ còn ở lại Assur cũng được làm rõ. Vợ và con gái của thương nhân dệt vải ở nhà rồi gửi đi[2][8]. Họ điều chỉnh sản lượng và mặc cả giá. Trong thư gửi chồng, họ cũng tính toán giá thị trường. Cécile Michel đã gom hơn 300 văn bản như vậy và dịch sang tiếng Anh. Michel làm việc tại Trung tâm Nghiên cứu Khoa học Quốc gia Pháp và Đại học Hamburg[2].

Đọc mà không bóc phong bì

Các thương nhân cổ đặt những bảng đất sét quan trọng vào phong bì bằng bùn rồi niêm phong. Muốn đọc bên trong thì phải đập vỡ phong bì. Nhưng trên phong bì cũng có dấu con dấu. Nếu đập vỡ phong bì, thông tin đó biến mất.

Phong bì là thiết bị chống giả mạo. Nội dung hợp đồng được viết trên bảng đất sét bên trong. Người ta phủ thêm bùn lên trên để làm phong bì. Trên mặt ngoài phong bì, họ viết lại cùng nội dung. Các nhân chứng lăn con dấu để in dấu. Nếu bên ngoài và bên trong khác nhau, việc sửa chữa sẽ lộ ra.

Cần có cách đọc bên trong mà vẫn giữ nguyên phong bì. Vào tháng 6 năm 2026, Bảo tàng Khảo cổ Kayseri đã cho thấy lời giải. Họ đọc được một bảng đất sét nằm trong phong bì mà không bóc nó ra. Đây là công việc có sự tham gia của Trung tâm Nghiên cứu Khoa học Quốc gia Pháp và Đại học Hamburg. Họ dùng phương pháp giống chụp CT, nhìn vào bên trong mà không cắt vật thể. Nội dung bên trong là cam kết giao dịch lúa mì và lúa mạch. Kulakoğlu nói rằng lần chụp này cho thấy cả một hạt lúa mạch[1]. Ảnh chữ thu được như vậy trở thành dữ liệu đầu vào cho máy dịch. Số văn bản đọc được phải tăng thì máy mới có thứ để học.

Cuộc thi dịch thuật trị giá 50.000 đô la

Một cuộc thi dịch thuật cũng đã được tổ chức. Một tổ chức tên là Deep Past Initiative chuẩn bị cuộc thi[3]. Cuộc thi diễn ra trên Kaggle, trang thi đấu về AI[4]. Nhiệm vụ là dịch các văn bản thương mại của Kanesh sang tiếng Anh[3]. Cuộc thi bắt đầu vào ngày 16 tháng 12 năm 2025. Hạn cuối tham gia là ngày 23 tháng 3 năm 2026. Giải thưởng là 50.000 đô la. Chi phí tổ chức do một công ty tài chính chi trả[7].

Dữ liệu cuộc thi được công khai để ai cũng có thể tải xuống[4]. Dữ liệu công khai vẫn còn sau khi cuộc thi kết thúc. Nhà nghiên cứu tiếp theo bắt đầu lại trên dữ liệu đó. Cấu trúc là một cuộc thi để lại một bộ dữ liệu.

Trong số các văn bản Kültepe, vẫn còn nhiều thứ chưa dịch được[3]. Việc khai quật đến nay vẫn tiếp tục hằng năm[1]. Khi bảng đất sét mới xuất hiện, nó được thêm vào danh mục. Nếu số người đọc không tăng, danh mục chỉ dài thêm. Đây là lý do cần sự trợ giúp của máy.

Tình trạng này không dừng lại ở một di tích. Ở nhiều bảo tàng trên thế giới, các văn bản tương tự đang chất lại. Ở đó cũng có những hộp chưa được đọc. Văn bản có mẫu cố định xuất hiện ở khắp nơi. Sổ ngũ cốc và sổ thuế là như vậy. Phương pháp đã có tác dụng ở Kanesh cũng có thể dùng cho các kho văn bản khác.

Thông báo cuộc thi đưa những nhân vật có thật trong văn bản lên phía trước. Đó là tên của các thương nhân như Ashur-nada và Ashur-idi. Tên Ishtar-lamassi và Ennam-Ashur cũng được nêu cùng[3]. Tên của một gia đình thương nhân 4.000 năm trước đã trở thành chủ đề cuộc thi.

Cách tổ chức cuộc thi có điểm tốt. Không chỉ vài phòng nghiên cứu, mà các nhà phát triển trên toàn thế giới cùng lao vào một vấn đề. Khi đo kết quả bằng cùng một thước, phương pháp nào tốt hơn sẽ hiện ra ngay. Tiền thưởng có sức kéo người tham gia.

Nhóm văn bản này là dữ liệu hợp với dịch máy. Mẫu văn bản cố định và câu ngắn. Cùng một cách diễn đạt lặp lại trong nhiều văn bản. Máy giỏi tìm ra các quy tắc lặp lại. Cũng có không ít văn bản đã được học giả dịch sẵn. Những bản dịch đó trở thành giáo trình cho máy.

Việc máy không làm được

Thành quả đã tích lại, nhưng những bức tường còn lại cũng rõ ràng. Bức tường thứ nhất là thiếu dữ liệu. Máy dịch của các ngôn ngữ ngày nay lớn lên nhờ hàng trăm triệu câu. Tiếng Akkad chỉ có vài chục nghìn câu ghép cặp[5]. Khi dữ liệu ít, máy dễ sai ở các địa danh và nhân danh hiếm gặp.

Bức tường thứ hai là nhiều nghĩa của ký hiệu. Cùng một ký hiệu có thể khiến người ta lẫn lộn giữa âm, từ và dấu chỉ[6]. Khi ngữ cảnh trước sau ngắn, máy cũng lúng túng.

Bức tường thứ ba là cách máy phán đoán. Máy tính xem từ nào có xác suất xuất hiện cao. Nó không thể phán đoán câu đó có đúng về mặt lịch sử hay không. Vì vậy bước kiểm chứng của học giả luôn cần ở cuối cùng[5]. Khi máy và người kết thành một cặp, bản dịch mới đáng tin.

Cách chia việc đang bén rễ. Máy quét qua hàng chục nghìn tấm và đưa ra nghĩa sơ bộ. Học giả chọn ra những văn bản quan trọng trong đó. Chỉ những văn bản đã chọn mới được đọc lại kỹ bằng tay. Làm như vậy, số văn bản được đọc tăng lên nhiều.

Bản dịch cần nhãn ghi chú. Cần ghi rõ đoạn nào do máy tạo ra. Chỗ con người sửa cũng phải được đánh dấu. Như vậy người sau mới có thể lần lại căn cứ. Nghiên cứu lịch sử bắt đầu từ việc lần lại căn cứ.

Tên gọi là phần khó trong dịch thuật. Tên người và tên đất không thể dịch nghĩa. Phải chuyển âm đúng như tiếng gọi. Nhưng cùng một tên đôi khi được viết khác nhau trong từng văn bản. Máy dễ đổi tên lần đầu thấy thành một từ lệch hẳn. Vì vậy người ta lập riêng danh sách tên và báo cho máy biết.

Dù vậy, hướng đi đã rõ. Trước đây một học giả đọc vài trăm tấm trong cả đời. Giờ đây máy tạo bản dịch nháp trước. Học giả dành thời gian sửa bản nháp đó. Những biên nhận 4.000 năm trước từng ngủ trong hộp đang chờ đến lượt. Tiếng nói của người bình thường, không phải của vua, bắt đầu vang lên.

Các thương nhân Kanesh không nghĩ thư của mình sẽ còn lại lâu đến vậy. Họ chỉ ghi phép tính của ngày hôm đó. Phép tính ấy nay cho thấy nền kinh tế của cả một thời đại. Thứ còn lại không phải lời khoe khoang, mà là ghi chép.

Tài liệu tham khảo

[1] Türkiye Today. (2026, 6월 13일). 4,000-year-old clay tablet sealed in mud envelope goes on display in central Türkiye. Türkiye Today. https://www.turkiyetoday.com/culture/4000-year-old-clay-tablet-sealed-in-mud-envelope-goes-on-display-in-central-turkiye-3221841 (truy cập ngày 3 tháng 9 năm 2026)

[2] Michel, C. (2020). Women of Assur and Kanesh: Texts from the Archives of Assyrian Merchants (Writings from the Ancient World 42). SBL Press.

[3] Deep Past Initiative. (2026). Deep Past Challenge: Introduction. deeppast.org. https://www.deeppast.org/challenge/intro/ (truy cập ngày 3 tháng 9 năm 2026)

[4] Kaggle. (2026). Deep Past Challenge: Translate Akkadian to English. kaggle.com. https://www.kaggle.com/competitions/deep-past-initiative-machine-translation (truy cập ngày 3 tháng 9 năm 2026)

[5] Gutherz, G., Gordin, S., Sáenz, L., Levy, O., & Berant, J. (2023). Translating Akkadian to English with neural machine translation. PNAS Nexus, 2(5), pgad096. https://doi.org/10.1093/pnasnexus/pgad096

[6] Gordin, S., Gutherz, G., Elazary, A., Romach, A., Jiménez, E., Berant, J., & Cohen, Y. (2020). Reading Akkadian cuneiform using natural language processing. PLoS ONE, 15(10), e0240511. https://doi.org/10.1371/journal.pone.0240511

[7] Deep Past Initiative. (2025, 12월 16일). Deep Past Initiative partners with Kaggle on prize challenge to translate 4,000-year-old tablets. EIN Presswire. https://www.einpresswire.com/article/875885564/deep-past-initiative-partners-with-kaggle-on-prize-challenge-to-translate-4-000-year-old-tablets (truy cập ngày 3 tháng 9 năm 2026)

[8] Bouscaren, D. (2023). Assyrian women of letters. Archaeology, 76(6). https://archaeology.org/issues/november-december-2023/features/assyrian-women-of-letters/ (truy cập ngày 3 tháng 9 năm 2026)

Chương 7. Tương lai của đế quốc từng ngủ trong yếm rùa: giáp cốt văn và ghép nối ảo 3D

7.1. Hồ sơ sự kiện

Những chữ viết 3.000 năm trước được phát hiện trong hiệu thuốc

Năm 1899, học giả Wang Yirong của nhà Thanh nhận ra một chữ cổ. Có một câu chuyện được truyền lại từ lâu về việc này. Chuyện kể rằng Wang Yirong mắc bệnh và nhận được một vị thuốc gọi là long cốt. Long cốt là thuốc bột làm bằng cách nghiền xương thú cổ. Người xưa tin những xương này là xương rồng. Người ta nói mảnh xương ông nhận được khi ấy vẫn chưa bị nghiền. Ông đã thấy những hoa văn lạ trên bề mặt của nó[1].

Câu chuyện này chưa được xác nhận là sự thật. Giới học thuật Trung Quốc cho rằng không có tư liệu nào chứng minh chuyện ấy. Nhà cổ văn tự học Li Xueqin đã bác bỏ câu chuyện này trong một bài viết năm 2007. Ông cho rằng long cốt bán làm thuốc đã bị đập vụn nên không thể nhìn thấy chữ. Trên thực tế, một người buôn đồ cổ đã mua xương khai quật ở Anyang rồi chuyển đi. Người ta cho rằng những mảnh xương đó đến tay nhà kim thạch học Wang Yirong để thẩm định[2].

Wang Yirong qua đời vào năm sau đó nên không tìm được nguồn gốc của xương. Bạn ông là Liu E đã xuất bản cuốn sách đầu tiên tập hợp các chữ này vào năm 1903. Năm 1908, học giả Luo Zhenyu đã xác định được nguồn gốc của xương[1]. Xương đang được tìm thấy trên các cánh đồng ở Anyang, tỉnh Hà Nam.

Vùng đất ấy ở Anyang chính là Ân Khư. Ân Khư là di chỉ kinh đô cuối cùng của nhà Thương. Nhà Thương cai trị nơi này từ thế kỷ 13 trước Công nguyên đến thế kỷ 11 trước Công nguyên[3]. Cuộc khai quật chính thức bắt đầu năm 1928. Trước đó, xương do nông dân đào được bị bán làm thuốc. Cũng có trường hợp người ta cạo bỏ mặt có khắc chữ rồi đem bán[1]. Những ghi chép đã tồn tại 3.000 năm suýt nữa biến mất thành thuốc.

Đất nước nơi nhà vua hỏi thần linh

Những chữ khắc trên xương ở Ân Khư được gọi là giáp cốt văn. Ở Đông Á, chưa có ghi chép bằng chữ viết nào được biết đến sớm hơn giáp cốt văn[3].

Người nhà Thương hỏi thần linh trước những việc lớn. Họ hỏi trời có mưa hay không. Họ hỏi mùa màng có tốt hay không. Họ hỏi có nên ra trận hay không. Họ cũng hỏi cả ngày vương phi sinh con. Nghi lễ hỏi ý thần linh như vậy được gọi là bói toán. Bói toán là công việc quốc gia do vương thất trực tiếp quản lý[3].

Vương phi Fu Hao của nhà Thương cũng là một chỉ huy quân đội. Tên của Fu Hao xuất hiện nhiều lần trong ghi chép giáp cốt văn. Năm 1976, mộ của Fu Hao được khai quật nguyên vẹn tại Ân Khư[4]. Đồ đồng và đồ ngọc trong mộ cho thấy địa vị của bà. Giáp cốt văn, cùng với các ngôi mộ, cho biết đời sống của người xưa như vậy.

Những câu hỏi ghi trong giáp cốt văn không chỉ dừng ở việc nước. Cũng có câu hỏi lo lắng về chứng đau răng của nhà vua. Cũng có câu hỏi về việc có nên đi săn hay không. Vì thế, những mảnh xương này trở thành ghi chép cho thấy đời sống hằng ngày của người nhà Thương[3].

Xương nung bằng lửa, chữ khắc bằng dao

Người nhà Thương dùng hai loại vật liệu để bói toán. Một là yếm phẳng của rùa. Loại kia là xương bả vai rộng của bò. Hai vật liệu này gộp lại được gọi là giáp cốt[3].

Quá trình chuẩn bị rất tỉ mỉ. Thư lại mài nhẵn bề mặt xương. Sau đó họ đục một hàng rãnh nhỏ ở mặt sau. Họ áp que nung nóng vào các rãnh đã đục. Xương gặp nhiệt thì mặt ngoài nứt ra. Nhà vua và thầy bói quan sát hướng và góc của vết nứt. Rồi họ đọc vết nứt đó như câu trả lời của thần linh[3].

Khi việc đọc kết thúc, thư lại cầm dao. Họ khắc từng nét quá trình bói toán lên bề mặt xương. Chữ được khắc ấy chính là giáp cốt văn.

Ghi chép tuân theo một trình tự cố định. Trước hết, người ta ghi ngày bói và tên người phụ trách. Tiếp theo, họ ghi câu hỏi đặt ra cho thần linh. Sau đó, họ ghi phán đoán tốt xấu của nhà vua. Cuối cùng, họ ghi việc đã thực sự xảy ra. Đây là cách điền lần lượt vào bốn phần. Tuy nhiên, những ghi chép còn đủ cả bốn phần thì hiếm. Xương vỡ nên một hoặc hai phần có thể bị mất. Cũng có trường hợp thư lại để trống ngay từ đầu[3]. Dù vậy, bản thân thứ tự ghi chép vẫn ổn định.

Sự ngay ngắn này về sau giúp ích rất nhiều. Vì khi biểu mẫu đã cố định, có thể thu hẹp phần bị bỏ trống. Nếu ô trước có ngày tháng, ô sau sẽ là câu hỏi. Máy tính cũng học thứ tự này để đoán chữ bị thiếu[3].

Cách ghi ngày tháng cũng được quy định. Nhà Thương ghép mười ký hiệu với mười hai ký hiệu để đếm ngày. Cặp ký hiệu này quay hết một vòng sau mỗi sáu mươi ngày. Dù trên xương chỉ còn hai ký hiệu ngày tháng, vẫn có thể xét theo thứ tự. Manh mối để thu hẹp thời điểm của ghi chép bị cắt nằm ở đây[3].

Ghi chép vỡ thành 150.000 mảnh

Sau khi nhà Thương sụp đổ, Ân Khư bị chôn trong đất. Trong suốt thời gian dài ấy, xương không thể còn nguyên vẹn. Sức nặng của đất đè lên những yếm mỏng và làm chúng nứt vỡ. Nước và vi sinh vật ăn mòn thớ xương. Một yếm lớn vốn là một tấm nguyên đã tách thành hàng chục mảnh.

Cho đến nay, số mảnh giáp cốt được thế giới biết đến là khoảng 150.000 mảnh[5]. Những di vật này không tập trung ở một nơi. Sau khi được phát hiện vào cuối thế kỷ 19, một phần lớn đã trôi vào thị trường đồ cổ[1]. Vì vậy, các mảnh vỡ tách ra từ cùng một mảnh xương lại nằm trong bảo tàng ở nhiều nước khác nhau[5]. Có di vật ở Anyang và Bắc Kinh của Trung Quốc. Cũng có di vật thuộc các cơ quan ở châu Âu và Hàn Quốc[6].

Việc đưa các mảnh vỡ rải rác trở lại thành một tấm ban đầu được gọi là ghép nối. Trong thời gian dài, ghép nối chỉ dựa vào mắt người. Các học giả nhìn hàng chục nghìn bản thác đen trắng qua kính lúp. Bản thác là bản sao được tạo bằng cách đặt giấy lên xương rồi dập mực. Nó giống như đặt giấy lên đồng xu rồi chà bút chì để hiện hoa văn. Các học giả so sánh thớ xương và độ cong của đường vỡ. Họ cũng xem xét từng nét bị cắt có nối tiếp nhau hay không[5].

Cách làm này chậm và vất vả. Có khi mất vài năm mới tìm được một cặp. Những mảnh chưa tìm được cặp vẫn nằm nguyên trong hộp. Không có cặp thì câu văn cũng vẫn bị cắt đứt. Các ghi chép chỉ còn câu hỏi và mất kết quả ngày càng chất đống.

Khi ước lượng quy mô công việc, khó khăn hiện rõ hơn. Nếu có 150.000 mảnh, số trường hợp chọn hai mảnh vượt quá 10 tỷ cặp. Dù một người so sánh một trăm cặp mỗi ngày, cũng không thể xem hết trong cả đời. Hơn nữa, các mảnh vỡ nằm rải rác ở nhiều nước[5]. Nhiều khi còn không có cơ hội đặt hai mảnh cạnh nhau để xem.

3.000 chữ vẫn chưa lên tiếng

Ngay cả khi ghép xong mọi mảnh vỡ, vấn đề vẫn còn. Đó là vì không đọc được chính các chữ. Trong toàn bộ giáp cốt văn, có khoảng 4.500 chữ khác nhau đã được xác nhận. Nhưng số chữ mà giới học thuật thống nhất về nghĩa chỉ khoảng 1.500 chữ. Khoảng 3.000 chữ còn lại vẫn chưa biết nghĩa[7].

Có nhiều lý do khiến việc đọc khó khăn. Vết dao đã mòn nên đường nét mờ đi. Khi xương vỡ, một nửa chữ biến mất. Mỗi thư lại có thói quen cầm dao khác nhau nên khắc cùng một chữ theo cách khác nhau.

Vì vậy, nhiều chữ có cùng nghĩa nhưng khác hình dạng đã xuất hiện. Những chữ như vậy được gọi là dị thể tự[3]. Từ điển chữ Hán hiện đại không thể giải thích những hình dạng từ 3.000 năm trước này.

Sự im lặng này cũng để lại lỗ hổng trong lịch sử nhà Thương. Nếu không đọc được tên tế lễ, không thể đếm được các loại tế lễ. Nếu không đọc được tên đất, không thể vẽ ranh giới đất nước. Nếu không đọc được tên người, không thể biết ai phụ trách việc gì. Sự im lặng của một chữ dẫn đến sự im lặng của một cảnh trong lịch sử.

Còn lại 3.000 chữ chưa biết nghĩa và hàng chục nghìn mảnh vỡ mất cặp. AI đã bước vào Ân Khư để phá vỡ sự im lặng này.

Tài liệu tham khảo

[1] Mark, E. (2016, 2월 26일). Oracle bones. World History Encyclopedia. https://www.worldhistory.org/Oracle_Bones/ (truy cập ngày 3 tháng 9 năm 2026)

[2] 姚小鸥. (2020, 7월 16일). 谁是甲骨文的最早发现者. 澎湃新闻. https://www.thepaper.cn/newsDetail_forward_8297008 (truy cập ngày 3 tháng 9 năm 2026)

[3] Chen, Z., Hua, W., Li, J., Zhu, Y., Zhi, X., Liu, Z., Chen, T., Zhang, W., & Zhai, G. (2026). Oracle bone inscriptions information processing: a comprehensive survey. npj Heritage Science, 14, 220. https://doi.org/10.1038/s40494-026-02511-w (truy cập ngày 3 tháng 9 năm 2026)

[4] Smarthistory. War and sacrifice: The tomb of Fu Hao. Smarthistory. https://smarthistory.org/tomb-of-fu-hao/ (truy cập ngày 3 tháng 9 năm 2026)

[5] Zhang, C., Wang, B., Chen, K., Zong, R., Mo, B., Men, Y., Almpanidis, G., Chen, S., & Zhang, X. (2022). Data-driven oracle bone rejoining: A dataset and practical self-supervised learning scheme. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 4482-4492). ACM. https://doi.org/10.1145/3534678.3539050

[6] 梁文艳. (2026, 5월 26일). 以科技赋能，让甲骨文"活"起来. 中国青年报. https://zqb.cyol.com/pc/content/202605/26/content_426168.html (truy cập ngày 3 tháng 9 năm 2026)

[7] Liu, Y., Guan, H., Wang, P., Wang, X., Wan, J., Zhang, K., Zheng, H., Liu, X., Kuang, Z., Yang, H., Li, B., Liu, Y., Jin, L., & Bai, X. (2026). AlphaOracle: Oracle bone script decipherment via human-workflow-inspired deep learning. The Innovation, 7(11), 101462. https://doi.org/10.1016/j.xinn.2026.101462

7.2. Truy vết của đội điều tra khoa học: mô hình đa phương thức chuyên cho giáp cốt và ghép nối ảo 3D

Gom tư liệu rải rác vào một màn hình

Cuộc điều tra bắt đầu từ việc thu thập tư liệu. Nghiên cứu giáp cốt cũng vậy. Năm 2019, nhóm nghiên cứu của Đại học Sư phạm Anyang và Viện Khoa học Xã hội Trung Quốc mở một nền tảng tư liệu tên là Yin Qi Wen Yuan. Yin Qi Wen Yuan tập hợp bản thác giáp cốt, thông tin chữ viết và luận văn vào một nơi rồi công khai miễn phí[1].

Nền tảng này không chỉ đăng ảnh. Nó chụp các mảnh giáp cốt ở độ phân giải cao và tạo thành mô hình ba chiều. Mô hình ba chiều là dữ liệu tái hiện vật thể trong máy tính dưới dạng lập thể bằng cách quét từ nhiều hướng. Khi xoay màn hình, có thể nhìn mặt trước, mặt sau và cạnh vỡ của xương. Nhóm nghiên cứu tạo các mô hình này với độ chính xác đến từng milimét. Độ sâu của những vết dao khắc nông cũng được đưa vào dữ liệu này[1].

Công việc đưa giáp cốt đã ra nước ngoài trở lại cũng nối với nền tảng này. Nhóm nghiên cứu chọn cách đưa dữ liệu ba chiều về thay cho hiện vật thật. Họ hợp tác với các nơi lưu giữ ở Đức, Hàn Quốc, Pháp và các nước khác để thu thập hơn 900 mẫu dữ liệu[1]. Hiện vật thật vẫn ở nơi lưu giữ ban đầu, chỉ có hình dạng của chúng trở về Anyang.

Tháng 10 năm 2025, một công cụ tên là Yin Qi Xing Zhi được gắn thêm lên nền tảng này. Công cụ này nhận dạng chữ, so sánh bản thác và tìm tài liệu liên quan[1]. Nhà kho từng chỉ chất tư liệu đã trở thành trợ lý trả lời câu hỏi.

Có thể kiểm tra lượng tư liệu đã tích lũy bằng các con số. Một bài tổng quan công bố năm 2026 giới thiệu 36 bộ dữ liệu công khai liên quan đến giáp cốt. Có bộ dữ liệu dùng cho nhận dạng chữ. Cũng có bộ dữ liệu dùng cho nghiên cứu ghép nối[2].

Cách dùng nhiều loại dữ liệu cùng nhau được gọi là đa phương thức. Trong nghiên cứu giáp cốt, từ này có nghĩa là xem đồng thời ba loại dữ liệu. Một là bản thác đen trắng được dập bằng mực. Một là ảnh màu còn giữ màu sắc và chất liệu. Loại còn lại là dữ liệu quét ba chiều chứa thông tin độ sâu. Bản thác làm nét chữ rõ nhưng mất độ sâu. Dữ liệu ba chiều chứa độ sâu nhưng nét chữ trông mờ. Khi chồng ba loại lên nhau, chúng có thể bù vào chỗ thiếu của nhau[2].

Bắt chước trình tự con người đọc chữ

AlphaOracle công bố năm 2026 là AI vận hành trên các dữ liệu này. Mô hình này do Đại học Khoa học và Công nghệ Hoa Trung, Đại học Sư phạm Anyang và các đơn vị khác cùng xây dựng. Nguyên tắc thiết kế là đi theo đúng trình tự làm việc của học giả con người[3].

Học giả con người trải qua bốn bước. Trước hết, họ xem bản thác và chỉnh lý các nét. Tiếp theo, họ phân tích hình dạng của chữ. Sau đó, họ tìm các câu khác có cùng chữ xuất hiện. Cuối cùng, họ dùng tri thức văn hiến cổ để xét nghĩa có phù hợp hay không. AlphaOracle được chia thành các bộ phận phụ trách từng bước trong bốn bước này[3].

Cấu trúc này có lý do. Một cỗ máy chỉ đưa ra đáp án đột ngột thì học giả khó tin. Khi chia thành các bước, có thể thấy đáp án rẽ hướng ở bước nào. Học giả có thể quay lại bước đó và đối chiếu ý kiến với máy[3].

Lượng tư liệu mà mô hình này đọc vượt quá lượng một người có thể xem trong cả đời. Nó xử lý 73.883 bản thác. Nó cũng xử lý 45.364 mẫu dữ liệu quét và 23.755 luận văn nghiên cứu. Nhóm nghiên cứu cho 86 chuyên gia dùng thử công cụ này. Kết quả là thời gian phân tích một chữ giảm 64 phần trăm. 79 phần trăm người tham gia trả lời rằng công cụ này hữu ích[3].

Nhìn chữ bằng cách tách thành bộ phận

Một nhóm nghiên cứu khác chọn cách không học thuộc cả chữ. Chữ giáp cốt văn được tạo từ các bộ phận hình vẽ lặp lại[4]. Những bộ phận như hình bàn tay, hình bàn chân, hình đồ đựng xuất hiện trong nhiều chữ. Nếu biết nghĩa của bộ phận, có thể đoán cả chữ chưa từng thấy.

Từ ý tưởng này, bộ dữ liệu OB-Radix được công bố vào tháng 4 năm 2026. OB-Radix chứa 934 chữ khác nhau. Số hình bộ phận tạo nên các chữ ấy là 1.853. Có 478 loại bộ phận. Mỗi bộ phận có phần giải thích đã được chuyên gia xác nhận[4].

Khi gặp chữ lạ, AI trước hết tách chữ thành các bộ phận. Sau đó, nó xét các bộ phận ấy nằm trong quan hệ nào. Hai bộ phận đặt trên dưới hay đặt cạnh nhau có thể làm nghĩa khác đi. Mô hình xem xét cả quan hệ này để đề xuất nghĩa[4].

Tháng 6 cùng năm, một mô hình tên OracleAnalyzer cũng ra đời. Mô hình này được huấn luyện lại trên dữ liệu giáp cốt văn từ một mô hình ngôn ngữ nhỏ quy mô 3 tỷ tham số. Nhóm nghiên cứu mới tạo và bổ sung dữ liệu suy luận giáp cốt văn và dữ liệu ưu tiên. Dữ liệu ưu tiên là dữ liệu tập hợp các câu trả lời mà con người chọn là tốt hơn. Kết quả là mô hình nhỏ đạt thành tích tốt hơn nhiều so với mô hình lớn hơn rất nhiều[5]. Điều đó có nghĩa là nếu chọn dữ liệu tốt, không cần chỉ dựa vào kích thước để thúc đẩy kết quả.

Mô hình vẽ lại các nét đã mờ

Cũng có nghiên cứu làm cho chữ hiện rõ trước khi đọc. OBSD công bố năm 2024 là một công cụ như vậy. Công cụ này dùng mô hình khuếch tán để vẽ lại hình ảnh[6]. Mô hình khuếch tán là cách tạo dần một hình ảnh rõ nét từ vết mờ.

OBSD nhận đầu vào là hình chữ giáp cốt đã mòn và mờ. Sau đó, nó đề xuất bằng hình ảnh xem chữ đó nối tiếp thành chữ Hán hiện đại nào. Nghiên cứu này được chọn là bài báo xuất sắc nhất tại một hội nghị quốc tế năm 2024[6]. Ý tưởng biến việc đọc chữ thành bài toán vẽ hình đã được công nhận.

Máy tính ghép mặt cắt của các mảnh vỡ

Song song với việc đọc chữ, nghiên cứu ghép mảnh vỡ cũng được tiến hành. OB-Rejoin công bố năm 2022 là bộ dữ liệu dùng cho nghiên cứu ghép nối. Nó chứa 998 bản thác giáp cốt thật. Cùng nhóm nghiên cứu cũng công bố mạng nơ-ron ghép nối tên S3-Net[7].

Mạng nơ-ron này tính toán độ giống nhau giữa các đường cong ở mép mảnh vỡ. Khi hai mép khớp với nhau, điểm số tăng cao. Máy sắp xếp các ứng viên theo thứ tự điểm cao. Cùng nhóm nghiên cứu cũng công bố một nghiên cứu trước đó vào năm 2020. Khi ấy, tỷ lệ đưa đáp án đúng vào 10 ứng viên đầu là 98.39 phần trăm[11].

Khi ghép tranh xếp hình, con người nhìn hình dạng mảnh vỡ trước khi nhìn bức tranh. Máy cũng dùng cách đó. Nhưng máy có thể so sánh hàng trăm nghìn lần trong một ngày.

Thứ máy đưa ra không phải đáp án đúng, mà là danh sách ứng viên. Có khi mục đứng đầu danh sách cũng sai. Vì vậy, nhóm nghiên cứu lấy việc đáp án đúng có nằm trong 10 mục đầu hay không làm tiêu chuẩn đánh giá hiệu năng[7]. Tình huống con người chỉ cần kiểm tra 10 mục khác với tình huống phải kiểm tra 100.000 mục.

Trung tâm Nghiên cứu giáp cốt văn của Đại học Sư phạm Thủ đô đã cùng Đại học Hà Nam làm ra một chương trình ghép nối. Tên của chương trình là Tiedada, và được dùng từ năm 2020. Chương trình này so sánh cùng lúc 180.000 bản dập để tìm cặp khớp. Trung tâm nghiên cứu đã dùng cách này để sắp xếp hồ sơ của khoảng 1.500 mảnh giáp cốt. Trước đây, việc tìm mảnh khớp với một mảnh có khi mất 8 đến 10 năm. Nay chỉ cần vài ngày là có ứng viên[8].

Bài kiểm tra chuyển sang đơn vị câu

Đọc một chữ và hiểu một câu là hai việc khác nhau. S-OBI, được công bố vào cuối tháng 6 năm 2026, là một bộ đề kiểm tra đo sự khác nhau này. Nhóm nghiên cứu lấy câu hỏi từ 95 bản dập gốc đã được chuyên gia xác nhận. Tổng số câu hỏi là 695, và người trả lời phải đáp ở cấp độ câu. Sau đó, họ cho nhiều AI lớn giải cùng bộ đề. Kết quả kiểm tra cho thấy điểm số giữa các mô hình khác nhau rõ rệt. Việc nắm dòng chảy của cả câu vẫn còn khó[9].

Bài kiểm tra này còn cho biết một sự thật nữa. Đó là AI thông thường không giải tốt bài toán giáp cốt văn[9]. Ngay cả mô hình đã học rộng nhiều loại văn bản trên đời cũng dao động trước câu văn 3.000 năm trước. Vì đó là chữ và ngữ pháp mà nó chưa từng học. Vì vậy, các nghiên cứu huấn luyện lại riêng cho giáp cốt vẫn tiếp tục[5].

Cũng đã có nghiên cứu nối ngữ cảnh vượt qua mặt cắt vỡ. Một nhóm nghiên cứu tạo ra nhiệm vụ phán đoán hai câu có vốn nằm trên cùng một mảnh xương hay không. Họ công bố bộ dữ liệu có tên OBID-ACR. Họ cũng đề xuất cách dùng chung thông tin hình dạng chữ và thông tin câu[10]. Nhờ vậy, đã mở ra con đường nối thử ngữ cảnh ngay cả trước khi thật sự ghép các mảnh lại.

Tài liệu tham khảo

[1] 梁文艳. (2026, 5월 26일). 以科技赋能，让甲骨文"活"起来. 中国青年报. https://zqb.cyol.com/pc/content/202605/26/content_426168.html (truy cập ngày 3 tháng 9 năm 2026)

[2] Chen, Z., Hua, W., Li, J., Zhu, Y., Zhi, X., Liu, Z., Chen, T., Zhang, W., & Zhai, G. (2026). Oracle bone inscriptions information processing: a comprehensive survey. npj Heritage Science, 14, 220. https://doi.org/10.1038/s40494-026-02511-w (truy cập ngày 3 tháng 9 năm 2026)

[3] Liu, Y., Guan, H., Wang, P., Wang, X., Wan, J., Zhang, K., Zheng, H., Liu, X., Kuang, Z., Yang, H., Li, B., Liu, Y., Jin, L., & Bai, X. (2026). AlphaOracle: Oracle bone script decipherment via human-workflow-inspired deep learning. The Innovation, 7(11), 101462. https://doi.org/10.1016/j.xinn.2026.101462

[4] Zhang, J., Li, R., Pang, H., Xia, D., Zhu, Z., Zhang, Q., Li, C., & Yang, X. (2026, 4월 8일). Specializing large models for oracle bone script interpretation via component-grounded multimodal knowledge augmentation. arXiv:2604.06711. https://arxiv.org/abs/2604.06711 (truy cập ngày 3 tháng 9 năm 2026)

[5] Song, Z., Wang, Y., Ma, Z., Yu, Z., Wang, T., Zhang, J., Wang, T., & Yu, K. (2026, 6월 24일). OracleAnalyser: Analysing implicit semantics of oracle bone scripts through MLLMs with post-training. arXiv:2606.25906. https://arxiv.org/abs/2606.25906 (truy cập ngày 3 tháng 9 năm 2026)

[6] Guan, H., Yang, H., Wang, X., Han, S., Liu, Y., Jin, L., Bai, X., & Liu, Y. (2024). Deciphering oracle bone language with diffusion models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.831/ (truy cập ngày 3 tháng 9 năm 2026)

[7] Zhang, C., Wang, B., Chen, K., Zong, R., Mo, B., Men, Y., Almpanidis, G., Chen, S., & Zhang, X. (2022). Data-driven oracle bone rejoining: A dataset and practical self-supervised learning scheme. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 4482-4492). ACM. https://doi.org/10.1145/3534678.3539050

[8] 北京日报. (2026, 8월 28일). 高校团队"唤醒"三千年前文明碎片. 新浪教育. https://edu.sina.com.cn/l/2026-08-28/doc-inipwhrw0020843.shtml (truy cập ngày 3 tháng 9 năm 2026)

[9] Li, Z., Chen, Z., Chen, T., & Zhai, G. (2026, 6월 30일). Beyond single character: Evaluating MLLMs for sentence-level oracle bone inscription understanding. arXiv:2606.31169. https://arxiv.org/abs/2606.31169 (truy cập ngày 3 tháng 9 năm 2026)

[10] Zhang, H., Wang, T., Zhang, Z., Li, B., Sun, H., Hou, C., Wang, N., Yu, Y., Jiao, Q., Xiong, J., & Liu, Y. (2026). A multi-modal dataset and method for bone-level association prediction in oracle bone inscriptions. npj Heritage Science, 14, 19. https://doi.org/10.1038/s40494-025-02282-w (truy cập ngày 3 tháng 9 năm 2026)

[11] Zhang, C., Zong, R., Cao, S., Men, Y., & Mo, B. (2020). AI-powered oracle bone inscriptions recognition and fragments rejoining. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20), 5309-5311. https://doi.org/10.24963/ijcai.2020/779

7.3. Sự thật được làm sáng tỏ

Khi một chữ thay đổi, bản đồ cũng thay đổi

Thành quả đầu tiên mà AlphaOracle đưa ra là một chữ. Nhóm nghiên cứu đưa vào hệ thống một chữ đã gây tranh luận lâu nay. Đó là chữ về sau nối tiếp thành chữ Hán lao (勞)[1]. Hệ thống sắp xếp hình dạng nét và cấu tạo bộ phận của chữ. Sau đó, nó tìm và liệt kê các câu khác có chữ đó. Tiếp theo, nó đối chiếu với tri thức văn hiến cổ để thu hẹp nghĩa.

Đáp án hệ thống đề xuất là chữ này là tên đất hoặc tên thị tộc. Thị tộc là nhóm gia đình thân tộc tách ra từ cùng một tổ tiên. Kết quả này khác với cách giải thích từng đọc chữ đó như một từ bổ nghĩa cho câu. Nhóm nghiên cứu cho biết cách giải thích này đã qua sự xem xét của các học giả cổ văn tự học[1].

Khi nghĩa của một chữ thay đổi, mọi câu có chữ đó cũng thay đổi. Nếu đó là tên người, câu ấy là chuyện về một cá nhân. Nếu đó là tên đất, cùng câu ấy trở thành ghi chép về hành chính địa phương. Cũng phải vẽ lại cách nhà Thương cai quản vùng đất đó. Nhóm nghiên cứu báo cáo rằng kết quả này khiến ta phải điều chỉnh cách nhìn về hành chính và xã hội nhà Thương[1].

Điều đáng chú ý trong nghiên cứu này không phải bản thân đáp án. Đó là hệ thống đưa ra cả căn cứ. AlphaOracle cho thấy nó đã nhìn vào câu nào và bộ phận nào để chọn đáp án đó[1]. Học giả có thể lần theo căn cứ ấy rồi quyết định đồng ý hay không. Trọng tâm của nghiên cứu này là đưa ra đáp án ở dạng con người có thể kiểm tra.

Việc ghép cặp từng mất nhiều năm đã rút xuống còn vài ngày

Phía ghép nối cũng có kết quả. Đó là thành quả của chương trình Tiedada đã nêu ở trên. Chương trình này tìm ra hơn 300 cặp mà các học giả đã bỏ lỡ trong thời gian dài[2]. Đó là con số được cộng thêm vào các cặp đã được tìm bằng tay hơn 100 năm qua.

Phần việc của con người vẫn còn nguyên. Học giả kiểm tra từng mục trong danh sách do máy đưa ra. Việc kiểm tra kết thúc bằng đối chiếu với hiện vật thật.

Phòng nghiên cứu của Đại học Sư phạm An Dương cũng công bố thành quả riêng. Đến tháng 5 năm 2026, số giáp cốt mới được ghép lại là 155 trường hợp. Số trường hợp cùng phòng nghiên cứu nhận ra ghi chép trùng lặp lên gần 7.000[3]. Từng có chuyện một mảnh xương được đăng riêng trong nhiều sách mục lục. Khi đó, một mảnh xương bị tính như hai di vật khác nhau. Số trường hợp phát hiện trùng lặp lớn hơn nhiều so với số mảnh mới được ghép. Điều đó có nghĩa là việc máy làm không chỉ là phát hiện mới, mà còn là kiểm tra lại hồ sơ cũ.

Khi các mảnh được ghép lại, câu văn dài hơn. Thường thì một bên chỉ còn câu hỏi, còn kết quả nằm ở bên kia. Hai mảnh phải gặp nhau thì ngày bói và việc thật sự xảy ra mới nối thành một dòng. Khi đó, nó không còn là nhiều ghi chép ngắn, mà trở thành ghi chép về một sự kiện.

Sự khác biệt này tác động ngay đến việc viết lịch sử. Nếu chỉ nhìn ghi chép bị cắt, sự kiện trông rời rạc. Nếu nhìn ghi chép đã nối, thứ tự trước sau hiện ra. Ta nắm được nhà vua hỏi khi nào và nhận kết quả khi nào.

7.000 ghi chép được nhận ra là trùng lặp cũng có trọng lượng như vậy[3]. Nếu tính cùng một mảnh xương thành hai, số lượng tư liệu đã sai ngay từ đầu. Có khi cùng một câu bị hiểu và trích dẫn như hai ghi chép khác nhau. Khi máy so sánh lại hàng loạt bản dập, nó lọc ra những trùng lặp này. Đây không phải phát hiện nổi bật, nhưng là việc dựng lại nền tư liệu cho đúng.

Những mảnh xương rải rác bên kia biển gặp nhau trên màn hình

Bức tường cản trở việc ghép nối là khoảng cách. Nếu các mảnh vỡ rơi ra từ cùng một mảnh xương lại nằm ở các nước khác nhau, không thể đặt chúng cạnh nhau để xem. Dù trao đổi ảnh, cũng không biết được độ dày và độ cong. Dữ liệu 3D đã hạ thấp bức tường này.

Hơn 900 tư liệu ở nước ngoài đã nêu ở trên là trường hợp vượt qua bức tường ấy[3]. Hiện vật vẫn ở nguyên nơi vốn có. Thứ được chuyển đi là thông tin hình dạng của bề mặt.

Nhà nghiên cứu ở An Dương có thể xoay mảnh vỡ đó trên màn hình để xem. Họ cũng có thể đặt mảnh ở Bắc Kinh và mảnh ở châu Âu trên cùng một màn hình rồi thử khớp mặt cắt.

Cách công bố tư liệu cũng thay đổi. Trước đây, phải đến cơ quan lưu giữ mới có thể xem di vật. Chỉ riêng việc xin phép có khi mất nhiều tháng. Nay sinh viên cũng có thể mở tư liệu giáp cốt ở nhà[3]. Khi số người xử lý tư liệu tăng, số con mắt kiểm tra cũng tăng. Khi số con mắt tăng, cơ hội lọc ra các cặp ghép sai cũng tăng.

Những việc máy vẫn chưa làm được

Cùng với thành quả, giới hạn cũng trở nên rõ ràng. Bài kiểm tra S-OBI công bố vào cuối tháng 6 năm 2026 đã cho thấy giới hạn ấy bằng con số. Nhiều AI lớn giải cùng các câu hỏi về câu giáp cốt. Điểm số giảm ở bài hiểu cả câu so với bài nhận ra từng chữ. Nguyên nhân cũng lộ ra. Mô hình nhìn sai ngay cả những chữ còn nguyên vẹn. Nó mang lỗi đó đi suy luận tiếp, nên đáp án bị lệch[4]. Việc nắm dòng chảy của câu văn 3.000 năm trước vẫn còn phụ thuộc nhiều vào con người.

Vấn đề cần thận trọng hơn là bịa ra. AI chọn đáp án theo xác suất. Trong lĩnh vực có ít dữ liệu học như giáp cốt văn, nó có thể tạo ra chữ trông có vẻ hợp lý nhưng thật ra chưa từng tồn tại. Ngôn ngữ nhà Thương cũng không có nhiều tư liệu ngôn ngữ láng giềng để so sánh[5]. Đây là tình huống khó tra từ điển để kiểm lại đáp án.

Sự lệch của tư liệu vẫn còn. Tư liệu giáp cốt được tập hợp đến nay nghiêng về một số thời kỳ và một số địa điểm khai quật. Máy học từ tư liệu lệch dễ mắc lỗi khi gặp giáp cốt ở vùng xa lạ. Một bài tổng thuật ra năm 2026 cũng nêu chất lượng và sự cân bằng của tư liệu là nhiệm vụ phía trước[5].

Nhìn bằng con số, đường đi vẫn còn dài. Số chữ chưa biết nghĩa vẫn còn khoảng 3.000 chữ. Chữ mà AlphaOracle giải được với đầy đủ căn cứ là một trong số đó[1]. Nghĩ đến công sức cần để giải một chữ, phần việc còn lại rất lớn. Nhưng khi phương pháp được sắp xếp rõ, chữ tiếp theo sẽ nhanh hơn một chút.

Phần việc con người và máy chia nhau đảm nhận

Vì vậy, cách nghiên cứu hiện nay đã định hình theo hướng đặt con người và máy cạnh nhau. Máy đưa ra rộng các ứng viên và sắp xếp căn cứ. Con người kiểm tra các ứng viên ấy bằng văn hiến và hiện vật.

Việc rút ngắn thời gian phân tích đã nêu ở trên cũng là kết quả của sự phân công này. Máy không thay con người phán đoán, mà lọc trước tư liệu để con người xem. Đây cũng là lý do các chuyên gia trả lời rằng công cụ này hữu ích[1].

Trong ghép nối, cùng một cách làm cũng đã định hình. Việc đối chiếu trước đây mất nhiều năm nay được chương trình hoàn tất trong vài ngày[2]. Học giả cầm các ứng viên ấy đi vào kho lưu trữ. Họ đặt xương thật sát nhau để kiểm tra mặt cắt có khớp hay không. Phán đoán cuối cùng diễn ra ở đầu ngón tay con người.

Sự phân công này đi kèm một điều kiện. Đó là con người phải thấy được căn cứ của đáp án do máy đưa ra. Nếu không thấy căn cứ, học giả không có cách kiểm tra. Đáp án không thể kiểm tra thì không thể đưa vào viết lịch sử. Đây cũng là lý do AlphaOracle chia quy trình thành các bước[1].

Vị vua nhà Thương 3.000 năm trước nhìn vết nứt trên xương để hỏi về tương lai. Nhà nghiên cứu hôm nay nhìn mặt cắt nứt vỡ của cùng mảnh xương để hỏi về quá khứ. Hướng của câu hỏi trái ngược nhau, nhưng dấu vết được nhìn vào vẫn là một.

Tài liệu tham khảo

[1] Liu, Y., Guan, H., Wang, P., Wang, X., Wan, J., Zhang, K., Zheng, H., Liu, X., Kuang, Z., Yang, H., Li, B., Liu, Y., Jin, L., & Bai, X. (2026). AlphaOracle: Oracle bone script decipherment via human-workflow-inspired deep learning. The Innovation, 7(11), 101462. https://doi.org/10.1016/j.xinn.2026.101462

[2] 北京日报. (2026, 8월 28일). 高校团队"唤醒"三千年前文明碎片. 新浪教育. https://edu.sina.com.cn/l/2026-08-28/doc-inipwhrw0020843.shtml (truy cập ngày 3 tháng 9 năm 2026)

[3] 梁文艳. (2026, 5월 26일). 以科技赋能，让甲骨文"活"起来. 中国青年报. https://zqb.cyol.com/pc/content/202605/26/content_426168.html (truy cập ngày 3 tháng 9 năm 2026)

[4] Li, Z., Chen, Z., Chen, T., & Zhai, G. (2026, 7월 1일). Beyond single character: Evaluating MLLMs for sentence-level oracle bone inscription understanding. arXiv:2606.31169. https://arxiv.org/abs/2606.31169 (truy cập ngày 3 tháng 9 năm 2026)

[5] Chen, Z., Hua, W., Li, J., Zhu, Y., Zhi, X., Liu, Z., Chen, T., Zhang, W., & Zhai, G. (2026). Oracle bone inscriptions information processing: a comprehensive survey. npj Heritage Science, 14, 220. https://doi.org/10.1038/s40494-026-02511-w (truy cập ngày 3 tháng 9 năm 2026)

Chương 8. Nét bút lông sống sót trở về từ hố bùn: thẻ gỗ Silla và AI đọc chữ viết thảo

8.1. Hồ sơ vụ việc

Thời đại viết chữ trên gỗ

Đã có một thời giấy rất quý. Người Đông Á cổ đại bào mỏng gỗ rồi viết chữ lên đó. Tấm gỗ có ghi chữ này được gọi là thẻ gỗ. Thẻ gỗ thường rộng bằng hai hoặc ba ngón tay đặt sát nhau. Nhiều thẻ dài hơn một gang tay một chút. Cũng có loại gỗ được đẽo như một cây trụ. Loại thẻ gỗ này có thể viết trên cả bốn mặt. Vì là tấm gỗ có chữ trên nhiều mặt, nó được gọi là thẻ gỗ nhiều mặt.

Thẻ gỗ có nhiều công dụng. Một loại là nhãn treo trên hàng hóa. Khi gửi thóc hoặc mắm từ địa phương về kinh đô, người ta buộc thẻ gỗ bằng dây rồi treo vào hàng. Trên đó ghi tên địa phương gửi, tên hàng và số lượng. Một loại khác là văn thư của quan phủ. Người ta tìm thấy thẻ gỗ ghi tên và cấp bậc của quan lại, đồ vật đã nhận và ngày tháng[1]. Cũng có thẻ gỗ dùng để luyện chữ, trên đó cùng một chữ Hán được viết nhiều lần.

Gỗ cứng hơn giấy nên có thể viết lại. Khi thẻ gỗ hết công dụng, người ta dùng dao bào mỏng lớp mặt ngoài. Khi mặt mới lộ ra, họ lại viết chữ lên đó. Những mảnh gỗ mỏng bị bào ra khi ấy cũng được tìm thấy cùng trong di tích. Những mảnh vụn này được gọi là saksseol. Từ này có nghĩa là mảnh gỗ bị bào ra để xóa chữ. Có thể thấy điều đó qua 329 hiện vật thuộc nhóm thẻ gỗ tìm thấy tại di tích Gwanbuk-ri ở Buyeo. Trong số đó, 247 hiện vật là saksseol. Điều này có nghĩa là gần đó từng có cơ sở tạo và sửa văn thư[1].

Các nhà sử học coi trọng thẻ gỗ vì có lý do. Sách sử thường được viết sau khi sự kiện đã kết thúc. Trong đó có phán đoán của người viết và hoàn cảnh của vương triều. Thẻ gỗ gần với ghi chép về công việc trong chính ngày đó. Đó không phải là bài viết được trau chuốt để cho đời sau xem. Ai gửi mấy seom thóc vào lúc nào được giữ lại gần như nguyên trạng. Loại ghi chép được tạo ra vào thời điểm sự kiện xảy ra như vậy được gọi là tư liệu bậc một. Trên thẻ gỗ cũng có chữ viết sai và chỗ sửa. Vì vậy, nhà nghiên cứu đặt nhiều thẻ cạnh nhau để đối chiếu.

Thẻ gỗ được tìm thấy cả trên bán đảo Triều Tiên và quần đảo Nhật Bản. Cả hai nơi đều là xã hội làm văn thư bằng chữ Hán. Đồng thời, họ cũng tìm cách ghi lại tiếng nói của nước mình. Người Silla trộn nghĩa và âm của chữ Hán để ghi trật tự từ của tiếng Hàn. Cách ghi này được gọi là Idu và Hyangchal. Người Nhật cũng mượn âm của chữ Hán để ghi tiếng của mình. Thẻ gỗ là bằng chứng vật chất cho biết những thử nghiệm này bắt đầu từ khi nào. Lý do là văn bản giấy hầu như không còn lại. Đây là lý do các nghiên cứu đặt thẻ gỗ của hai nước cạnh nhau để so sánh vẫn tiếp tục.

Một nghìn năm được bùn gìn giữ

Nơi tìm thấy thẻ gỗ thường khá cố định. Đó là đáy ao bên cạnh cung điện. Đó là vùng đất thấp ứ nước bên trong thành. Đó là lớp bùn lấp mương. Gỗ mà con người dùng rồi bỏ đi chìm xuống nước và bị chôn nguyên như vậy.

Trong bùn dưới nước, không khí hầu như không lọt vào được. Thứ thường làm gỗ mục là vi sinh vật sống nhờ không khí. Khi không khí bị chặn, các vi sinh vật này không hoạt động được. Vẫn còn những vi sinh vật sống không cần không khí. Tuy vậy, tốc độ gỗ bị phân hủy chậm đi rất nhiều. Vì thế, gỗ giữ được hình dạng suốt hơn một nghìn năm. Ngay cả ở vùng đất nơi da và giấy biến mất không dấu vết, thẻ gỗ vẫn còn lại. Không phải mộ, mà chính hố rác đã gìn giữ ghi chép.

Thay vào đó, một biến đổi khác đã xảy ra. Nước thấm vào trong gỗ và lấp đầy thành tế bào. Sắt và nhiều khoáng chất trong đất cũng đi vào cùng nước. Gỗ trở thành vật chứa đầy nước. Loại gỗ này được gọi là gỗ ngậm nước, tức gỗ bão hòa nước.

Không được lấy gỗ bão hòa nước ra rồi để khô tự nhiên. Khi nước thoát ra, gỗ co rút nhanh và nứt. Đây là lý do tại hiện trường khai quật, thẻ gỗ được cho ngay vào thùng nước. Trong phòng bảo tồn, nước trong gỗ được thay dần bằng hóa chất. Chỉ riêng quá trình này đã mất từ vài tháng đến vài năm. Việc đọc chữ phải hoàn tất trước đó. Lý do là màu bề mặt có thể thay đổi trong quá trình xử lý.

Chữ đen trên gỗ đen

Gỗ bị khoáng chất thấm vào sẽ sẫm màu. Nhiều mảnh chuyển từ nâu sang đen kịt. Hiện tượng gỗ chuyển sang màu đen này được gọi là đen hóa. Vấn đề nằm ở vật liệu viết chữ. Người xưa đốt gỗ thông lấy muội, rồi trộn với keo da để làm mực tàu. Thành phần chính của mực tàu là carbon và màu của nó là đen.

Có chữ được viết bằng mực đen trên gỗ đã chuyển đen. Dù chiếu đèn sáng đến đâu, hai thứ vẫn không phân biệt được. Hơn nữa, trong thời gian dài nằm dưới nước, nhiều hạt mực đã bị rửa trôi. Những vệt dọc sẫm của thớ gỗ lẫn vào dấu bút. Đất và chất mục lấp vào rãnh nơi bút đi qua. Bằng mắt người, thậm chí khó biết có chữ hay không.

Khi việc đọc khó, các giả thuyết cũng chia rẽ. Cùng một thẻ gỗ nhưng mỗi học giả có thể đọc thành chữ Hán khác nhau. Khi một chữ thay đổi, nghĩa của câu thay đổi và cách giải thích lịch sử cũng thay đổi. Vì vậy, các nhà nghiên cứu thẻ gỗ từ lâu đã tìm một đôi mắt mới. Họ cần công cụ bắt được những gì mắt người bỏ sót.

Số thẻ gỗ cần đọc vẫn tiếp tục tăng. Tại thành Seongsan ở Haman, tỉnh Gyeongnam, thẻ gỗ Silla liên tục được tìm thấy. Vào tháng 8 năm 2025, hai thẻ gỗ có minh văn mới cũng được xác nhận[2]. Minh văn có nghĩa là chữ được ghi trên di vật. Tại di tích Gwanbuk-ri ở Buyeo, tỉnh Chungnam, nhiều hiện vật thẻ gỗ thời Sabi của Baekje được tìm thấy thành từng nhóm lớn[1]. Tại thành Daemosan ở Yangju, tỉnh Gyeonggi, một thẻ gỗ được cho là thuộc thế kỷ 5 đã được tìm thấy. Nếu đúng, đó sẽ là một trong những tư liệu chữ viết Baekje sớm nhất từng được biết đến[3].

Tình hình ở Nhật Bản cũng như vậy. Viện Nghiên cứu Quốc gia Nara về Di sản Văn hóa tập hợp và công bố thông tin về thẻ gỗ tìm thấy ở nhiều nơi. Cơ sở dữ liệu này có tên là Mokkan-ko[4]. Ở đây, người dùng không chỉ tìm bằng chữ ghi trên thẻ gỗ. Họ cũng có thể tìm theo công dụng, niên đại, chất liệu và nơi khai quật[5]. Số gỗ cần đọc đã chất lên như vậy.

Tài liệu tham khảo

[1] 파이낸셜뉴스. (2026, 2월 5일). 부여 관북리 '백제 피리' 첫 확인..목간 329점 무더기 발견. 파이낸셜뉴스. https://www.fnnews.com/news/202602051448564101 (truy cập ngày 3 tháng 9 năm 2026)

[2] 경남일보. (2025, 8월 7일). 함안 성산산성서 새로운 목간 2점 명문 확인. 경남일보. https://www.gnnews.co.kr/news/articleView.html?idxno=616028 (truy cập ngày 3 tháng 9 năm 2026)

[3] 뉴시스. (2025, 11월 20일). 1500년 전 '백제 문자' 깨어났다…양주 대모산성서 5세기 목간 출토. 뉴시스. https://www.newsis.com/view/NISX20251120_0003410107 (truy cập ngày 3 tháng 9 năm 2026)

[4] 奈良文化財研究所. 木簡庫(MOKKAN-KO). https://mokkanko.nabunken.go.jp/ja/ (truy cập ngày 3 tháng 9 năm 2026)

[5] カレントアウェアネス・ポータル. (2018). 奈良文化財研究所、木簡に関する情報を検索するシステム「木簡庫」(MOKKAN-KO)を公開. 国立国会図書館. https://current.ndl.go.jp/car/35714 (truy cập ngày 3 tháng 9 năm 2026)

8.2. Cuộc truy tìm của đội điều tra khoa học 1: kính lúp mực số

Dùng ánh sáng mắt không nhìn thấy

Mắt người chỉ nhìn thấy một dải ánh sáng hẹp. Bên ngoài màu đỏ của cầu vồng cũng có ánh sáng. Ánh sáng này được gọi là tia hồng ngoại. Vì có bước sóng dài, mắt không bắt được nó, nhưng máy ảnh có thể chụp được. Tín hiệu điều khiển từ xa gửi đến tivi cũng là tia hồng ngoại.

Lý do dùng tia hồng ngoại để đọc thẻ gỗ nằm ở tính chất của vật liệu. Hạt carbon trong mực hấp thụ tia hồng ngoại tốt. Thớ gỗ bị đen hóa phản xạ lại nhiều tia hồng ngoại. Hai chất cùng đen như nhau trước mắt người lại tách ra trước máy ảnh hồng ngoại. Phần hấp thụ được chụp thành màu tối, phần phản xạ được chụp thành màu sáng. Dấu bút tưởng như đã biến mất hiện lên trên ảnh.

Ánh sáng thường dùng là cận hồng ngoại có bước sóng khoảng 800 nanomet. Nanomet là độ dài bằng 1 mét chia cho 1 tỷ. Ánh sáng đỏ mà con người nhìn thấy kết thúc quanh 700 nanomet. Điều đó có nghĩa là người ta dùng ánh sáng ngay bên ngoài vùng ấy. Việc chụp được thực hiện trong phòng chặn ánh sáng bên ngoài. Ánh sáng được tỏa đều để không dồn về một phía. Nếu chỉ chiếu từ một phía, bóng sẽ kéo dài ở từng rãnh của thớ gỗ.

Chụp hồng ngoại đã trở thành quy trình cơ bản trong nghiên cứu thẻ gỗ. Bên cạnh ảnh thẻ gỗ trong báo cáo khai quật, thường có ảnh hồng ngoại đặt song song. Viện Nghiên cứu Quốc gia Nara về Di sản Văn hóa từ sớm đã phát triển thiết bị dùng máy ảnh hồng ngoại để làm rõ hình ảnh chữ trên thẻ gỗ[1].

Từ hồng ngoại đến siêu phổ

Không phải mọi chữ đều sống lại chỉ bằng một bức ảnh hồng ngoại. Bước sóng giúp chữ hiện rõ thay đổi theo vị trí vết bẩn và trạng thái thớ gỗ. Vì vậy, phương pháp ảnh siêu phổ đã xuất hiện.

Ảnh thông thường chỉ lưu ba giá trị đỏ, lục và lam. Ảnh siêu phổ thì khác. Nó chia một cảnh thành hàng chục đến hàng trăm dải bước sóng hẹp để chụp. Có thể hình dung như tách ánh sáng thành các màu cầu vồng thật nhỏ rồi chụp riêng từng phần. Với mỗi điểm trên màn hình, các giá trị độ sáng theo từng bước sóng được ghi nối tiếp. Vì vậy, thay vì một tấm ảnh, ta có một bó ảnh dày.

Khi đối chiếu các giá trị này, có thể phân biệt nơi còn mực với nơi có vết đất. Lý do là mực và đất thay đổi độ sáng theo từng bước sóng theo cách khác nhau. Có chữ hiện rõ quanh 900 nanomet. Có chữ sống lại ở dải khác. Nhà nghiên cứu chọn từng dải bước sóng để tìm tổ hợp dễ đọc nhất. Máy tính cũng tự động đối chiếu nhiều dải để lọc ra ứng viên. Công việc trước đây con người phải kiểm tra từng ảnh bằng mắt đã được rút ngắn bằng tính toán.

Năm 2025, Viện Nghiên cứu Quốc gia về Di sản Văn hóa Gaya lần đầu dùng phương pháp này. Viện đã đưa ảnh siêu phổ vào việc đọc thẻ gỗ ở thành Seongsan, Haman. Viện cho biết phương pháp này chính xác hơn chụp hồng ngoại trước đây. Giải thích của viện là nó có thể làm sống lại cả những chữ khó nhận ra bằng mắt hoặc bằng chụp ảnh thông thường[2].

Tìm chữ bằng hình dạng chữ

Ngay cả khi ảnh đã rõ hơn, vấn đề tiếp theo vẫn còn. Cần xác định dấu bút mờ là chữ Hán nào. Chữ viết theo lối thảo biến đổi hình dạng rất lớn. Nhiều chữ chỉ còn vài nét. Chữ Hán có đến hàng chục nghìn chữ, nên con người cũng có giới hạn khi tra tự điển.

Viện Nghiên cứu Quốc gia Nara về Di sản Văn hóa và Viện Biên soạn Sử liệu thuộc Đại học Tokyo đã tạo ra một hệ thống hỗ trợ việc này. Tên của nó là Mojizo. Người dùng cắt hình chữ muốn đọc rồi tải lên. Hệ thống tìm trong kho hình ảnh chữ được thu thập từ thẻ gỗ và cổ thư. Sau đó, nó tìm và hiển thị những hình giống nhau[3]. Con người xem các ứng viên rồi phán đoán. Đây không phải là máy đưa ra đáp án, mà là công cụ tìm kiếm lấy ra các trường hợp tương tự.

Điều mà công cụ như vậy thay đổi là trình tự làm việc. Trước đây, học giả dựa vào trí nhớ và tự điển để nghĩ ra ứng viên. Giờ đây, máy chọn trước những chữ giống nhau trong hàng chục nghìn trường hợp. Con người xét ngữ cảnh trong phạm vi ứng viên đã thu hẹp để chọn. Thời gian cần cho việc đọc chữ giảm mạnh.

Phân biệt nét do máy vẽ

AI tạo sinh ngày nay làm cho hình mờ trở nên rõ hơn. Nó nối các đường bị đứt và lấp chỗ trống. Nếu dùng công nghệ này cho ảnh thẻ gỗ, kết quả nhìn đẹp hơn. Nhưng ở đây có một cái bẫy.

Nét do máy lấp vào không phải là mực thật còn lại trên gỗ. Đó là suy đoán rút ra từ mức trung bình của hàng chục nghìn mẫu chữ đã học. Nếu vân gỗ còn lại thành hình tròn, máy có thể đọc nó thành nét tròn. Khi đó, một chữ vốn không tồn tại có thể xuất hiện rất có vẻ thật. Tư liệu bị ô nhiễm và cách giải thích lịch sử bị lệch.

Vì vậy, tại hiện trường nghiên cứu có quy tắc. Ảnh gốc và kết quả chỉnh sửa phải luôn được đăng cạnh nhau. Dấu mực có thể xác nhận thật và điểm ảnh do máy tạo phải được đánh dấu phân biệt. Đáp án của máy được xử lý như danh sách ứng viên, không phải kết luận. Con người thực hiện lần đọc cuối cùng. Khi công bố phương án đọc, cũng phải nêu rõ ảnh được chụp bằng thiết bị nào và trong điều kiện nào.

Việc chuẩn bị ở cấp chính phủ cũng đang được tiến hành. Ngày 16 tháng 4 năm 2026, Hội đồng Tư vấn Khoa học và Công nghệ Quốc gia đã thông qua kế hoạch thực hiện của Cục Di sản Quốc gia. Đây là kế hoạch đưa AI, robot và cảm biến vào khảo sát và phục hồi di sản quốc gia từ năm 2026 đến năm 2030. Quy mô đầu tư trong năm nay là 14 tỷ won. Trong đó, 4,4 tỷ won được phân bổ cho phát triển công nghệ bảo tồn tiên tiến thông minh[4]. Ngày 23 tháng 6 năm 2026, Bảo tàng Quốc gia Gimhae đã tổ chức hội thảo học thuật. Chủ đề là góc nhìn khoa học trong nghiên cứu di sản Gaya. Có bảy bài trình bày về phân tích không phá hủy, phục hồi số và công nghệ kính hiển vi X-ray 3D[5]. Các phương pháp nhìn vào bên trong di vật mà mắt thường không thấy được đang mở rộng ra ngoài lĩnh vực thẻ gỗ.

Tài liệu tham khảo

[1] 奈良文化財研究所. 赤外線カメラを利用した木簡文字画像鮮明化システムの開発. 学術機関リポジトリデータベース. https://irdb.nii.ac.jp/01144/0004944317 (truy cập ngày 3 tháng 9 năm 2026)

[2] 경남일보. (2025, 8월 7일). 함안 성산산성서 새로운 목간 2점 명문 확인. 경남일보. https://www.gnnews.co.kr/news/articleView.html?idxno=616028 (truy cập ngày 3 tháng 9 năm 2026)

[3] 奈良文化財研究所, 東京大学史料編纂所. 木簡・くずし字解読システム MOJIZO. https://aimojizo.nabunken.go.jp/ (truy cập ngày 3 tháng 9 năm 2026)

[4] 뉴시스. (2026, 4월 16일). "AI, 블록체인으로 국가유산 보존관리"...과기자문회의, 유산청 시행계획 의결. 뉴시스. https://www.newsis.com/view/NISX20260416_0003594216 (truy cập ngày 3 tháng 9 năm 2026)

[5] 경남일보. (2026, 6월 17일). 국립김해박물관, 23일 '가야학술제전' 심포지엄 개최. 경남일보. https://www.gnnews.co.kr/news/articleView.html?idxno=639314 (truy cập ngày 3 tháng 9 năm 2026)

8.3. Cuộc truy tìm của đội điều tra khoa học 2: KuroNet và Miwo

Chữ viết nối liền không thể tách ra

Ở Nhật Bản còn lại hàng triệu văn bản cổ được dùng đến thời Edo. Những văn bản này được viết bằng lối chữ thảo gọi là kuzushiji. Kuzushiji là kiểu chữ viết tay được dùng hơn một nghìn năm từ thế kỷ 8. Năm 1900, Nhật Bản thay đổi hệ thống chữ viết. Sau đó, nhà trường không còn dạy kiểu chữ này. Vì vậy, phần lớn người Nhật ngày nay không đọc được cả sách cách đây 150 năm[1].

Lý do khó đọc không chỉ nằm ở hình dạng chữ. Có một phép viết trong đó nhiều chữ được viết nối liền mà bút không rời khỏi giấy. Nét cuối của chữ trước và nét đầu của chữ sau nối thành một. Khi đó, giữa chữ này và chữ kia không còn đường ranh giới. Cả một dòng trông như một đường dài duy nhất.

Nhận dạng chữ bằng máy tính trong thời gian dài hoạt động theo cùng một trình tự. Trước hết, từng chữ được cắt ra thành các ô vuông. Sau đó, máy đoán hình trong ô bị cắt là chữ gì. Cách này cần có đường ranh giới mới hoạt động. Trước chữ thảo viết nối liền, nó sụp đổ ngay từ bước đầu. Nếu phần trước sai, phần sau cũng sai, nên không có cách nâng độ chính xác.

KuroNet không tách chữ

Nhóm nghiên cứu của Trung tâm Dữ liệu mở Nhân văn Nhật Bản đã mở ra một hướng khác. Tarin Clanuwat, Alex Lamb và Asanobu Kitamoto cùng xây dựng một mô hình. Tên của nó là KuroNet. Mô hình này được công bố lần đầu tại hội nghị quốc tế về phân tích và nhận dạng văn bản năm 2019[1]. Năm 2020, nhóm tinh chỉnh cấu trúc và đăng lại kết quả trên tạp chí học thuật[2].

KuroNet không cắt chữ trước. Nó nhận đầu vào là toàn bộ một trang cổ văn. Cấu trúc trung tâm của mô hình là U-Net. U-Net thu nhỏ hình ảnh dần để rút ra đặc trưng. Sau đó, nó là mạng nơ-ron phóng hình ảnh trở lại kích thước ban đầu. Phần thu nhỏ và phần phóng to đối xứng nhau, nên được vẽ như hình chữ U. Trong lúc thu nhỏ, nó học độ dày và hướng của nét bút. Trong lúc phóng to, nó đưa thông tin đó trở lại vị trí ban đầu.

Mạng nơ-ron này dự đoán cùng lúc hai việc trên toàn bộ trang. Một là tâm của chữ nằm ở đâu. Việc còn lại là chữ ở vị trí đó là chữ gì. Kết quả hiện ra dưới dạng bản đồ xác suất. Đó là hình đánh dấu khắp màn hình những nơi có khả năng có chữ. Đi theo những nơi có xác suất cao thì vị trí và tên chữ được xác định cùng lúc. Vì không có bước cắt chữ, nó không sụp đổ cả với chữ viết nối.

Dữ liệu học dùng tư liệu do Viện Nghiên cứu Văn học Quốc văn Nhật Bản tạo ra. Tên của nó là bộ dữ liệu Kuzushiji cổ điển Nhật Bản. Trong đó có hơn 1 triệu mẫu chữ[2]. Cùng nhóm nghiên cứu cũng công bố một bản thu nhỏ để các nhà nghiên cứu học máy dễ dùng. Đó là Kuzushiji-MNIST, làm theo mẫu dữ liệu chữ số viết tay. Họ cũng đưa ra bản tăng số chữ và bản chỉ gom chữ Hán[3]. Nhờ vậy, cả những nhà nghiên cứu không biết cổ điển Nhật Bản cũng có thể tham gia bài toán này.

KuroNet đã chuyển giao nhiệm vụ vào tháng 10 năm 2022. Tên của mô hình nhận dạng đổi thành Ruri[4]. Ruri là mô hình chỉnh kỹ thuật phát hiện vật thể cho phù hợp với chữ viết thảo. Ngay cả khi màu hoặc hoa văn phía sau chữ phức tạp, nó vẫn đọc tốt hơn trước. Nó được học bằng cách tăng cường cùng bộ dữ liệu và đọc được những chữ trước đây không đọc được[5].

Miwo trong lòng bàn tay

Đưa mô hình từng chạy trên máy chủ vào điện thoại di động là nhiệm vụ tiếp theo. Ứng dụng ra đời như vậy là Miwo. Nó được công bố lần đầu vào ngày 30 tháng 8 năm 2021[5]. Cách viết tiếng Nhật là みを, và cách viết tiếng Anh là Miwo[6].

Cách dùng đơn giản. Người dùng chụp tài liệu cũ bằng camera điện thoại. Một lúc sau, kết quả chuyển sang chữ hiện nay xuất hiện trên màn hình. Có thể dùng thanh trượt để chồng bản gốc và kết quả lên nhau rồi so sánh. Ứng dụng cũng hiển thị hentaigana vốn là chữ nào. Hentaigana là dạng hiragana cũ hiện nay không còn dùng. Nó cũng có chức năng xuất kết quả thành tệp văn bản. Ảnh đã chụp không được lưu trên máy chủ. Có thể dùng miễn phí trên cả Android và iPhone[6].

Từ bản 1.1 ra ngày 26 tháng 10 năm 2022, mô hình nhận dạng đổi thành Ruri[5]. Cũng trong tháng đó, ứng dụng này nhận Giải Thiết kế Tốt năm 2022[7]. Lịch sử sử dụng tiếp tục tích lũy. Tính đến ngày 3 tháng 9 năm 2026, số ảnh được nhận dạng từ khi công bố là 3.523.373 ảnh[6]. Điều đó có nghĩa là mọi người đang chụp tài liệu cũ ở thư viện, chùa chiền và kho tư liệu địa phương.

Công dân cùng đọc

Cách máy đọc trước rồi con người sửa cũng được dùng ở nơi khác. Có một dự án phiên khắc trực tuyến tên là Minna de Honkoku. Phiên khắc là việc chép chữ cũ sang chữ hiện nay. Cách viết là みんなで翻刻, nghĩa là cùng nhau chép lại. Dự án bắt đầu năm 2017 tại Hội Nghiên cứu Cổ Địa chấn của Đại học Kyoto. Hiện nay, Bảo tàng Quốc gia Lịch sử và Dân tục học cùng Viện Nghiên cứu Động đất của Đại học Tokyo vận hành dự án[8].

Cách tham gia được mở rộng. Chỉ cần có trình duyệt internet, ai cũng có thể chép lại tài liệu cũ. Trên màn hình hiện kết quả nháp do AI đọc trước. Con người chỉ cần sửa những chữ sai. Trang này dùng dịch vụ nhận dạng Kuzushiji của Trung tâm Sử dụng Chung Dữ liệu Mở Nhân văn. Cho đến nay, khoảng 10.800 người tham gia đã chép lại hơn 48 triệu chữ[8].

Bài chép lại không chỉ được tích lũy. Đáp án đúng do con người sửa lại trở thành dữ liệu học. Khi dữ liệu học tăng lên, mô hình tốt hơn. Khi mô hình tốt hơn, kết quả nháp chính xác hơn và con người phải sửa ít hơn. Khi vòng tuần hoàn này bắt đầu chạy, tốc độ sẽ tăng.

Cách dùng các tài liệu đã đọc được cũng rộng. Dự án này do các nhà nghiên cứu động đất bắt đầu. Trong nhật ký cũ và văn thư quan phủ có ghi ngày tháng và thiệt hại của động đất, sóng thần và mất mùa. Khi những ghi chép từng không thể đụng tới vì chữ viết thảo biến thành dữ liệu, có thể đếm chu kỳ thiên tai trong quá khứ. Chữ bút lông cũ trở thành tư liệu phòng chống thiên tai hôm nay.

Đến đây là câu chuyện về cổ văn giấy của Nhật Bản. KuroNet, Ruri và Miwo đều là mô hình đọc Kuzushiji viết trên giấy. Chưa xác nhận được trường hợp mô hình này tự động đọc thẻ gỗ trong bùn. Thẻ gỗ có ít chữ, nền là gỗ, và mực đã bị xóa nhiều. Dữ liệu đáp án dùng để học cũng chưa tích lũy nhiều như cổ văn giấy. Nếu dữ liệu thẻ gỗ của Hàn Quốc và Nhật Bản được gom thêm, tình hình có thể khác. Hiện nay đó là kỳ vọng, chưa phải thành quả.

Tài liệu tham khảo

[1] Clanuwat, T., Lamb, A., & Kitamoto, A. (2019). KuroNet: Pre-modern Japanese Kuzushiji character recognition with deep learning. 2019 International Conference on Document Analysis and Recognition (ICDAR), 607-614. https://doi.org/10.1109/ICDAR.2019.00103

[2] Lamb, A., Clanuwat, T., & Kitamoto, A. (2020). KuroNet: Regularized residual U-Nets for end-to-end Kuzushiji character recognition. SN Computer Science, 1, 177. https://doi.org/10.1007/s42979-020-00186-z

[3] Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., & Ha, D. (2018). Deep learning for classical Japanese literature. arXiv. https://arxiv.org/abs/1812.01718

[4] ROIS-DS 人文学オープンデータ共同利用センター. KuroNetくずし字認識サービス(AI OCR). CODH. https://codh.rois.ac.jp/kuronet/ (truy cập ngày 3 tháng 9 năm 2026)

[5] ROIS-DS 人文学オープンデータ共同利用センター. みを(miwo)とは？ CODH. https://codh.rois.ac.jp/miwo/about/ (truy cập ngày 3 tháng 9 năm 2026)

[6] ROIS-DS 人文学オープンデータ共同利用センター. みを(miwo)：AIくずし字認識アプリ. CODH. https://codh.rois.ac.jp/miwo/ (truy cập ngày 3 tháng 9 năm 2026)

[7] 国立情報学研究所. (2022, 10월 26일). 世界初のAIくずし字認識アプリ「みを(miwo)」が2022年度グッドデザイン賞を受賞. 国立情報学研究所. https://www.nii.ac.jp/news/release/2022/1026.html (truy cập ngày 3 tháng 9 năm 2026)

[8] みんなで翻刻. みんなで翻刻について. https://honkoku.org/ (truy cập ngày 3 tháng 9 năm 2026)

8.4. Sự thật được làm sáng tỏ

Đời sống hoàng thất do thẻ gỗ Wolji cho biết

Donggung và Wolji ở Gyeongju là biệt cung và ao của hoàng thất Silla Thống Nhất. Thẻ gỗ lần đầu xuất hiện trong cuộc khai quật Wolji năm 1975. Sau đó, khảo sát tiếp tục và số lượng tăng lên. Bảo tàng Quốc gia Gyeongju trưng bày di vật tìm thấy tại đây ở Nhà Wolji[1].

Thẻ gỗ Wolji cho thấy một ngày của hoàng thất. Một số thẻ gỗ là nhãn treo vào chum. Thẻ gỗ dạng nhãn có rãnh khoét sang hai bên ở đầu thẻ. Người ta quấn dây vào rãnh đó rồi buộc vào đồ vật. Công dụng của nó giống vận đơn giao hàng ngày nay. Trên thẻ gỗ tìm thấy ở Wolji có các chữ nokgap và ong. Nokgap là mắm làm từ hươu, còn ong là chum. Điều đó có nghĩa là hoàng thất Silla ghi loại mắm và thời điểm muối để quản lý[2].

Những ghi chép như vậy không xuất hiện trong sách sử. Samguk Sagi ghi vua, chiến tranh và chế độ. Nó không ghi thứ gì đã vào bếp. Thẻ gỗ lấp khoảng trống đó. Một mảnh gỗ trong bùn trở thành sổ quản lý đời sống cung điện.

Hội Thẻ gỗ Hàn Quốc và Bảo tàng Quốc gia Gyeongju đã tổ chức hội thảo quốc tế vào ngày 3 và ngày 4 tháng 4 năm 2026. Chủ đề là không gian và đời sống chữ viết của kinh đô Silla, Gyeongju. Hội thảo diễn ra trong hai ngày tại Trung tâm Hội nghị Hwabek Gyeongju. Tư liệu từ hào Wolseong, Donggung và Wolji được đặt cạnh nhau. Đó là nỗ lực vẽ lại diện mạo kinh đô Silla bằng thẻ gỗ, bia đá và chữ mực. Các nhà nghiên cứu Hàn Quốc và Nhật Bản cùng trình bày[3].

Báo cáo xử phạt trên thẻ gỗ Seongsansanseong

Seongsansanseong ở Haman là sơn thành do Silla xây vào thế kỷ 6. Thẻ gỗ tiếp tục xuất hiện ở vùng đất ướt bên trong tường thành. Tính đến tháng 8 năm 2025, số thẻ gỗ có chữ mực tìm thấy tại đây là 247 thẻ. Con số này chiếm hơn một nửa số thẻ gỗ cổ đại tìm thấy ở Hàn Quốc[4].

Hai thẻ mới được đọc trong đợt khai quật lần thứ 18 được làm từ nhóm gỗ thông. Một thẻ là thẻ gỗ nhiều mặt có chữ viết trên bốn mặt. Thẻ còn lại là thẻ gỗ hai mặt[4]. Thẻ gỗ nhiều mặt chứa nội dung hành chính trên ba trong bốn mặt. Đó là ghi chép về việc xử phạt một người[5]. Thời điểm tạo tác được xem là giữa thế kỷ 6[4].

Cách đọc này là kết quả thu được bằng ảnh siêu phổ đã xem ở trên. Những chữ không thể nhận ra bằng mắt thường hay bằng ảnh thông thường đã sống lại[4]. Báo cáo của một viên chức cấp thấp Silla nằm trong bùn hơn nghìn năm đã trở lại thành câu văn. Trên thẻ gỗ Seongsansanseong cũng ghi nhiều tên làng đã gửi ngũ cốc. Đây là tư liệu cho thấy Silla thế kỷ 6 chia và cai trị địa phương như thế nào. Cũng có thể lần theo đường thuế đã đi tới thành.

Đống văn thư của vương cung Baekje

Di tích Gwanbuk-ri ở Buyeo được biết đến là nền vương cung thời Sabi của Baekje. Viện Nghiên cứu Di sản Văn hóa Quốc gia Buyeo đã thực hiện cuộc khai quật lần thứ 16 trong năm 2024 và 2025. Tại đây tìm thấy 329 hiện vật thuộc nhóm thẻ gỗ. Có 82 thẻ gỗ có chữ và 247 mảnh bào[6]. Đây là số lượng lớn nhất tại một di tích đơn lẻ trong nước, và thuộc giai đoạn sớm trong số thẻ gỗ thời Sabi[7].

Trong các thẻ gỗ có lẫn thẻ ghi niên đại. Xuất hiện can chi gyeongsinnyeon và gyehaenyeon. Chúng lần lượt được xem là năm 540 và năm 543. Điều đó có nghĩa đây là ghi chép ngay sau khi Baekje dời đô từ Ungjin về Sabi. Hồ sơ nhân sự của quan lại, sổ ghi tài chính quốc gia và tên chức quan cùng được xác nhận[7]. Hiện vật nhạc cụ hơi Baekje làm bằng tre cũng lần đầu xuất hiện[8].

Vào ngày 22 và ngày 23 tháng 7 năm 2026, một hội thảo quốc tế đã được tổ chức. Viện Nghiên cứu Di sản Văn hóa Quốc gia Buyeo, Hội Thẻ gỗ Hàn Quốc và Hội Nghiên cứu Baekje cùng chuẩn bị. Chủ đề là nhạc

cụ và thẻ gỗ tại di tích Gwanbuk-ri, Buyeo. Hội thảo diễn ra ở quận Haeundae, Busan. Nó bắt đầu bằng bài giảng chính của Park Soon-bal, giáo sư danh dự Đại học Quốc gia Chungnam. Mười hai bài được trình bày trong ba phân ban. Các nhà nghiên cứu Hàn Quốc, Trung Quốc và Nhật Bản cùng xem xét ý nghĩa của thẻ gỗ Gwanbuk-ri[8].

Phải thận trọng trước bia đá

Phương pháp có tác dụng với thẻ gỗ không có nghĩa là có tác dụng với mọi chữ cổ. Khi nền chuyển thành đá, tình hình khác đi. Bia Gwanggaeto Đại vương là một ví dụ. Bia dựng năm 414 này khắc lịch sử Goguryeo. Mặt bia đã chịu mưa gió lâu ngày nên chữ bị mòn. Ở đây phải đọc rãnh khắc, không phải mực.

Vì vậy, các nhà nghiên cứu so sánh bản dập. Bản dập nguyên thạch là bản dập được làm trước khi bôi vôi lên mặt bia. Người ta đặt giấy trực tiếp lên đá gốc rồi gõ để lấy bản dập. Đến nay xác nhận được hơn mười bản. Cuối thế kỷ 19 xuất hiện loại khác. Đó là bản dập vôi, được làm sau khi bôi vôi lên mặt bia để chữ hiện rõ. Cũng có nhiều bản gọi là ssanggugamukbon. Đó là bản đặt giấy lên văn bia rồi vẽ đường quanh chữ. Sau đó phần trống được tô bằng mực đậm để trông như bản dập. Đây không phải là bản dập được gõ lấy, mà là bản sao chép bằng vẽ lại. Trong quá trình làm, chữ có thể bị chuyển sai[9]. Vì vậy, cùng một chữ có khi trông khác nhau giữa các bản dập.

Trung tâm tranh luận là câu thuộc đoạn năm Sinmyo. Cách giải thích câu này đã bị tranh cãi lâu dài vì gắn với lịch sử quan hệ Hàn - Nhật cổ đại. Vấn đề là chữ có bị bàn tay con người thay đổi hay không[10]. Muốn biết chữ vốn có, phải so sánh bản dập nguyên thạch sớm với nhiều bản dập vôi. Việc đầu tiên là đặt ra tiêu chuẩn để phân loại từng bản dập[9].

Có lời lan truyền rằng đo đạc 3D và AI đã đưa ra câu trả lời cho cuộc tranh luận này. Khi kiểm tra, nhiều trường hợp không tìm được nghiên cứu được nêu làm căn cứ. Kỳ vọng rằng phương pháp đã có tác dụng với thẻ gỗ cũng sẽ có tác dụng với bia đá khác với kết quả nghiên cứu thực tế. Công nghệ đã xem ở trên làm cho thứ không nhìn thấy trở nên nhìn thấy. Việc xác định ý nghĩa của thứ đã được làm cho nhìn thấy là bước tiếp theo.

Vì vậy, kết luận có hai lớp. Ảnh hồng ngoại và ảnh siêu phổ đã làm sống lại mực biến mất trên thẻ gỗ Silla và Baekje. Mô hình nhận dạng không phân đoạn đã đọc chữ viết thảo không thể tách rời trong cổ văn giấy Nhật Bản. Đối tượng mà hai công nghệ xử lý khác nhau, nhưng hướng đi giống nhau. Máy chỉ ra trước những nét mà mắt người bỏ lỡ. Đồng thời, nếu không có kỷ luật phân biệt nét do máy vẽ với nét còn lại trên bản gốc, cùng công nghệ ấy sẽ tạo ra lịch sử không từng tồn tại. Máy là bên đánh thức chữ bút lông, còn con người là bên chịu trách nhiệm về chữ đó.

Tài liệu tham khảo

[1] 국립경주박물관. 월지관. 국립경주박물관. https://gyeongju.museum.go.kr/kor/html/sub02/0201.html (truy cập ngày 3 tháng 9 năm 2026)

[2] 국사편찬위원회. 목간. 우리역사넷. https://contents.history.go.kr/mobile/tt/view.do?levelId=tt_b36 (truy cập ngày 3 tháng 9 năm 2026)

[3] 경북일보. (2026, 3월 31일). 1500년 전 신라의 목소리 깨어난다…경주서 국제 목간 학술대회. 경북일보. https://www.kyongbuk.co.kr/news/articleView.html?idxno=4068614 (truy cập ngày 3 tháng 9 năm 2026)

[4] 경남일보. (2025, 8월 7일). 함안 성산산성서 새로운 목간 2점 명문 확인. 경남일보. https://www.gnnews.co.kr/news/articleView.html?idxno=616028 (truy cập ngày 3 tháng 9 năm 2026)

[5] 경남도민일보. (2025, 8월 7일). 누군가를 처벌한 신라 행정 보고, 함안 성산산성 목간서 확인. 경남도민일보. https://www.idomin.com/news/articleView.html?idxno=943461 (truy cập ngày 3 tháng 9 năm 2026)

[6] 파이낸셜뉴스. (2026, 2월 5일). 부여 관북리 '백제 피리' 첫 확인..목간 329점 무더기 발견. 파이낸셜뉴스. https://www.fnnews.com/news/202602051448564101 (truy cập ngày 3 tháng 9 năm 2026)

[7] 경향신문. (2026, 2월 17일). 부여 관북리서 발견된 목간 329점, '백제 피리'만큼이나 중요한 이유. 경향신문. https://www.khan.co.kr/article/202602171427001 (truy cập ngày 3 tháng 9 năm 2026)

[8] 이데일리. (2026, 7월 21일). 백제 왕궁터 출토 목간 329점·횡적 발굴 성과, 국제학술대회서 조명. 이데일리. https://edaily.co.kr/News/Read?mediaCodeNo=257&newsId=03276726645516816 (truy cập ngày 3 tháng 9 năm 2026)

[9] 국립중앙박물관. 광개토대왕비 탁본. 국립중앙박물관 큐레이터 추천 소장품. https://www.museum.go.kr/MUSEUM/contents/M0501000000.do?schM=view&relicRecommendId=254462 (truy cập ngày 3 tháng 9 năm 2026)

[10] 국사편찬위원회. 광개토대왕릉비. 우리역사넷. https://contents.history.go.kr/mobile/kc/view.do?levelId=kc_r100300 (truy cập ngày 3 tháng 9 năm 2026)

Chương 9. Bí ẩn thất lạc của biển Aegea: chữ tuyến tính A và lưới hình học

9.1. Hồ sơ vụ việc

Ba loại chữ từ cung điện trên đảo Crete

Ở phía nam biển Aegea có đảo Crete. Vào thời đại đồ đồng, nền văn minh Minoa đã phát triển trên hòn đảo này. Người Minoa đóng thuyền và đi lại trên Địa Trung Hải. Họ chở dầu ô liu, rượu vang và ngũ cốc. Cung điện thu các món hàng này rồi phân phát cho người dân.

Khoảng năm 1900, nhà khảo cổ học người Anh Arthur Evans khai quật di tích cung điện Knossos. Ông tìm thấy những bảng đất sét có khắc chữ dưới lòng đất. Chữ viết không chỉ có một loại. Evans chia các chữ tìm được thành ba loại. Đó là chữ tượng hình Crete, chữ tuyến tính A và chữ tuyến tính B[1].

Nền văn minh Minoa kéo dài hơn một nghìn năm. Rồi đến cuối thời đại đồ đồng, nó trải qua biến động lớn. Một biến động là núi lửa phun trào trên đảo Santorini ở phía bắc. Thời điểm phun trào được cho là từ thế kỷ 17 đến thế kỷ 16 trước Công nguyên. Niên đại chính xác đến nay vẫn còn gây tranh luận[7]. Tro núi lửa được cho là đã bay tới đảo Crete. Sóng thần cũng được cho là đã đánh vào bờ biển. Khoảng năm 1450 trước Công nguyên, người Hy Lạp Mycenae chiếm hòn đảo này[1]. Giữa núi lửa và sự cai trị của Mycenae có khoảng cách hơn một trăm năm. Khó có thể nối ngay hai sự kiện thành nguyên nhân và kết quả. Sau đó, ngôn ngữ của người Minoa biến mất khỏi ghi chép.

Ba loại chữ được dùng vào các thời kỳ khác nhau. Chữ tuyến tính A được dùng từ khoảng năm 1750 trước Công nguyên đến khoảng năm 1450 trước Công nguyên. Chữ tuyến tính B xuất hiện muộn hơn. Đó là chữ được dùng sau khi người Hy Lạp Mycenae chiếm đảo Crete. Hai loại chữ này cùng có nhiều ký hiệu giống nhau về hình dạng[1].

Việc các bảng đất sét còn lại là chuyện ngẫu nhiên. Người Minoa không nung bảng đất sét. Họ khắc chữ lên đất ướt, phơi khô, dùng khoảng một năm rồi bỏ. Nhưng khi cung điện bị cháy, các bảng đất được nung như đồ gốm. Chỉ những bảng bị lửa nung mới chịu được ba nghìn năm trăm năm. Tư liệu còn lại hôm nay là tư liệu do lửa để lại.

Trong ba loại chữ mà Evans phân chia, chữ tượng hình Crete xuất hiện sớm nhất. Đó là loại chữ còn giữ rõ hình đầu người hoặc hình chiếc bát. Chữ tuyến tính A gần với hình ảnh đã rút gọn các hình vẽ ấy thành nét. Hai loại chữ này được dùng cùng thời trong gần hai trăm năm. Không phải một loại biến mất rồi loại kia mới sinh ra. Có quan điểm cho rằng chữ tuyến tính A được tạo ra theo mẫu chữ tượng hình Crete. Tuy nhiên, không thể khẳng định ba loại chữ này nối nhau theo một đường thẳng. Rõ ràng chúng có chung một số ký hiệu. Nhưng con đường nối tiếp vẫn chưa được làm sáng tỏ[1].

Chữ tuyến tính B đã được đọc

Chữ tuyến tính B được đọc vào thập niên 1950. Kiến trúc sư người Anh Michael Ventris đã làm được việc đó. Ông chứng minh rằng ngôn ngữ được ghi bằng chữ tuyến tính B là một dạng cổ của tiếng Hy Lạp. Nhà cổ điển học John Chadwick hỗ trợ nghiên cứu của ông[1].

Ventris bắt đầu bằng cách vẽ một bảng. Hàng ngang đặt phụ âm, hàng dọc đặt nguyên âm[8]. Bảng như vậy được gọi là lưới âm tiết. Đó là hình vẽ chuẩn bị sẵn các ô cho từng âm tiết. Ventris dùng thống kê để thu hẹp ký hiệu nào sẽ đi vào ô nào[1].

Ông đưa các địa danh như Knossos và Phaistos vào những ô đã thu hẹp. Khi địa danh khớp, các ký hiệu còn lại cũng lần lượt được giải. Các học giả thử áp dụng cùng phương pháp cho chữ tuyến tính A. Họ

chuyển nguyên giá trị âm của chữ tuyến tính B sang các ký hiệu giống nó[3]. Âm đọc xuất hiện. Nhưng không ai hiểu âm đó là ngôn ngữ gì[1].

Tình huống đã chuyển giá trị âm mà vẫn không ra nghĩa là điều lạ. Vì đọc chữ và hiểu lời nói là hai việc khác nhau. Người biết chữ Hangeul có thể đọc một ngoại ngữ lạ nếu nó được ghi bằng Hangeul. Âm đọc xuất hiện, nhưng nghĩa thì không biết. Hoàn cảnh của các học giả đứng trước chữ tuyến tính A cũng như vậy.

Không có tấm đá ghi bằng hai ngôn ngữ

Khi đọc chữ cổ, các học giả thường tìm tài liệu đối chiếu. Đó là văn bản ghi cùng một nội dung bằng hai ngôn ngữ đặt cạnh nhau. Văn bản như vậy được gọi là bia song ngữ. Chữ tượng hình Ai Cập được giải nhờ đá Rosetta. Trên đá Rosetta có khắc cùng một sắc lệnh bằng ba loại chữ. Một trong số đó là tiếng Hy Lạp cổ mà các học giả đã biết đọc.

Chữ hình nêm Mesopotamia cũng đi theo con đường tương tự. Trên vách đá Behistun ở Iran còn lại một bia khắc bằng ba ngôn ngữ. Ban đầu cả ba ngôn ngữ trên bia này đều không đọc được. Trong đó, tiếng Ba Tư cổ được giải trước. Vì ký hiệu ít và ranh giới giữa các từ được tách ra. Các học giả lấy tiếng Ba Tư cổ đã giải được làm điểm tựa để đọc phần còn lại.

Trên đảo Crete không xuất hiện văn bản như vậy. Không có di vật nào ghi bản tiếng Hy Lạp đối chiếu bên cạnh chữ tuyến tính A. Chữ viết mà người Minoa để lại chỉ được ghi bằng ngôn ngữ của người Minoa. Không có cặp để so sánh, nên con đường kiểm chứng ý nghĩa bị chặn lại[1].

Cũng không thấy dấu hiệu nào có thể làm manh mối. Với Champollion, người đọc chữ tượng hình Ai Cập, có cartouche. Đó là khung dài bao quanh tên vua. Người ta đã biết trước rằng chữ nằm trong khung này là tên người. Từ đó, Champollion lấy ra vài giá trị âm để giải phần còn lại. Trên bảng đất sét chữ tuyến tính A không có khung như vậy[1].

Còn một vấn đề nữa. Đó là không biết gốc rễ của ngôn ngữ mà chữ tuyến tính A ghi lại. Các học giả gọi ngôn ngữ này là tiếng Minoa. Tiếng Minoa thuộc ngữ hệ nào vẫn chưa được làm rõ. Có ý kiến cho rằng đó là ngữ hệ Ấn-Âu. Cũng có ý kiến cho rằng đó là hệ ngôn ngữ Tây Á. Trong số đó, các giả thuyết được xem là có căn cứ khá hơn là hệ Semit và hệ Lycia. Tiếng Lycia là một ngôn ngữ cổ từng được dùng ở phía tây nam Thổ Nhĩ Kỳ ngày nay[1]. Không giả thuyết nào nhận được sự đồng thuận của giới học thuật.

Chữ tuyến tính A không chỉ xuất hiện trên đảo Crete. Nó cũng được tìm thấy trên nhiều đảo ở biển Aegea và trên lục địa Hy Lạp[1]. Nó còn xuất hiện tại di tích đảo Santorini bị chôn vùi dưới núi lửa. Chữ viết cũng lan theo các tuyến đường biển mà người Minoa đi qua. Đây là tư liệu cho thấy Địa Trung Hải thời đại đồ đồng được kết nối rộng đến mức nào.

Đó cũng là lý do các học giả bám sát chữ tuyến tính A. Chữ tuyến tính A thuộc nhóm chữ viết cổ xuất hiện trên đất châu Âu. Đây là ghi chép duy nhất mà người Minoa để lại bằng ngôn ngữ của chính họ. Nếu chữ tuyến tính A được giải, ta sẽ biết họ tin vào điều gì và mua bán những gì. Cội nguồn của câu chuyện Minotaur còn lại trong thần thoại Hy Lạp cũng nằm ở đây.

1.400 sổ sách ngắn

Lượng tư liệu còn lại cũng không dồi dào. Minh văn là chữ được khắc hoặc vẽ trên di vật. Số minh văn chữ tuyến tính A đã được xác nhận đến nay vào khoảng 1.370. Tổng số ký hiệu ghi trong đó cũng chỉ khoảng 7.400[1]. Đây là lượng không thể so với tư liệu ghi chép của Ai Cập hay Mesopotamia.

Độ dài văn bản cũng ngắn. Phần lớn bảng đất sét là sổ kho. Chúng ghi món gì nhập vào bao nhiêu và xuất cho ai bao nhiêu. Chỉ có tên người, tên đồ vật và con số nối tiếp nhau. Những câu đúng nghĩa rất hiếm. Khi không có câu, cách tìm ra ngữ pháp cũng giảm đi.

Khi đếm số ký hiệu, cần chia thành từng lớp. Ký hiệu ghi âm tiết có 97 cái. Ký hiệu biểu thị trọn tên đồ vật có khoảng 50 cái. Có 4 ký hiệu ghi số và 17 ký hiệu ghi phân số. Khoảng 150 tổ hợp tạo bằng cách chồng hai ký hiệu lên nhau cũng đã được biết đến[6]. Danh mục ký hiệu chuẩn gom tất cả những thứ này lên tới 300 ký hiệu[2]. Trong số đó, ký hiệu xuất hiện thường xuyên không nhiều. Phần còn lại chỉ xuất hiện vài lần rồi thôi. Muốn giao thống kê cho máy tính, cùng một ký hiệu phải xuất hiện nhiều lần. Những ký hiệu chỉ xuất hiện vài lần thì thống kê không bắt được.

Vì vậy, các nhà nghiên cứu quyết định làm lại chính tư liệu. Đó là công việc chuyển các hình sao chép bằng tay sang dạng máy tính có thể xử lý[2]. Năm 2026, cũng xuất hiện phương pháp đọc lại vết khắc bằng ánh sáng[4]. Phương pháp này làm lộ cả những rãnh nông mà mắt thường bỏ sót[5]. Công việc đánh thức sự im lặng của biển Aegea bắt đầu lại từ việc chỉnh lý tư liệu như vậy.

Tài liệu tham khảo

[1] Braović, M., Krstinić, D., Štula, M., & Ivanda, A. (2024). A systematic review of computational approaches to deciphering Bronze Age Aegean and Cypriot scripts. Computational Linguistics, 50(2), 725-779. https://doi.org/10.1162/coli_a_00514

[2] Salgarella, E., & Castellan, S. (2021). SigLA: The Signs of Linear A. A palaeographical database. In Grapholinguistics in the 21st Century 2020: Proceedings (Grapholinguistics and Its Applications, Vol. 5, pp. 945-962). Fluxus Editions. https://sigla.phis.me/ (truy cập ngày 3 tháng 9 năm 2026)

[3] Younger, J. G. (n.d.). Linear A texts and inscriptions in phonetic transcription. University of Kansas. http://people.ku.edu/~jyounger/LinearA/ (truy cập ngày 3 tháng 9 năm 2026)

[4] Giorgi, L., & Lopez, S. (2026). Aegean Bronze Age writing systems through RTI: New potential and perspectives. Digital Applications in Archaeology and Cultural Heritage, 41, e00508. https://www.sciencedirect.com/science/article/abs/pii/S2212054826000159 (truy cập ngày 3 tháng 9 năm 2026)

[5] GreekReporter. (2026, 3월 14일). New technology reveals hidden secrets of 3,500-year-old Bronze Age Aegean scripts. greekreporter.com. https://greekreporter.com/2026/03/14/new-technology-reveals-hidden-secrets-aegean-bronze-age-scripts/ (truy cập ngày 3 tháng 9 năm 2026)

[6] Del Freo, M. (2022). Linear A. Mnamon: Ancient writing systems in the Mediterranean. Scuola Normale Superiore. https://doi.org/10.25429/sns.it/lettere/mnamon022 (truy cập ngày 3 tháng 9 năm 2026)

[7] Pearson, C. L., Brewer, P. W., Brown, D., Heaton, T. J., Hodgins, G. W. L., Jull, A. J. T., Lange, T., & Salzer, M. W. (2018). Annual radiocarbon record indicates 16th century BCE date for the Thera eruption. Science Advances, 4(8), eaar8241. https://doi.org/10.1126/sciadv.aar8241

[8] Stavrakakis, E. (2024). Ventris' Grids. Bulletin of the Institute of Classical Studies, 67(1), 86-103. https://doi.org/10.1093/bics/qbae033

9.2. Cuộc truy tìm của đội điều tra khoa học: khớp lưới âm vị không giám sát

Cỗ máy tìm cặp mà không cần từ điển

Muốn làm máy dịch, thông thường cần câu đối dịch. Người ta xếp song song câu tiếng Hàn và câu tiếng Anh rồi cho máy xem. Tư liệu như vậy được gọi là ngữ liệu song song. Chữ tuyến tính A không có ngữ liệu song song[3].

Vì vậy, các nhà nghiên cứu chọn cách học không cần bảng đáp án. Đó là cách không đưa trước đáp án, mà để máy tìm quy luật chỉ bằng thống kê. Cách này được gọi là học không giám sát. Máy đếm xem các ký hiệu xuất hiện theo thứ tự nào. Rồi nó so sánh với thống kê của hệ chữ khác để tìm cặp tương ứng.

Năm 2019, nhóm nghiên cứu của Viện Công nghệ Massachusetts ở Hoa Kỳ đưa ra phương pháp này. Đây là mô hình do Jiaming Luo, Yuan Cao và Regina Barzilay cùng xây dựng. Họ biến việc nối hai hệ chữ thành bài toán dòng chảy chi phí tối thiểu. Đó là cách đặt dòng nước giữa ký hiệu này và ký hiệu kia, rồi gán giá trị cho dòng nước chảy. Dòng nước có chi phí thấp nhất có khả năng cao là cặp đúng[1].

Nhóm nghiên cứu thử mô hình này với chữ Ugarit. Nó vượt kết quả tốt nhất trước đó 5 điểm phần trăm. Họ cũng thử với chữ tuyến tính B. Trong các từ có cùng gốc với tiếng Hy Lạp, mô hình chuyển đúng 67,3%[1]. Tuy nhiên, chữ tuyến tính B đã có đáp án. Nhóm nghiên cứu đã biết trước rằng chữ đó ghi tiếng Hy Lạp.

Có lý do khiến nhóm nghiên cứu chọn chữ tuyến tính B làm bàn thử nghiệm. Chỉ với bài toán đã biết đáp án mới có thể đo kết quả của mô hình[1]. Nếu đưa ngay vào bài toán chưa có đáp án, không có cách nào biết kết quả đúng hay sai. Vì vậy, trước hết họ đo hiệu năng bằng chữ đã đọc được, rồi chuyển sang chữ viết chưa được giải mã[3].

Thứ mô hình đưa ra không phải là bản dịch hoàn chỉnh. Đó là danh sách các ứng viên giá trị âm có thể gắn với từng ký hiệu[1]. Con người nhận danh sách này rồi lọc bằng cách so với di vật. Máy tính đảm nhận việc giảm khối lượng mà con người phải xem[3].

Biến hình dạng âm thanh thành tọa độ

Chữ tuyến tính A có hoàn cảnh khác. Ngay từ đầu vẫn chưa rõ phải so sánh nó với ngôn ngữ nào[3]. Năm 2021, cùng phòng nghiên cứu đó đưa ra một mô hình xử lý điều kiện này. Luo, Frederic Hartmann và Enrico Santos đã xây dựng nó cùng Barzilay và Cao[2].

Mô hình này xử lý cùng lúc hai khó khăn. Khó khăn thứ nhất là không có ranh giới từ. Chữ cổ thường không có khoảng cách giữa các từ. Khó khăn thứ hai là chưa xác định được ngôn ngữ họ hàng[2].

Nhóm nghiên cứu biến hình dạng của âm thanh thành tọa độ số rồi đưa vào mô hình. Họ đặt đặc tính của phụ âm và nguyên âm lên tọa độ dựa trên Bảng ký hiệu ngữ âm quốc tế. Các âm phát bằng môi nằm gần nhau trên tọa độ. Chúng ở xa các âm phát bằng gốc lưỡi. Khi có tọa độ này, máy ít tạo ra các cặp sai lệch hơn[2].

Nhóm nghiên cứu thử mô hình này với tiếng Goth và chữ Ugarit. Sau đó họ áp dụng cho tiếng Iberia, ngôn ngữ vẫn chưa được giải nghĩa. Giá trị âm của các ký hiệu trong chữ Iberia đã được biết khá nhiều. Thứ chưa được giải là ngôn ngữ mà các âm đó ghi lại. Nhóm nghiên cứu không tìm thấy bằng chứng cho thấy tiếng Iberia có họ hàng với tiếng Basque. Kết quả này không trái với quan điểm chủ lưu của giới học thuật. Họ cũng trình bày quy trình so sánh các ứng viên có thể là ngôn ngữ họ hàng[2]. Cùng phương pháp đó có thể dùng nguyên như vậy trong nghiên cứu chữ tuyến tính A[3].

Quy luật do ô trống cho biết

Thứ máy so sánh không chỉ là hình dạng ký hiệu. Thói quen ký hiệu đi cùng các ký hiệu lân cận cũng được đưa vào tính toán. Có ký hiệu chỉ xuất hiện ở đầu từ. Có ký hiệu chỉ gắn ở cuối. Cũng có trường hợp hai ký hiệu hoàn toàn không đứng cạnh nhau[3].

Thống kê như vậy được gọi là mức độ cùng xuất hiện, tức thống kê đồng hiện. Khi gom thống kê đồng hiện, các ô trống trong bảng lộ ra. Các ký hiệu nằm trên cùng hàng ngang có khả năng cao có cùng phụ âm. Các ký hiệu nằm trên cùng hàng dọc có khả năng cao có cùng nguyên âm[8]. Máy tính đang lặp lại nhanh hơn rất nhiều công việc mà Ventris từng làm bằng tay[3].

Máy tính hoàn tất phép tính này nhanh hơn con người rất nhiều. Nếu có 300 ký hiệu, số cặp tương ứng lên tới hàng chục nghìn. Đó là lượng quá lớn để con người vừa ghi ra giấy vừa đếm. Ventris đã làm việc này trong nhiều năm[3].

Đến đây mới chỉ là việc vẽ cấu trúc âm thanh. Dù vẽ được cấu trúc, ý nghĩa của âm thanh ấy vẫn chưa hiện ra. Việc lấp đầy lưới và việc hiểu lời nói là hai việc riêng biệt[3].

Biến hình dạng chữ thành dữ liệu

Muốn máy đưa ra thống kê, tư liệu phải ở dạng máy có thể đọc. Chỉ dùng hình in trong sách giấy thì không thể bắt máy tính toán. Cổng tư liệu đảm nhận việc này có tên là SigLA. Esther Salgarella ở Cambridge và Simon Castellan ở Rennes cùng xây dựng nó[4].

SigLA chứa 300 ký hiệu chuẩn và hơn 400 minh văn được sao chép bằng tay. Trong các minh văn này có thể tìm thấy hơn 3.000 ký hiệu riêng lẻ. Dữ liệu và hình ảnh được công khai để ai cũng có thể dùng. Cũng có thể tách riêng nét của từng ký hiệu để so sánh[4].

Nghiên cứu công bố vào tháng 3 năm 2026 đã mở rộng dữ liệu thêm một lớp. Lavinia Giorgi và Sara Lopez đã áp dụng kỹ thuật chụp ảnh biến đổi phản xạ cho chữ viết Aegea. Đây là phương pháp chiếu ánh sáng từ nhiều hướng lên cùng một hiện vật và chụp nhiều ảnh. Khi ghép các ảnh này lại, những chỗ lồi lõm nhỏ trên bề mặt hiện ra. Thậm chí có thể xem xét dao khắc đã khắc nét nào trước[7].

Khi loại dữ liệu này tích lũy, nó cũng trở nên hữu ích cho các chữ viết khác. Năm 2022, nhóm nghiên cứu của Đại học Bologna đã xem xét lại chữ Cypro-Minoan. Đây là nghiên cứu của Michele Corazza, Fabio Tamburini, Miguel Valério và Silvia Ferrara. Họ so sánh hình dạng ký hiệu bằng mạng nơ-ron không đưa kiến thức có sẵn vào trước. Kết quả là họ kết luận cần xem xét lại cách phân loại cũ vốn chia thành ba nhóm. Theo họ, xem đó là một hệ chữ viết duy nhất phù hợp hơn với thống kê[5].

Kết quả do máy tính đưa ra không trở thành bài báo ngay. Nó phải trải qua quá trình được nhà khảo cổ học và nhà ngữ văn học xem xét lại[3]. Họ đối chiếu tấm đất sét nào được tìm thấy trong phòng nào. Họ so sánh thói quen của nét chữ để xem có phải cùng một thư lại viết hay không[7]. Chỉ sau khi vượt qua bước kiểm tra này, nó mới được đưa vào tư liệu học thuật[3].

Các nghiên cứu xử lý cùng lúc ba chữ viết Aegea cũng đang tiếp tục. Đó là cách đặt chữ tượng hình Crete, chữ tuyến tính A và chữ Cypro-Minoan cạnh nhau. Khi so sánh ba chữ viết, những quy tắc không thấy được khi chỉ nhìn một chữ viết sẽ hiện ra. Máy tính phát huy sức mạnh trong kiểu so sánh quy mô lớn này[3].

Việc máy móc không thể làm thay

Những phương pháp này vẫn chưa làm sáng tỏ ý nghĩa của chữ tuyến tính A. Vẫn còn ba bức tường phải vượt qua.

Thứ nhất là đối tượng để so sánh. Căn chỉnh không giám sát phát huy tác dụng khi có quan hệ thân thuộc giữa chữ viết chưa được giải mã và chữ viết đã biết[1]. Tiếng Minoan không có ngôn ngữ họ hàng nào còn sống sót[3]. Tọa độ có thể được vẽ ra, nhưng không có điểm chuẩn để cố định các tọa độ đó.

Thứ hai là lượng dữ liệu. 1.370 minh văn là quá ít để dùng cho học máy[3]. Khi dữ liệu ít, mạng nơ-ron lớn sẽ tạo ra những thứ có vẻ hợp lý nhưng không có thật. Lỗi này được gọi là ảo giác.

Thứ ba là kiểm chứng. Một bài viết đăng trên The Conversation vào tháng 7 năm 2026 đã chỉ ra vấn đề này. Bài viết do Jane Adkins của Dublin City University viết. Bà viết rằng việc khớp mẫu bằng thống kê không thể tạo ra ý nghĩa vốn không tồn tại. Bà chỉ ra rằng cần có chỗ dựa như một ngôn ngữ đã biết hoặc văn bản song ngữ để phân biệt sự trùng hợp ngẫu nhiên với quy tắc thật. Máy chỉ giúp thu hẹp ứng viên, còn việc phân định đúng sai vẫn thuộc về con người[6].

Tài liệu tham khảo

[1] Luo, J., Cao, Y., & Barzilay, R. (2019). Neural decipherment via minimum-cost flow: From Ugaritic to Linear B. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. https://aclanthology.org/P19-1303/ (truy cập ngày 3 tháng 9 năm 2026)

[2] Luo, J., Hartmann, F., Santus, E., Barzilay, R., & Cao, Y. (2021). Deciphering undersegmented ancient scripts using phonetic prior. Transactions of the Association for Computational Linguistics, 9, 69-81. https://doi.org/10.1162/tacl_a_00354

[3] Braović, M., Krstinić, D., Štula, M., & Ivanda, A. (2024). A systematic review of computational approaches to deciphering Bronze Age Aegean and Cypriot scripts. Computational Linguistics, 50(2), 725-779. https://doi.org/10.1162/coli_a_00514

[4] Salgarella, E., & Castellan, S. (2021). SigLA: The Signs of Linear A. A palaeographical database. In Grapholinguistics in the 21st Century 2020: Proceedings (Grapholinguistics and Its Applications, Vol. 5, pp. 945-962). Fluxus Editions. https://sigla.phis.me/ (truy cập ngày 3 tháng 9 năm 2026)

[5] Corazza, M., Tamburini, F., Valério, M., & Ferrara, S. (2022). Unsupervised deep learning supports reclassification of Bronze age cypriot writing system. PLOS ONE, 17(7), e0269544. https://doi.org/10.1371/journal.pone.0269544

[6] Adkins, J. (2026, 7월 27일). Cracking the code: can AI help us decipher ancient languages? The Conversation. https://theconversation.com/cracking-the-code-can-ai-help-us-decipher-ancient-languages-288238 (truy cập ngày 3 tháng 9 năm 2026)

[7] Giorgi, L., & Lopez, S. (2026). Aegean Bronze Age writing systems through RTI: New potential and perspectives. Digital Applications in Archaeology and Cultural Heritage, 41, e00508. https://www.sciencedirect.com/science/article/abs/pii/S2212054826000159 (truy cập ngày 3 tháng 9 năm 2026)

[8] Stavrakakis, E. (2024). Ventris' Grids. Bulletin of the Institute of Classical Studies, 67(1), 86-103. https://doi.org/10.1093/bics/qbae033

9.3. Sự thật đã được làm sáng tỏ

Số và phân số đã được giải

Phần được giải trước trong chữ tuyến tính A là cách đếm số. Người Minoan dùng hệ thập phân, tức cách đếm theo nhóm mười. Một vạch dọc là 1, một vạch ngang là 10. Hình tròn là 100, còn hình tròn có tia tỏa ra là 1000[3]. Chỉ nhìn hình cũng có thể đoán được quy tắc này.

Phần khó là ký hiệu phân số. Trên các tấm đất sét xuất hiện mười bảy ký hiệu phân số[8]. Nếu tính cả những ký hiệu viết chồng hai ký hiệu lên nhau, con số gần ba mươi[9]. Trong thời gian dài, người ta không giải được ký hiệu nào có nghĩa là bao nhiêu.

Năm 2021, nhóm nghiên cứu của Đại học Bologna đã xử lý vấn đề này. Đây là nghiên cứu của Michele Corazza, Silvia Ferrara, Barbara Montecchi, Fabio Tamburini và Miguel Valério. Họ kết hợp phương pháp cũ, tức xem xét hình dạng ký hiệu, với tính toán bằng máy tính[1].

Trước hết, nhóm nghiên cứu sắp xếp các quy tắc mà các ký hiệu tuân theo trên tấm đất sét. Họ đếm ký hiệu nào xuất hiện cạnh ký hiệu nào và ký hiệu nào không đi cùng nhau. Sau đó, họ dùng máy tính xem xét các tổ hợp giá trị. Số ứng viên gần 4 triệu. Họ thu hẹp các ứng viên này bằng kiểm định thống kê và so sánh với hệ phân số của các nền văn minh khác[1].

Kết quả còn lại như sau. Trong hệ phân số Minoan, giá trị nhỏ nhất là 1 phần 60. Hệ này có thể ghi phần lớn các giá trị lấy 60 làm mẫu số. Ký hiệu có nghĩa là 1 phần 10 được truyền nguyên sang chữ tuyến tính B. Trong chữ tuyến tính B, ký hiệu này được dùng làm đơn vị đo ngũ cốc khô. Giữa ghi chép của hai chữ viết có độ lệch thời gian khoảng 250 năm. Dù vượt qua khoảng cách ấy, cách tính vẫn còn tồn tại[1].

Khi cách tính được giải, cấu trúc của sổ sách cũng bắt đầu hiện ra. Các tấm đất sét tuân theo cách ghi ghép ký hiệu mặt hàng với số lượng. Cùng một mẫu biểu lặp lại ở nhiều di tích. Điều đó có nghĩa là các cung điện trên đảo Crete cùng chia sẻ một quy tắc kế toán[3]. Dù không biết ý nghĩa, tính chất của văn bản vẫn hiện ra như vậy.

Địa danh đọc được, nhưng ý nghĩa chưa được giải

Ở phương diện âm thanh cũng có thu hoạch. John Younger của Đại học Kansas, Hoa Kỳ, đã công bố danh mục chuyển các minh văn chữ tuyến tính A thành âm. Đây là bản phiên âm thử, trong đó ông gán cùng giá trị âm cho các ký hiệu có hình dạng giống trong chữ tuyến tính B[2].

Trong bản phiên âm này xuất hiện nhiều từ có vẻ là địa danh cổ của đảo Crete. Đó là những tên còn lại ở các thời kỳ sau, như Knossos và Phaistos[2]. Việc địa danh khớp cho thấy giả thuyết về giá trị âm không hoàn toàn sai[3].

Giả thuyết về giá trị âm cũng có giới hạn rõ ràng. Dù hình dạng trong hai chữ viết trông giống nhau, không có gì bảo đảm âm cũng giống nhau. Những người tạo ra chữ tuyến tính B đã mượn chữ của người khác để ghi tiếng của mình. Trong quá trình vay mượn, âm có thể đã chuyển đổi[3]. Vì vậy, bản phiên âm chỉ được dùng như một cách đọc tạm thời[2].

Nhưng chỉ đến đó. Sau khi loại địa danh ra, ý nghĩa của các từ còn lại không nối lại được. Có thể đọc theo âm, nhưng không biết chúng có nghĩa gì. Cây cầu giữa âm và nghĩa vẫn chưa được bắc[3].

Sự xôn xao năm 2026

Mùa hè năm 2026, có tuyên bố rằng chữ tuyến tính A đã được đọc. Đó là tuyên bố của kỹ sư AI Tom Di Mino. Ông không phải là người được đào tạo chính quy về ngôn ngữ học[4].

Di Mino cho biết ông viết mã Python để rà soát hai nguồn dữ liệu. Đó là GORILA, nơi tập hợp dữ liệu sao chép bằng tay, và SigLA đã nói ở trên. Trong công bố đầu tiên, ông nói rằng mình đã gán giá trị âm cho 40 ký hiệu và lập danh sách 408 từ[4]. Đến tháng 8, ông đưa ra một bài viết dài 41 trang và mở rộng danh sách lên 508 mục. Ông đăng kèm cách đọc của 42 ký hiệu và bản dịch của 443 minh văn. Ông cho rằng tiếng Minoan thuộc ngữ hệ Semitic, cùng hệ với tiếng Hebrew và tiếng Aramaic[5].

Giới học thuật không chấp nhận tuyên bố này. Lý do là chưa có bài báo nào qua bình duyệt được công bố[4]. Có tin rằng Đại học Rutgers và Đại học Cambridge đang tiến hành xem xét. Tuy nhiên, kết luận vẫn chưa có[5]. Trong một trăm năm qua, đã nhiều lần có tuyên bố đọc được chữ tuyến tính A. Vì vậy, các nhà nghiên cứu tiếp cận tuyên bố mới một cách thận trọng[6].

Cũng trong tháng 8, một cách diễn giải khác xuất hiện. Nó liên quan đến minh văn trên một bàn đá tìm thấy tại khu thánh trên đỉnh núi Juktas. Đó là bàn dùng để rót rượu dâng thần. Đề xuất này cho rằng hai tên được khắc trên đá nối với các vị thần Hy Lạp về sau. Đề xuất này cũng vẫn chỉ là giả thuyết. Lý do là chữ tuyến tính A chưa được giải chắc chắn[7].

Câu chuyện này cũng để lại điều gì đó cho toàn bộ nghiên cứu AI. Máy móc giỏi tìm quy tắc ở nơi có quy tắc. Kết quả đã xuất hiện trong dữ liệu buộc phải khớp trước sau, như ký hiệu phân số[1]. Ngược lại, có những nơi chỉ quy tắc thôi thì không chạm tới được. Việc âm nào gắn với nghĩa nào là vấn đề quy ước. Khi tất cả những người biết quy ước ấy đã biến mất, con đường khôi phục bằng tính toán trở nên hẹp lại[6].

Việc để ngỏ phần dữ liệu còn lại

Phương pháp nghiên cứu cũng đã thay đổi. Trước đây, hình minh văn chỉ có trong những tập sách ảnh đắt tiền. Chỉ người có được vài cuốn sách ấy mới có thể làm thống kê. Hiện nay, dữ liệu được công khai trên internet. Sinh viên cũng có thể tải xuống và đếm ký hiệu[3].

Khi dữ liệu được mở, việc kiểm chứng cũng nhanh hơn. Nếu ai đó đưa ra cách đọc mới, người khác có thể kiểm tra bằng cùng dữ liệu[6]. Trong sự xôn xao mùa hè năm 2026, quá trình kiểm chứng cũng vận hành như vậy[4]. Bên đưa ra tuyên bố và bên xem xét đang nhìn vào cùng một dữ liệu[5].

Nhờ dữ liệu được công khai, ngưỡng tham gia nghiên cứu cũng thấp hơn. Người không thuộc trường đại học cũng có thể thử thống kê ký hiệu. Khi có thêm những con mắt mới, đôi khi những quy tắc từng bị bỏ lỡ được phát hiện. Nhưng cùng với đó, các tuyên bố chưa được kiểm chứng cũng tăng lên. Công khai và kiểm chứng phải đi cùng nhau[6].

Những chữ vẫn chưa đọc được

AI đã làm được bốn việc trong chữ tuyến tính A. Máy móc đã gom dữ liệu rải rác về một chỗ. Nó biến hình dạng và nét của ký hiệu thành số để có thể so sánh[3]. Giá trị của ký hiệu phân số được thu hẹp nhờ tính toán máy tính hỗ trợ cách đọc của con người. Nó giảm một bài toán có hàng triệu ứng viên xuống quy mô con người có thể xem xét[1].

Những việc máy móc chưa làm được cũng rõ ràng. Tiếng Minoan là ngôn ngữ nào thì vẫn chưa được làm sáng tỏ. Văn bản có thể đóng vai trò như Phiến đá Rosetta hoặc vẫn còn nằm dưới đất, hoặc không tồn tại[3]. Dù chuyển âm ra chữ viết, ta vẫn chưa biết từ đó chỉ điều gì[6].

Trường hợp chữ Cypro-Minoan cho thấy rõ khác biệt này. Mạng nơ-ron đưa ra kết quả rằng cách phân loại cũ bị lung lay[10]. Nhưng như vậy không có nghĩa là nó đã đọc được chữ viết đó. Thứ máy móc sắp xếp được là cấu trúc của chữ, chứ không phải ý nghĩa của lời nói.

Con đường còn lại dường như có hai hướng. Một là hiện vật mới xuất hiện. Chỉ một văn bản viết bằng hai ngôn ngữ cũng có thể làm thay đổi cục diện. Hướng còn lại là tinh chỉnh dữ liệu kỹ hơn. Nếu ghi lại đến từng nét, nghiên cứu sau này có thể bắt đầu từ nền tảng đó[3].

Chữ tuyến tính A đến nay vẫn là chữ chưa đọc được. Trước nó, máy móc không mở cửa, mà chỉ vẽ ra hình dạng của ổ khóa. Việc mở cửa phụ thuộc vào hiện vật mới và phán đoán của con người[6]. Những tấm đất sét Aegea vẫn chưa tự nói lên lời của mình.

Tài liệu tham khảo

[1] Corazza, M., Ferrara, S., Montecchi, B., Tamburini, F., & Valério, M. (2021). The mathematical values of fraction signs in the Linear A script: A computational, statistical and typological approach. Journal of Archaeological Science, 125, 105214. https://doi.org/10.1016/j.jas.2020.105214

[2] Younger, J. G. (n.d.). Linear A texts and inscriptions in phonetic transcription. University of Kansas. http://people.ku.edu/~jyounger/LinearA/ (truy cập ngày 3 tháng 9 năm 2026)

[3] Braović, M., Krstinić, D., Štula, M., & Ivanda, A. (2024). A systematic review of computational approaches to deciphering Bronze Age Aegean and Cypriot scripts. Computational Linguistics, 50(2), 725-779. https://doi.org/10.1162/coli_a_00514

[4] GreekReporter. (2026, 7월 30일). AI engineer claims to have deciphered Linear A. greekreporter.com. https://greekreporter.com/2026/07/30/ai-engineer-claim-decipher-linear-a/ (truy cập ngày 3 tháng 9 năm 2026)

[5] GreekReporter. (2026, 8월 13일). AI engineer claims to have deciphered Linear A, identifies Minoan language as Semitic. greekreporter.com. https://greekreporter.com/2026/08/13/ai-engineer-claim-linear-a-minoans-semitic-language/ (truy cập ngày 3 tháng 9 năm 2026)

[6] Adkins, J. (2026, 7월 27일). Cracking the code: can AI help us decipher ancient languages? The Conversation. https://theconversation.com/cracking-the-code-can-ai-help-us-decipher-ancient-languages-288238 (truy cập ngày 3 tháng 9 năm 2026)

[7] GreekReporter. (2026, 8월 14일). Linear A inscription points to Minoan origins of Greek gods Dionysos and Demeter. greekreporter.com. https://greekreporter.com/2026/08/14/minoan-linear-a-inscription-greek-gods-dionysos-demeter/ (truy cập ngày 3 tháng 9 năm 2026)

[8] Del Freo, M. (2022). Linear A. Mnamon: Ancient writing systems in the Mediterranean. Scuola Normale Superiore. https://doi.org/10.25429/sns.it/lettere/mnamon022 (truy cập ngày 3 tháng 9 năm 2026)

[9] Salgarella, E., & Castellan, S. (2021). SigLA: The Signs of Linear A. A palaeographical database. In Grapholinguistics in the 21st Century 2020: Proceedings (Grapholinguistics and Its Applications, Vol. 5, pp. 945-962). Fluxus Editions. https://sigla.phis.me/ (truy cập ngày 3 tháng 9 năm 2026)

[10] Corazza, M., Tamburini, F., Valerio, M., & Ferrara, S. (2022). Unsupervised deep learning supports reclassification of Bronze age cypriot writing system. PLOS ONE, 17(7), e0269544. https://doi.org/10.1371/journal.pone.0269544

Chương 10. Hack mật mã chu kỳ mặt trời và con dấu động vật: âm thanh của chữ viết Ấn Hà và rongorongo

10.1. Hồ sơ vụ án

Bốn ký hiệu khắc trên con dấu

Văn minh Ấn Hà kéo dài từ khoảng năm 2600 TCN đến khoảng năm 1900 TCN[1]. Địa bàn của nó là Pakistan và tây bắc Ấn Độ ngày nay. Nếu nhìn vào quy mô và số lượng đô thị, đây là một trong những nền văn minh tiêu biểu của thời đại đồ đồng. Thời đại đồ đồng là thời xưa khi con người dùng đồng nhiều hơn sắt. Vùng đất có đô thị và làng mạc phân bố được cho là rộng hơn 1 triệu km2[2].

Trong thập niên 1920, các cuộc khai quật lớn bắt đầu tại Harappa và Mohenjo-daro[1]. Người phụ trách toàn bộ công việc này là nhà khảo cổ học người Anh John Marshall. Ông là người đứng đầu Cơ quan Khảo cổ Ấn Độ. Khu Harappa do Daya Ram Sahni khai quật vào năm 1921. Khu Mohenjo-daro do Rakhaldas Banerji phụ trách vào năm 1922. Marshall đã tập hợp các cuộc khai quật này và công bố với thế giới sự tồn tại của một nền văn minh mới[3].

Từ lòng đất xuất hiện những con đường gạch thẳng và hệ thống thoát nước. Rất nhiều con dấu vuông nhỏ hơn lòng bàn tay cũng được tìm thấy[1].

Chỉ nhìn hình dáng đô thị cũng có thể biết nhiều điều. Đường phố chạy thẳng như bàn cờ. Cống thoát nước nối đến từng nhà. Gạch được nung theo kích thước thống nhất. Quả cân dùng để cân đo cũng có chuẩn. Điều đó có nghĩa đây là một xã hội biết lập kế hoạch và tuân thủ quy tắc. Nhưng chữ viết họ để lại thì chưa đọc được.

Con dấu được làm bằng một loại đá mềm gọi là steatit. Steatit là loại đá có thể bị móng tay để lại vết. Ở phần dưới con dấu có khắc hình động vật. Có hình con thú một sừng, bò có u, và voi. Trên khoảng trống phía trên hình vẽ có một hàng ký hiệu[1].

Con dấu là vật dùng để ấn in. Khi ấn lên đất sét mềm, hình vẽ và ký hiệu được chuyển nguyên sang đó. Nếu đắp đất sét lên dây buộc hàng rồi đóng dấu, nó trở thành niêm phong. Trong các đô thị Ấn Hà, người ta cũng tìm thấy các niêm phong bằng đất sét có dấu in[1].

Các học giả gọi nhóm ký hiệu này là chữ viết Ấn Hà hoặc chữ viết Harappa[1]. Tên gọi có chữ viết, nhưng đến nay vẫn chưa đọc được. Không biết âm thanh. Cũng không biết ý nghĩa. Tình trạng đó đã kéo dài hơn một trăm năm từ khi phát hiện[2].

Nói chưa đọc được có hai nghĩa chồng lên nhau. Một là không biết âm thanh của ký hiệu. Hai là không biết ý nghĩa của ký hiệu. Hiện nay cả hai đều chưa biết.

Vấn đề là độ dài. Trên một con dấu thường chỉ có bốn hoặc năm ký hiệu[1]. Ngay cả bản khắc dài trên một mặt cũng không vượt quá mười bảy ký tự. Phải đếm cả bản khắc chia trên ba mặt thì mới được hai mươi sáu ký tự[4]. Đó là lượng chữ còn không đủ lấp đầy một dòng vở trong lớp học.

Những dòng chữ khắc trên con dấu hoặc đồ gốm như vậy được gọi là minh văn. Số minh văn Ấn Hà đã thu thập đến nay vào khoảng 3.000 hiện vật[5]. Số lượng không ít, nhưng từng văn bản lại quá ngắn. Chỉ gom các văn bản ngắn thì khó tìm ra cấu trúc câu. Vì không có cơ hội nhìn thấy chủ ngữ và vị ngữ được đặt ra sao.

Ký hiệu trên tấm gỗ bên kia biển

Giữa Nam Thái Bình Dương có một hòn đảo tên là Rapa Nui. Với chúng ta, tên gọi đảo Phục Sinh quen thuộc hơn. Nó cách vùng đất có người ở gần nhất hơn 2.000 km. Nơi này cũng nổi tiếng với các tượng đá lớn moai. Trên đảo này, người ta cũng tìm thấy các tấm gỗ có khắc những ký hiệu chưa rõ nghĩa. Những ghi chép này ngày nay được gọi là rongorongo[6].

Các ký hiệu không phải chữ viết bằng bút lông. Chúng được khắc vào gỗ bằng mảnh obsidian sắc hoặc công cụ làm từ xương. Hình khắc trông giống người, chim và cá[6]. Hình người giơ tay chân xuất hiện thường xuyên.

Hướng đọc cũng xa lạ. Người đọc bắt đầu từ đầu bên trái ở phía dưới tấm bảng. Sau khi đọc xong một dòng, người đọc xoay tấm bảng 180 độ. Ở trạng thái bị đảo ngược trên dưới, người đọc đọc dòng tiếp theo. Cách lật bảng theo từng dòng này được gọi là kiểu viết ngược chiều bò cày. Nói dễ hiểu là cách đọc bằng cách lật ngược. Ở Hy Lạp cổ đại cũng từng có loại chữ đổi hướng theo từng dòng. Trường hợp đó không xoay tấm bảng, mà chỉ đổi hướng chữ.

Cách này thử thách nhà nghiên cứu thêm một lần nữa. Phải xác định dòng nào là dòng đầu thì mới sắp được thứ tự[6]. Nếu xác định sai thứ tự, toàn bộ chuỗi nối tiếp của ký hiệu sẽ bị đảo lộn. Đây là phần con người phải phán đoán trước khi đưa dữ liệu vào máy tính.

Số di vật còn lại không nhiều. Số hiện vật thật nằm rải rác trong các bảo tàng và bộ sưu tập tư nhân trên thế giới là 26. Tất cả đều ở ngoài đảo Rapa Nui[6]. Dù cộng toàn bộ ký hiệu khắc trên đó, số lượng cũng chỉ khoảng 15.000 ký tự[7]. Ít hơn cả một cuốn sách.

Có lý do khiến di vật ít như vậy. Vào thế kỷ 19, dân số trên đảo giảm mạnh, và nhiều tấm gỗ cũng biến mất. Những gì còn lại bị phân tán đến các bảo tàng và tu viện ở châu Âu. Trong số đó, bốn tấm được lưu giữ tại một dòng tu ở Roma[8].

Thời điểm các tấm gỗ này được làm ra cũng là vấn đề gây tranh luận lâu dài. Đã có nghi ngờ rằng chúng được làm giả theo kiểu bắt chước sau khi người châu Âu đến đảo. Năm 2024, nhóm nghiên cứu của Silvia Ferrara đã đo niên đại bốn tấm gỗ ở Roma[8]. Họ dùng phương pháp đo lượng carbon còn lại trong gỗ để xác định tuổi. Phương pháp này được gọi là định tuổi bằng carbon phóng xạ.

Trong bốn tấm, ba tấm là gỗ thế kỷ 19. Tấm còn lại cho kết quả là gỗ cổ hơn nhiều. Nó có trước hơn 200 năm so với thập niên 1720, khi tàu châu Âu đến đảo[8]. Khả năng người Rapa Nui tự tạo ra chữ viết đã tăng lên.

Nhóm nghiên cứu đặt thêm một điều kiện. Thời điểm chặt cây không giống thời điểm khắc ký hiệu. Tuổi của gỗ chỉ cho biết việc khắc diễn ra sau đó. Vì có thể người ta đã dùng gỗ cũ để khắc vào thời gian muộn hơn. Một điểm khác còn lại là chỉ có một tấm gỗ cho niên đại sớm[8].

Đây có thật là chữ viết không

Giới học thuật đã tranh luận lâu dài về hai hệ thống ký hiệu này. Cuộc tranh luận không bắt đầu từ việc giải mã. Trước hết họ tranh luận xem đây có đúng là chữ dùng để ghi lời nói hay không[1][9].

Năm 2004, Steve Farmer, Richard Sproat và Michael Witzel cùng công bố một bài báo. Họ cho rằng ký hiệu Ấn Hà không phải là chữ viết. Căn cứ họ nêu là các ghi chép quá ngắn và không có một văn bản dài nào. Nghĩa là các ký hiệu này gần với dấu hiệu của dòng họ, dấu sở hữu đồ vật, hoặc bùa chú hơn[9].

Phía phản đối đề nghị tìm câu trả lời bằng thống kê. Họ muốn đếm tần suất xuất hiện của ký hiệu và quy tắc gắn trước sau. Lời nói của con người có thói quen thống kê giống như thói quen nói năng. Cách tiếp cận này là đo xem thói quen ấy có trong ký hiệu hay không.

Ưu điểm của phương pháp này là có thể dùng dù chưa biết nghĩa. Chỉ cần đổi ký hiệu thành số và đếm thứ tự. Đây cũng là việc máy tính làm tốt.

Ba bức tường chặn đường giải mã

Cả hai hệ thống ký hiệu đều chưa được giải mã. Có ba bức tường chắn phía trước.

Không có văn bản để đối chiếu. Chữ tượng hình Ai Cập được giải nhờ đá Rosetta. Trên đá Rosetta, cùng một nội dung cũng được viết bằng tiếng Hy Lạp. Học giả Pháp Champollion đặt chữ đã biết cạnh chữ chưa biết và ghép cặp chúng. Văn bản viết song song cùng một nội dung bằng hai ngôn ngữ được gọi là bản song ngữ. Ở di tích Ấn Hà và đảo Rapa Nui, loại ghi chép như vậy chưa xuất hiện[2][6].

Cũng không biết các ký hiệu ghi lại ngôn ngữ nào. Ngôn ngữ thực tế mà ký hiệu dùng để ghi lại được gọi là ngôn ngữ nền. Về ngôn ngữ nền của chữ viết Ấn Hà, giả thuyết thuộc hệ Dravidian và giả thuyết thuộc hệ Indo-Aryan đối lập nhau. Cũng có ý kiến cho rằng nó không thuộc bất kỳ ngữ hệ nào đã biết[2].

Bức tường còn lại là lượng dữ liệu. AI ngày nay lớn lên nhờ ăn khối lượng văn bản khổng lồ. Nhưng ở đây chỉ có vài nghìn con dấu Ấn Hà và 26 tấm gỗ[1][6]. Khi thiếu dữ liệu, AI tạo ra những lời nói dối có vẻ hợp lý. Hiện tượng này được gọi là ảo giác.

Ba bức tường này níu giữ lẫn nhau. Nếu biết ngôn ngữ nền, ta có thể thử khớp giá trị âm thanh. Nếu có bản song ngữ, ta có thể tìm ra ngôn ngữ nền. Nếu có nhiều dữ liệu, ta có thể nắm manh mối bằng thống kê. Vì một điểm chưa được giải, những điểm còn lại vẫn đứng nguyên tại chỗ.

Ngày 5 tháng 1 năm 2025, bang Tamil Nadu của Ấn Độ treo giải thưởng. Đây là thông báo được đưa ra tại lễ khai mạc hội thảo quốc tế kỷ niệm 100 năm phát hiện văn minh Ấn Hà, tổ chức ở Chennai[10]. Nội dung là trao 1 triệu đô la Mỹ cho người giải được chữ viết Ấn Hà đến mức được các chuyên gia khảo cổ công nhận[10][11]. Tính đến tháng 9 năm 2026, chưa có người đoạt giải nào được biết đến.

Tài liệu tham khảo

[1] Rao, R. P. N., Yadav, N., Vahia, M. N., Joglekar, H., Adhikari, R., & Mahadevan, I. (2009). Entropic evidence for linguistic structure in the Indus script. Science, 324(5931), 1165. https://doi.org/10.1126/science.1170391

[2] Khan, M. A., Khurshid, S. K., & Aslam, M. (2025). Modern algorithms for ancient scripts: A review of AI-based techniques in Indus civilization research. Cultural and Historical Heritage: Preservation, Presentation, Digitalization, 11(2), 11-21. https://doi.org/10.55630/KINJ.2025.110201

[3] Grand Valley State University. (n.d.). Archaeologist of the month: John Marshall. gvsu.edu. https://www.gvsu.edu/archaeology/module-spotlight-view.htm?entryId=A748C17B-A353-F70E-D7422D41F7889433 (truy cập ngày 3 tháng 9 năm 2026)

[4] Farmer, S. (n.d.). Claims concerning the longest Indus "inscription". safarmer.com. https://safarmer.com/indus-longestinscription/ (truy cập ngày 3 tháng 9 năm 2026)

[5] Yadav, N., Joglekar, H., Rao, R. P. N., Vahia, M. N., Adhikari, R., & Mahadevan, I. (2010). Statistical analysis of the Indus script using n-grams. PLoS ONE, 5(3), e9506. https://doi.org/10.1371/journal.pone.0009506

[6] Australian Museum. (2026, 5월 28일). Kohau Rongorongo: Talking tablet. australian.museum. https://australian.museum/learn/cultures/pasifika-collections/rapa-nui-collections/kohau-rongorongo-talking-tablet/ (truy cập ngày 3 tháng 9 năm 2026)

[7] Ross, L. (2026, 3월 31일). Reading Rongorongo (Version 2). Zenodo. https://doi.org/10.5281/zenodo.19362491 (truy cập ngày 3 tháng 9 năm 2026)

[8] Ferrara, S., Tassoni, L., Kromer, B., Wacker, L., Friedrich, M., Tonini, F., Lastilla, L., Ravanelli, R., & Talamo, S. (2024). The invention of writing on Rapa Nui (Easter Island): New radiocarbon dates on the Rongorongo script. Scientific Reports, 14, 2794. https://doi.org/10.1038/s41598-024-53063-7

[9] Farmer, S., Sproat, R., & Witzel, M. (2004). The collapse of the Indus-script thesis: The myth of a literate Harappan civilization. Electronic Journal of Vedic Studies, 11(2), 19-57. https://doi.org/10.11588/ejvs.2004.2.620

[10] The Tribune. (2025, 1월 6일). Stalin announces $1 mn prize for deciphering Indus Valley script. tribuneindia.com. https://www.tribuneindia.com/news/india/stalin-announces-1-mn-prize-for-deciphering-indus-valley-script/ (truy cập ngày 3 tháng 9 năm 2026)

[11] Smithsonian Magazine. (2025, 1월 30일). Officials are offering $1 million to anyone who can decode this ancient script. smithsonianmag.com. https://www.smithsonianmag.com/smart-news/officials-are-offering-1-million-to-anyone-who-can-decode-this-ancient-script-180985922/ (truy cập ngày 3 tháng 9 năm 2026)

10.2. Cuộc truy tìm của đội điều tra khoa học: máy đo entropy thông tin

Thước đo dấu vân tay của chữ viết

Con dấu Ấn Hà và tấm gỗ Rapa Nui chưa từng gặp nhau. Nhưng các nhà nghiên cứu ở hai nơi đã cầm lên cùng một công cụ. Vì những việc có thể làm trước các chuỗi ký hiệu chưa biết nghĩa là tương tự nhau.

Trước những chữ chưa đọc được, các nhà nghiên cứu chọn một câu hỏi khác. Thay vì hỏi chúng có nghĩa gì, họ hỏi chúng được sắp xếp ra sao. Vì dù chưa biết nghĩa, ta vẫn có thể đếm thứ tự.

Công cụ dùng lúc này là entropy thông tin. Entropy là giá trị đo mức độ khó biết điều gì sẽ đến tiếp theo. Giá trị lớn nghĩa là khó dự đoán yếu tố tiếp theo. Giá trị nhỏ nghĩa là yếu tố tiếp theo gần như đã được định sẵn.

Khái niệm này xuất phát từ kỹ thuật truyền thông. Nó là thước đo được tạo ra để đo lượng thông tin được chở trong tín hiệu khi gửi qua đường dây điện thoại. Nó tính toán chỉ bằng thứ tự, không liên quan đến nghĩa. Đây là lý do có thể áp dụng nó cho chữ viết chưa đọc được.

Có thể hiểu dễ hơn nếu nghĩ đến chỗ ngồi trong lớp học. Nếu ngồi theo bốc thăm, ta không biết ai sẽ ngồi bên cạnh. Khi đó entropy cao. Nếu ngồi theo thứ tự số, người ngồi bên cạnh được xác định. Khi đó entropy thấp.

Lời nói của con người nằm giữa hai cực đó. Ngữ pháp ràng buộc thứ tự, nhưng điều muốn nói được chọn khá tự do. Vì vậy entropy của ngôn ngữ con người nằm ở vùng giữa. Lấy vùng giữa này làm thước đo là điểm xuất phát của nghiên cứu.

Ta cũng có thể kiểm tra bằng tiếng Hàn. Sau hakgyo-e có thể có nhiều từ. Ganda, gatda, danyeowatda đều có thể. Nhưng không phải từ nào cũng đứng được. Trạng thái có lựa chọn nhưng đã bị thu hẹp này là đặc điểm của ngôn ngữ.

Đo ký hiệu Ấn Hà

Năm 2009, nhóm nghiên cứu gồm Rajesh Rao và Nisha Yadav đã áp dụng thước đo này cho ký hiệu Ấn Hà. Họ tính độ bất định của ký hiệu tiếp theo khi đã biết ký hiệu phía trước[1]. Giá trị đo mức độ có thể dự đoán ký hiệu tiếp theo khi biết ký hiệu phía trước này được gọi là entropy có điều kiện.

Họ cũng chuẩn bị đối tượng so sánh. Một bên là ngôn ngữ của con người như tiếng Anh, tiếng Phạn và tiếng Sumer. Bên kia là những thứ không phải ngôn ngữ con người. Đó là chuỗi gene của sinh vật, mã chương trình máy tính và chuỗi ký hiệu tạo ngẫu nhiên[1].

Kết quả là ký hiệu Ấn Hà nằm gần nhóm ngôn ngữ con người. Nó có quy tắc rõ hơn chuỗi ký hiệu ngẫu nhiên và tự do hơn mã máy tính. Nhóm nghiên cứu đăng kết quả này trên Science[1].

Cùng nhóm nghiên cứu cũng đưa ra một mô hình xác suất. Đó là cách lập bảng xác suất ký hiệu tiếp theo xuất hiện khi ký hiệu trước đã được xác định. Mô hình chỉ nhìn ký hiệu trước để tính ký hiệu tiếp theo như vậy được gọi là mô hình Markov. Khi nhìn vào bảng, ta thấy ký hiệu nào thường xuất hiện ở đầu văn bản. Ta cũng thấy ký hiệu nào thường xuất hiện ở cuối văn bản[2].

Năm sau, cùng nhóm nghiên cứu khảo sát cả tần suất ký hiệu. Họ tập hợp khoảng 3.000 minh văn và chọn 1.548 hiện vật có tình trạng tốt. Một số ít ký hiệu chiếm phần lớn lượng sử dụng, còn các ký hiệu còn lại xuất hiện hiếm. Hình thái này là phân bố luôn thấy trong ngôn ngữ con người[3].

Phân bố này có tên riêng. Nó được gọi là định luật Zipf. Đây là quy luật nói rằng một vài từ thường dùng lấp đầy phần lớn văn bản. Trong tiếng Hàn, những từ như eun, neun, i, ga cũng chiếm hàng đầu. Ngược lại, có rất nhiều từ chỉ xuất hiện một lần rồi thôi. Ký hiệu Ấn Hà cũng cho thấy cùng hình thái đó[3].

Quy tắc cũng xuất hiện trong thứ tự đặt ký hiệu. 23 ký hiệu chiếm 80 phần trăm vị trí cuối văn bản. Ngược lại, để lấp đầy 80 phần trăm vị trí đầu văn bản, cần 82 ký hiệu[3]. Điều đó có nghĩa các ký hiệu đứng cuối bị giới hạn hơn nhiều. Nó giống hiện tượng trợ từ hoặc vĩ tố dồn về cuối câu trong tiếng Hàn.

Nhóm nghiên cứu cũng tạo mô hình đoán ký hiệu tiếp theo từ ký hiệu phía trước[2]. Mô hình này khôi phục các ký hiệu bị mòn và khó đọc với xác suất khoảng 75 phần trăm[3]. Đây là kết quả chỉ có thể xuất hiện khi có quy tắc.

Không nên hiểu lầm từ khôi phục. Điều mô hình điền vào chỉ là hình dạng ký hiệu. Ký hiệu đó từng có âm thanh gì thì vẫn chưa biết. Nó chỉ điền một ô còn thiếu của mảnh ghép, chứ không đọc bức tranh.

Phản bác và phản bác lại

Mô hình này cũng có giới hạn. Vì nó đoán ký hiệu tiếp theo chỉ bằng một hoặc hai ký hiệu phía trước[2]. Câu của con người có sự hô ứng giữa những từ cách xa nhau hơn nhiều. Nếu dữ liệu ngắn, mô hình không thể học sự hô ứng xa đó.

Kết quả này lập tức bị phản bác. Richard Sproat công bố bài phê bình vào năm 2010. Ông chỉ ra rằng chỉ riêng việc giá trị entropy giống nhau không đủ để khẳng định đó là chữ viết. Theo ông, các hệ thống ký hiệu không phải ngôn ngữ con người cũng có thể tạo ra giá trị tương tự[4].

Cùng năm đó, nhóm nghiên cứu của Rao đã đưa ra câu trả lời trên cùng tạp chí học thuật. Họ nói rằng họ không chứng minh các ký hiệu Ấn Hà là chữ viết. Họ chỉ giải thích rằng khả năng đó được thống kê hỗ trợ[5].

Năm 2014, Sproat đẩy phê phán đi xa hơn một bước. Ông gom nhiều bộ tư liệu ký hiệu không phải lời nói của con người. Đó là dấu hiệu thần linh của Mesopotamia, cột totem ở Bắc Mỹ, và hoa văn ngôi sao vẽ trên tường nhà kho. Ông cũng đưa vào các hình dùng trong dự báo thời tiết. Ông so sánh các tư liệu này với ký hiệu Ấn Hà bằng cùng một thước đo[6].

Quy trình phân loại do ông tạo ra xếp ký hiệu Ấn Hà vào phía không phải lời nói của con người. Sproat viết rằng chỉ thống kê là chưa đủ. Ông cho rằng phải có giải mã thành công, hoặc phải có chứng cứ khảo cổ riêng cho thấy nền văn hóa đó đã dùng chữ viết[6]. Cuộc tranh luận này vẫn chưa kết thúc.

Một phân tích mới công bố vào tháng 4 năm 2026 cũng đứng trước cùng câu hỏi. Ashish Nair thiết kế lại phép kiểm tra với 1.916 minh văn Ấn Hà. Ông dùng máy tính tạo ra tư liệu giả mô phỏng hệ thống ký hiệu không phải lời nói của con người. Ông đặt thống kê của tư liệu Ấn Hà thật cạnh thống kê của tư liệu giả rồi so sánh[7].

Tư liệu Ấn Hà không khớp hẳn với bất kỳ tư liệu giả nào. Nó nằm ở giữa hai tư liệu giả. Không có hệ thống ký hiệu phi ngôn ngữ nào đang tồn tại có thể tái hiện nguyên dạng thống kê của Ấn Hà. Nair viết rằng cần tạm hoãn phán đoán đây có phải chữ viết hay không. Bài viết này là bản công bố trước được đưa lên mạng trước khi qua bình duyệt của tạp chí học thuật[7].

Con mắt nhận ra ký hiệu

Có một việc phải làm trước thống kê. Đó là chuyển những vết mờ khắc trên đá thành ký hiệu để chép lại. Công việc này trong thời gian dài đã dựa vào mắt và tay của con người.

Việc chép lại khó hơn ta tưởng. Con dấu đã bị mài mòn trong 4.000 năm. Dù là cùng một ký hiệu, hình dạng vẫn khác nhau một chút tùy người khắc. Đôi khi cũng có ý kiến khác nhau về việc nên xem là hai ký hiệu hay một ký hiệu. Khi phán đoán này thay đổi, thống kê theo sau cũng thay đổi.

Năm 2017, Satish Palaniappan và Ronojoy Adhikari giao việc này cho máy tính. Họ tạo ra một quy trình xử lý. Khi đưa ảnh con dấu vào, hệ thống xuất ra danh sách ký hiệu[8].

Năm 2025, nhóm nghiên cứu của Vaishnavi Dixit kế thừa quy trình này. Họ tạo riêng một mạng nơ-ron nhận ra ký hiệu trong ảnh con dấu và một mạng nơ-ron nhận ra hình động vật. Họ chia tư liệu thành năm phần rồi lần lượt kiểm tra. Cả hai mạng nơ-ron đều cho độ chính xác trung bình khoảng 85 phần trăm[9].

Muốn giảm vấn đề này, cần công khai tiêu chuẩn phán đoán và so sánh với nhau. Năm 2025, nhóm nghiên cứu của Muzaffar Ali Khan đã hệ thống hóa công việc đó. Họ gom các nghiên cứu AI liên quan đến chữ viết Ấn Hà từ năm 2009 đến năm 2025 và chia theo loại. Họ cũng ghi lại những lỗi thường lặp lại[10].

Năm 2025, Atul Sharma và Subhajit Roy Chowdhury công bố một công cụ. Tên của nó là AI-EPIGRAPHY. Khi người dùng nhập chuỗi ký hiệu, công cụ trả về các ứng viên diễn giải có xác suất cao. Đây là công cụ kết hợp phân tích tần suất và tính xác suất chuỗi[11].

Vào tháng 5 năm 2026, một nỗ lực xây lại chính nền tảng dữ liệu được công bố. Nhóm nghiên cứu của Beata Megyesi tạo ra một cơ sở dữ liệu tập hợp các chữ viết hiếm và chưa đọc được. Ảnh độ phân giải cao, ký hiệu chép lại, chú giải và thông tin nguồn được đưa vào một khuôn chung. Họ cho rằng việc gom tư liệu rải rác là bài toán cần làm trước giải mã[12].

Tài liệu tham khảo

[1] Rao, R. P. N., Yadav, N., Vahia, M. N., Joglekar, H., Adhikari, R., & Mahadevan, I. (2009). Entropic evidence for linguistic structure in the Indus script. Science, 324(5931), 1165. https://doi.org/10.1126/science.1170391

[2] Rao, R. P. N., Yadav, N., Vahia, M. N., Joglekar, H., Adhikari, R., & Mahadevan, I. (2009). A Markov model of the Indus script. Proceedings of the National Academy of Sciences, 106(33), 13685-13690. https://doi.org/10.1073/pnas.0906237106

[3] Yadav, N., Joglekar, H., Rao, R. P. N., Vahia, M. N., Adhikari, R., & Mahadevan, I. (2010). Statistical analysis of the Indus script using n-grams. PLoS ONE, 5(3), e9506. https://doi.org/10.1371/journal.pone.0009506

[4] Sproat, R. (2010). Ancient symbols, computational linguistics, and the reviewing practices of the general science journals. Computational Linguistics, 36(3), 585-594. https://doi.org/10.1162/coli_a_00011

[5] Rao, R. P. N., Yadav, N., Vahia, M. N., Joglekar, H., Adhikari, R., & Mahadevan, I. (2010). Entropy, the Indus script, and language: A reply to R. Sproat. Computational Linguistics, 36(4), 795-805. https://doi.org/10.1162/coli_c_00030

[6] Sproat, R. (2014). A statistical comparison of written language and nonlinguistic symbol systems. Language, 90(2), 457-481. https://doi.org/10.1353/lan.2014.0031

[7] Nair, A. (2026, 4월 20일). How non-linguistic is the Indus sign system? A synthetic-baseline scorecard. arXiv:2604.17828. https://arxiv.org/abs/2604.17828 (truy cập ngày 3 tháng 9 năm 2026)

[8] Palaniappan, S., & Adhikari, R. (2017). Deep learning the Indus script. arXiv:1702.00523. https://arxiv.org/abs/1702.00523

[9] Dixit, V., Hussain, N., Basak, S., Atturu, D., Mitra, D., & Bhattacharya, U. (2025). Deep learning in archiving Indus script and motif information. Journal of Computer Applications in Archaeology, 8(1), 156-169. https://doi.org/10.5334/jcaa.175

[10] Khan, M. A., Khurshid, S. K., & Aslam, M. (2025). Modern algorithms for ancient scripts: A review of AI-based techniques in Indus civilization research. Cultural and Historical Heritage: Preservation, Presentation, Digitalization, 11(2), 11-21. https://doi.org/10.55630/KINJ.2025.110201

[11] Sharma, A., & Roy Chowdhury, S. (2025). AI-EPIGRAPHY: An interactive tool for computational decipherment of the Indus Valley script. Proceedings of IndiaHCI '25. https://doi.org/10.1145/3768633.3770145

[12] Megyesi, B., Rattenborg, R., Lang, B., Waldispuhl, M., & Heder, M. (2026). Building a corpus and database for rare and undeciphered scripts. Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026) @ LREC 2026, 184-196. ELRA Language Resources Association. https://aclanthology.org/2026.lt4hala-1.17/ (truy cập ngày 3 tháng 9 năm 2026)

10.3. Sự thật đã được làm sáng tỏ

Tấm gỗ đếm trăng

Trong số các tấm gỗ của Rapa Nui có một tấm mang tên Mamari. Trên tấm gỗ này có một đoạn có hình dạng khác với những chỗ khác. Các ký hiệu hình trăng non nối tiếp nhau thành hàng[1].

Học giả Đức Thomas Barthel chú ý đến đoạn này vào giữa thế kỷ 20. Ông xem đây là ghi chép về ngày tháng theo trăng[1]. Năm 1990, Jacques Guy sắp xếp lại tư liệu và đưa ra cùng kết luận[2].

Năm 2011, Paul Horley củng cố thêm cách giải thích này. Lịch trên tấm gỗ Mamari gồm ba mươi đêm. Con số này khớp với chu kỳ thực của Mặt Trăng. Nó cũng khớp với lịch của nhiều đảo Polynesia[1].

Horley cũng đặt lịch của Tahiti và New Zealand cạnh nhau để so sánh. Tên gọi gắn với từng đêm ở các đảo khá giống nhau. Lịch của Rapa Nui lệch khoảng ba ngày so với các đảo khác. Có vẻ đó là khác biệt hình thành trong thời gian hòn đảo ở xa các nơi khác[1].

Cách sắp xếp ký hiệu cũng theo quy tắc. Giữa các ký hiệu trăng non có những ký hiệu khác chen vào, chia thành tám nhóm. Hai ký hiệu trăng non cuối cùng chỉ những đêm không thấy trăng[1].

Cách sắp xếp tương tự cũng xuất hiện ngoài tấm gỗ. Trong các tranh khắc đá trên đảo, có hình khắc hoa văn rùa cùng với trăng non. Ở đó, hai trăng non cuối cùng cũng nằm trong mai rùa. Cách này giống cách tấm gỗ tách riêng những đêm không trăng để vẽ[1]. Tấm gỗ và tảng đá dùng chung một cách đếm.

Cách giải thích này được thừa nhận rộng rãi trong nghiên cứu rongorongo. Dù vậy, điều đã đọc ra là thứ tự ngày tháng. Ta không biết người Rapa Nui gọi những đêm đó là gì.

Câu trả lời do AI đưa ra và sức nặng của nó

Việc đếm trăng và việc đếm năm gắn với nhau. Nhật thực xảy ra khi Mặt Trời và Mặt Trăng chồng lên nhau. Vì vậy, tuyên bố đã đọc được lịch nhanh chóng lan thành tuyên bố đã đọc được toàn bộ bầu trời.

Vào tháng 3 năm 2026, Logan Ross công bố một bài tiền ấn phẩm. Ông chạy mô hình ngôn ngữ trên máy tính cá nhân để phân tích rongorongo. Ông tạo một từ điển gồm 89 ký hiệu và tuyên bố nó bao phủ gần 97 phần trăm phần ghi chép còn lại. Ông cũng cho rằng tấm gỗ Mamari là bảng dự báo nhật thực và nguyệt thực[3].

Tuyên bố này chưa qua kiểm chứng của giới học thuật. Bài viết của Ross không phải là luận văn đã vượt qua bình duyệt của tạp chí học thuật. Đây là bản công bố trước được đưa lên kho lưu trữ trên mạng[3]. Ai cũng có thể đưa bản công bố trước lên. Nội dung có đúng hay không phải được các nhà nghiên cứu khác kiểm tra tiếp.

Khi xử lý những tuyên bố như vậy, cần xem cả phương pháp kiểm tra. Các ghi chép rongorongo còn lại gồm 26 hiện vật, khoảng 15.000 ký tự[3]. Ở nơi chỉ có chừng ấy tư liệu, rất dễ xuất hiện kết quả tình cờ khớp. Nếu tự do gán giá trị âm cho 89 ký hiệu, người ta có thể tạo ra bất kỳ câu nào. Vì vậy, giới học thuật trước hết hỏi liệu người khác có dùng cùng phương pháp để thu được cùng kết quả hay không.

Nếu là giải mã thật, có một phép thử phải vượt qua. Đó là lấy một tấm gỗ khác không dùng khi lập phương án giải mã và thử đọc. Ở đó cũng phải xuất hiện những câu có nghĩa. Chữ tuyến tính B ghi tiếng Hy Lạp Mycenae là ví dụ đã vượt qua phép thử đó[4]. rongorongo và chữ viết Ấn Hà vẫn chưa vượt qua ngưỡng này.

Việc máy làm được và không làm được

Đến nay, việc AI đã làm được với hai chữ viết này là rõ ràng.

Máy nhận ra ký hiệu từ các vết mờ[5]. Việc con người phải mất vài tháng để đối chiếu bằng mắt có thể hoàn tất trong vài giờ. Máy cũng khôi phục các ký hiệu bị mài mòn, khó đọc bằng ngữ cảnh trước sau[6]. Nó gom ảnh và ghi chép của các di vật rải rác vào một nơi để có thể tìm kiếm[7].

Việc gom tư liệu không dễ thấy. Nhưng nếu không có việc này, cũng không có các phép tính tiếp theo. Nếu độ sáng và góc chụp của ảnh khác nhau, không thể so sánh ký hiệu. Nếu quy tắc chép lại khác nhau tùy nhà nghiên cứu, thống kê cũng lệch. Cơ sở dữ liệu công bố năm 2026 là nỗ lực thống nhất tiêu chuẩn đó[7].

Phía thống kê cũng có thành quả. Nhiều nghiên cứu đã xác nhận rằng ký hiệu Ấn Hà không phải là chuỗi ngẫu nhiên[6][8]. Kết luận rằng cách sắp xếp có quy tắc đã được tái lập nhiều lần. Nhưng điều đó không có nghĩa đây là chữ viết. Nói có quy tắc khác với nói đã ghi lại lời nói.

Tuy nhiên, việc máy chưa làm được vẫn còn lớn hơn. Ta vẫn chưa biết giá trị âm của từng ký hiệu Ấn Hà[9]. Ta cũng chưa biết giá trị âm của ký hiệu rongorongo[3]. Cả hai chữ viết đều chưa có từ nào được xác định nghĩa. Những gì viết trên con dấu Ấn Hà là tên người hay tên đồ vật cũng chưa được định rõ[9].

Lý do nằm ở tính chất của tư liệu. Nếu có văn bản song ngữ, việc giải mã dễ hơn nhiều. Ở di tích Ấn Hà cũng như trên đảo Rapa Nui đều không có loại văn bản ấy[9]. Không phải không có trường hợp được giải mà không có văn bản song ngữ. Chữ tuyến tính B đã được giải như vậy. Nhưng khi đó có manh mối để đoán ngôn ngữ nền, và tư liệu cũng dồi dào[4]. Với chữ viết Ấn Hà, ngôn ngữ nền là gì cũng chưa được xác định[9]. Với rongorongo, phần chữ còn lại quá ít[3].

Phân tích tháng 4 năm 2026 một lần nữa cho thấy tình hình này bằng con số. 1.916 minh văn Ấn Hà không khớp hẳn với thống kê của lời nói con người, cũng không khớp hẳn với thống kê của ký hiệu không phải lời nói con người[10]. Điều đó có nghĩa câu hỏi kéo dài hơn một trăm năm vẫn còn bỏ ngỏ.

Đứng về phía chờ đợi

Nghiên cứu về chữ viết chưa đọc được đòi hỏi hai thái độ. Một là thái độ đẩy công cụ mới đi đến cùng. Hai là thái độ nghi ngờ kết quả đã có.

AI làm tốt việc thứ nhất. Nó đếm, so sánh và xếp ra các ứng viên. Việc thứ hai thuộc về con người. Phải phán đoán ứng viên nào có thể đã tồn tại trong thời đại đó. Phán đoán ấy cần tính đến vị trí di tích, đồ vật tìm thấy trong mộ, và truyền khẩu còn lại.

Cách giải thích lịch của Rapa Nui là một ví dụ. Đếm ký hiệu trăng non là tính toán. So sánh ba mươi đêm ấy với lịch của nhiều đảo Polynesia là khảo chứng[1]. Khi hai công việc ăn khớp với nhau, cách giải thích có sức thuyết phục.

Hình động vật trên con dấu Ấn Hà cũng là loại tư liệu như vậy. Ta không biết con thú một sừng có nghĩa gì. Ba mươi đêm trên tấm gỗ Mamari cũng vậy. Ta đã biết thứ tự, nhưng không biết người ta đã làm gì trong những đêm đó.

Nói rằng mình không biết cũng có giá trị. Phải vạch ranh giới giữa phần đã được làm sáng tỏ và phần bắt đầu là suy đoán thì nghiên cứu tiếp theo mới bắt đầu. Nếu làm mờ ranh giới, một câu chuyện có vẻ hợp

lý sẽ đóng vai sự thật. Trong nghiên cứu về chữ viết chưa đọc được, việc vạch ranh giới này quan trọng như việc đưa ra kết quả.

Hai chữ viết được bàn trong chương này chưa từng gặp nhau. Chúng cách nhau hơn 3.000 năm và nơi cư trú cũng ở hai phía đối diện của Trái Đất. Nhưng khó khăn mà các nhà nghiên cứu gặp phải gần như giống nhau. Tư liệu ngắn, không có văn bản song ngữ, và không biết ngôn ngữ nền. Đó là lý do cùng một công cụ được dùng ở cả hai nơi.

Giải thưởng 1 triệu đô la do bang Tamil Nadu treo ra cũng phản chiếu tình hình ấy[11]. Dù đã hơn một năm kể từ khi giải thưởng được đặt ra, vẫn chưa có người nhận. Máy tính đã nhanh hơn, nhưng chìa khóa vẫn chưa xuất hiện.

Tài liệu tham khảo

[1] Horley, P. (2011). Lunar calendar in rongorongo texts and rock art of Easter Island. Journal de la Societe des Oceanistes, 132, 17-38. https://doi.org/10.4000/jso.6314

[2] Guy, J. B. M. (1990). The lunar calendar of Tablet Mamari. Journal de la Societe des Oceanistes, 91(2), 135-149. https://doi.org/10.3406/jso.1990.2882

[3] Ross, L. (2026, 3월 31일). Reading Rongorongo (Version 2). Zenodo. https://doi.org/10.5281/zenodo.19362491 (truy cập ngày 3 tháng 9 năm 2026)

[4] Chadwick, J. (1958). The decipherment of Linear B. Cambridge University Press.

[5] Palaniappan, S., & Adhikari, R. (2017). Deep learning the Indus script. arXiv:1702.00523. https://arxiv.org/abs/1702.00523

[6] Yadav, N., Joglekar, H., Rao, R. P. N., Vahia, M. N., Adhikari, R., & Mahadevan, I. (2010). Statistical analysis of the Indus script using n-grams. PLoS ONE, 5(3), e9506. https://doi.org/10.1371/journal.pone.0009506

[7] Megyesi, B., Rattenborg, R., Lang, B., Waldispuhl, M., & Heder, M. (2026). Building a corpus and database for rare and undeciphered scripts. Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026) @ LREC 2026, 184-196. ELRA Language Resources Association. https://aclanthology.org/2026.lt4hala-1.17/ (truy cập ngày 3 tháng 9 năm 2026)

[8] Rao, R. P. N., Yadav, N., Vahia, M. N., Joglekar, H., Adhikari, R., & Mahadevan, I. (2009). Entropic evidence for linguistic structure in the Indus script. Science, 324(5931), 1165. https://doi.org/10.1126/science.1170391

[9] Khan, M. A., Khurshid, S. K., & Aslam, M. (2025). Modern algorithms for ancient scripts: A review of AI-based techniques in Indus civilization research. Cultural and Historical Heritage: Preservation, Presentation, Digitalization, 11(2), 11-21. https://doi.org/10.55630/KINJ.2025.110201

[10] Nair, A. (2026, 4월 20일). How non-linguistic is the Indus sign system? A synthetic-baseline scorecard. arXiv:2604.17828. https://arxiv.org/abs/2604.17828 (truy cập ngày 3 tháng 9 năm 2026)

[11] The Tribune. (2025, 1월 6일). Stalin announces $1 mn prize for deciphering Indus Valley script. tribuneindia.com. https://www.tribuneindia.com/news/india/stalin-announces-1-mn-prize-for-deciphering-indus-valley-script/ (truy cập ngày 3 tháng 9 năm 2026)

Chương 11. AI đọc lại những bia ký Hy Lạp bị vỡ: mô hình nền tảng và Ithaca

11.1. Hồ sơ vụ việc

Ghi chép quốc gia khắc trên đá

Người Hy Lạp cổ đại khắc những việc quan trọng lên đá. Khi ban hành luật mới, họ khắc nội dung lên cột đá và dựng ở quảng trường. Sổ sách thu thuế cũng được để lại trên đá. Danh mục đồ vật dâng cho đền thờ cũng được để lại trên đá. Những văn bản khắc trên đá hoặc đồng như vậy được gọi là bia ký. Ngành học thu thập và đọc bia ký là kim thạch học. Nghĩa là ngành nghiên cứu chữ viết khắc trên đá hoặc kim loại[4].

Bia ký phần lớn là ghi chép do người cùng thời trực tiếp để lại. Ít có dấu vết người đời sau chép lại rồi sửa đổi. Vì vậy, đó là tư liệu quý đối với nhà sử học. Tuy nhiên, không phải lúc nào chúng cũng còn nguyên dạng ban đầu. Có những bia đá mà về sau chữ bị xóa hoặc được khắc sửa. Cũng còn những trường hợp người ta mài mặt đá rồi viết lại. Bia ký cho biết Athens đã liên minh với thành bang nào và vào lúc nào[4].

Bia ký là văn bản được làm ra để nhiều người xem. Chúng được dựng ở quảng trường hoặc trước đền thờ để người đi ngang đọc. Công dụng của chúng giống thông báo dán ở hành lang trường học. Vì vậy, bia ký ghi rõ ràng tên quan chức và ngày tháng của hôm đó. Nhà sử học lấy những tên và ngày tháng này làm manh mối[4].

Chữ bị xóa, phần khuyết thiếu

Vấn đề là thời gian. Bia đá đã đứng trước mưa gió suốt hàng nghìn năm. Bề mặt đá mòn dần, còn rãnh chữ nông đi. Bia đá bị chôn dưới đất thì bị độ ẩm và vi sinh vật ăn mòn. Có bia đá bị vỡ, chỉ còn lại một phần[1].

Kết quả là trong bia ký xuất hiện những ô trống nơi chữ đã biến mất. Các học giả gọi những ô trống này là phần khuyết thiếu. Trong tiếng Latin, đó là lacuna[1]. Phần khuyết thiếu khoét rỗng giữa câu. Nếu tên người biến mất, ta không biết ai đã làm việc đó. Nếu con số biến mất, ta không biết đã nộp bao nhiêu. Nếu tên thành bang biến mất, ta không biết đó là cam kết với nước nào.

Từ lâu, các học giả đã cố lấp những ô trống này. Họ tìm các bia ký khác cùng thời kỳ và có nội dung tương tự để so sánh. Bia ký Hy Lạp có những khuôn câu cố định. Văn bản công bố luật thường sắp câu theo gần như cùng một trật tự. Vì vậy, nếu phần trước và phần sau còn lại, có thể đoán từ đã mất[1].

Bia đá nặng nên không thể mang về phòng nghiên cứu. Vì vậy, các học giả đặt giấy ướt lên bia đá rồi ép xuống. Khi giấy khô, rãnh chữ hiện lên nguyên dạng. Bản sao bằng giấy này gọi là bản dập. Các học giả mang bản dập về và đọc chữ tại bàn làm việc. Trong các tập danh mục in cũ, thay vì ảnh, chỉ có chữ được chép lại. Vì vậy, người đời sau phải đọc chữ mà không nhìn thấy hình dạng của đá[4].

Việc phỏng đoán có quy tắc. Trước hết phải đếm xem ô trống có bao nhiêu chữ. Có thể biết đại khái bằng cách đo khoảng cách giữa các chữ còn lại trên đá. Sau đó chỉ đưa vào danh sách ứng viên những từ có số chữ phù hợp. Chữ được điền như vậy phải luôn được ghi trong ngoặc vuông. Dưới những chữ có hình dạng hư hỏng và khó đọc, người ta đặt dấu chấm nhỏ. Quy ước này gọi là ký hiệu Leiden. Đây là quy tắc được tạo ra để phân biệt chữ còn lại trên đá với suy đoán của con người[1, 4].

Cuộc tranh luận quanh sigma ba nét

Muốn đọc bia ký, cũng phải biết chữ đó được khắc vào thời nào. Phương pháp cổ điển là nhìn hình dạng chữ để đo niên đại. Lý do là hình dạng chữ thay đổi từng chút theo thời đại[2].

Một ví dụ thường gặp trong bia ký Athens là sigma. Sigma là một chữ cái trong bảng chữ cái Hy Lạp. Ở thời kỳ sớm, người ta khắc sigma có ba nét. Về sau, sigma có bốn nét được dùng rộng rãi. Trong thời gian dài, các học giả lấy sigma ba nét làm tiêu chí. Nếu có chữ như vậy, họ xem đó là bia ký trước khoảng năm 446 trước Công nguyên[2].

Tuy nhiên, đã có học giả nghi ngờ tiêu chí này. Tốc độ thay đổi kiểu chữ khác nhau tùy từng vùng. Có vùng vẫn giữ kiểu chữ cũ trong thời gian dài. Ngay trong một thành bang, mỗi thợ khắc cũng có thói quen khắc khác nhau[2].

Vì vậy, tranh luận kéo dài về niên đại các bia ký hiệp ước giữa Athens và các thành bang khác. Một bên cho rằng chúng có trước khoảng năm 446 trước Công nguyên. Bên kia cho rằng đó là chữ được khắc vào thập niên 420 trước Công nguyên[2]. Thập niên 420 trước Công nguyên là lúc Chiến tranh Peloponnesus đang diễn ra ác liệt. Cuộc chiến này bắt đầu vào năm 431 trước Công nguyên. Nếu niên đại lệch đi vài chục năm, mạch truyện lịch sử cũng thay đổi. Bởi thời điểm Athens dùng sức mạnh để ép các thành bang đồng minh sẽ bị dời đi toàn bộ.

Việc lấp ô trống luôn có rủi ro. Từ mà một học giả điền vào có thể được thế hệ sau đọc như nguyên văn. Chỉ một bản in bỏ sót ngoặc vuông cũng có thể biến suy đoán thành sự thật[1]. Vì vậy, các nhà kim thạch học luôn để lại dấu hiệu cho suy đoán của mình.

Cũng phải tìm ra bia ký ban đầu được dựng ở đâu. Mỗi thành bang Hy Lạp dùng một phương ngữ khác nhau. Cùng một từ cũng có cách viết khác nhau tùy thành bang[2]. Các học giả dựa vào khác biệt chính tả và tên chức vụ để đoán nơi khai quật[4]. Nơi khai quật là nơi bia đá được đưa lên khỏi đất.

Con người không thể đọc hết

Số bia ký cần đọc cứ tiếp tục chất chồng. Số bia ký Hy Lạp do Viện Nhân văn Packard thu thập là 178.551 hiện vật[2]. Tư liệu bia ký Latin thời La Mã cũng vượt quá 176.000 hiện vật[3]. Trong khi đó, số học giả có thể đọc bia ký trên thế giới không nhiều[4].

Ngay cả việc chỉnh lý dữ liệu số cũng không dễ. Khi xử lý dữ liệu này để máy có thể đọc, còn lại 78.608 hiện vật. Hơn một nửa tư liệu bị loại vì quá ngắn hoặc tình trạng quá xấu. Chỉ riêng việc chọn dữ liệu để đưa cho máy học đã là một công việc lớn[2].

Lấp ô trống cũng khó ngay cả với con người. Có một thử nghiệm đưa cho các nhà kim thạch học được đào tạo những bia ký bị xóa chữ và yêu cầu họ điền vào. Tỷ lệ đáp án đầu tiên của con người là đúng chỉ đạt 25 phần trăm[2]. Nghĩa là ba trong bốn lần đều sai. Trong một thí nghiệm trước đó, tỷ lệ lỗi của chữ do con người điền là 57,3 phần trăm[1].

Tình hình giấy cói còn tệ hơn. Giấy cói là tên một loài cỏ mọc ven nước ở Ai Cập. Người xưa thái mỏng phần lõi của loài cỏ đó. Họ xếp các lát đã thái theo chiều ngang và chiều dọc, rồi đập, ép và phơi khô. Vật liệu ghi chép được làm như vậy cũng gọi là giấy cói. Hàng trăm nghìn tờ đã được tìm thấy trong cát khô của Ai Cập. Số giấy cói tiếng Hy Lạp chưa ai đọc được ước tính lên tới một triệu hiện vật[5]. Trên loại giấy này có biên lai thuế, thư từ và hợp đồng. Đó là ghi chép chứa nguyên vẹn một ngày của người xưa.

Việc đọc không kết thúc bằng cách mở hộp. Bởi giấy đã rách và mực đã phai. Trước hết phải xử lý bảo tồn và ghép các mảnh rời. Sau đó mới có thể đọc chữ. Không có đủ người để làm tất cả những việc này[4].

Cách đọc tư liệu cũng là một vấn đề. Trong thời gian dài, tư liệu bia ký chỉ được tích lũy dưới dạng danh sách chữ trong sách. Ảnh của đá hoặc thông tin về chất liệu không được lưu kèm[4]. Điều đó có nghĩa là thông tin có thể đưa vào máy chỉ là chữ.

Việc khai quật vẫn tiếp tục đến nay. Mỗi năm, bia đá mới và giấy cói mới lại được đưa lên khỏi đất. Tư liệu cần đọc tăng lên, nhưng số người đọc vẫn như cũ[4]. Khoảng cách này ngày càng rộng hơn theo từng năm.

Vì vậy, nhiều ghi chép vẫn nằm trong hộp bảo tàng mà chưa được đọc. Có đá và giấy, nhưng thiếu người làm sống lại các câu viết trên đó. Các nhà nghiên cứu đã đưa máy móc vào để khai thông điểm nghẽn này[4].

Tài liệu tham khảo

[1] Assael, Y., Sommerschield, T., & Prag, J. (2019). Restoring ancient text using deep learning: a case study on Greek epigraphy. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 6368-6375). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1668

[2] Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., & de Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. Nature, 603(7900), 280-283. https://doi.org/10.1038/s41586-022-04448-z

[3] Assael, Y., Sommerschield, T., Cooley, A., Pavlopoulos, J., Shillingford, B., Herms, B., Suresh, P., Maynard, B., Grayston, J., Wulgaert, R., Prag, J., Mullen, A., & Mohamed, S. (2025). Contextualizing ancient texts with generative neural networks. Nature, 645(8079), 141-147. https://doi.org/10.1038/s41586-025-09292-5

[4] Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I., & de Freitas, N. (2023). Machine learning for ancient languages: A survey. Computational Linguistics, 49(3), 703-747. https://doi.org/10.1162/coli_a_00481

[5] Reply. (2026, 3월 19일). Austrian Academy of Sciences is developing the Ancient Greek AI "Apollo" with Mistral AI and Reply. reply.com. https://www.reply.com/en/newsroom/news/austrian-academy-of-sciences-is-developing-the-ancient-greek-ai-apollo-with-mistral-ai-and-reply (truy cập ngày 3 tháng 9 năm 2026)

11.2. Truy vết của đội điều tra khoa học: thiết bị tự động hoàn thành văn bản cổ Ithaca và Apollo

Máy học qua bài toán điền ô trống

Cách AI đọc chữ cổ giống bài điền ô trống trong môn ngữ văn. Người ta cho máy xem thật nhiều bia ký hoàn chỉnh. Sau đó che vài chữ và yêu cầu máy đoán phần cần điền. Nếu đáp án sai, các con số bên trong máy được chỉnh từng chút[4]. Khi lặp lại quá trình này hàng triệu lần, máy học được thói quen của câu tiếng Hy Lạp.

Máy học theo cách này không học thuộc lòng từ ngữ. Nó ghi nhớ bằng xác suất rằng sau một từ nào đó, từ nào thường xuất hiện[4]. Vì vậy, ngay cả với bia ký chưa từng thấy, nó vẫn có thể đưa ra các ứng viên phù hợp với ô trống.

Từ Pythia đến Ithaca

Năm 2019, các nhà nghiên cứu công bố một mô hình tên là Pythia. Mô hình này do Yannis Assael, Thea Sommerschield và Jonathan Prag tạo ra. Pythia là một mạng nơ-ron dự đoán chữ bị xóa trong bia ký Hy Lạp[1]. Mạng nơ-ron là cấu trúc tính toán tự học quy tắc bằng cách xem dữ liệu. Nó được mô phỏng theo kết nối giữa các tế bào thần kinh trong não người.

Pythia đọc cùng lúc ở cấp chữ cái và cấp từ. Lý do là để xử lý tư liệu có nửa từ bị vỡ, cần hai con mắt. Khi tỷ lệ lỗi của chữ do con người điền là 57,3 phần trăm, Pythia ghi nhận 30,1 phần trăm. Trường hợp đáp án đúng nằm trong 20 ứng viên của Pythia đạt 73,5 phần trăm[1].

Năm 2022, Ithaca xuất hiện. Google DeepMind và nhiều trường đại học cùng tạo ra mô hình này. Đại học Oxford, Đại học Ca' Foscari Venezia và Đại học Kinh tế và Kinh doanh Athens đã tham gia. Kết quả nghiên cứu được đăng trên tạp chí Nature vào tháng 3 năm 2022[2]. Tên gọi lấy từ đảo Ithaca trong sử thi Odyssey. Nó mang ý nghĩa trở về nhà sau một cuộc phiêu bạt dài.

Ithaca phán đoán như thế nào

Ithaca dùng cấu trúc gọi là transformer. Transformer là mạng nơ-ron xem xét các từ trong câu liên quan với nhau đến mức nào. Phần chữ bị xóa được đưa vào dưới dạng ký hiệu chỉ ô trống. Mô hình tính xác suất của chữ sẽ thay thế ký hiệu đó[2].

Ithaca làm ba việc. Nó lấp chữ bị xóa, đoán bia ký ban đầu ở đâu, và ước tính chữ được khắc khi nào. Nó không học ba việc riêng rẽ, mà học cùng lúc. Phương ngữ và thói quen chính tả là manh mối về địa điểm. Cách dùng từ và khuôn câu là manh mối về thời kỳ. Khi một mạng nơ-ron xử lý các manh mối này cùng nhau, chúng hỗ trợ lẫn nhau[2].

Ithaca không nhìn ảnh của đá. Nó chỉ đọc chữ đã được chép lại rồi phán đoán[2]. Con đường này khác với phương pháp cũ vốn đo niên đại bằng hình dạng nét chữ.

Cách đưa ra niên đại cũng khác. Ithaca không chỉ đích danh một năm duy nhất. Nó chia giai đoạn từ năm 800 trước Công nguyên đến năm 800 sau Công nguyên thành từng khoảng mười năm. Rồi nó gán xác suất cho mỗi khoảng và hiển thị bằng đồ thị[2]. Học giả nhìn đỉnh của đồ thị để phán đoán thời kỳ nào có khả năng cao.

Dữ liệu dùng để huấn luyện là bộ sưu tập bia ký của Viện Nhân văn Packard. Đó chính là 78.608 hiện vật đã nói ở trên. Chỉ tính những từ xuất hiện hơn mười lần cũng đã có 35.884 từ[2].

Khi tạo Ithaca, nhóm nghiên cứu đặt ra ba mục tiêu. Làm việc cùng học giả, hỗ trợ phán đoán, và cho thấy quá trình phán đoán. Mục tiêu không phải là một cỗ máy thay thế con người. Họ nhắm tới một công cụ đi vào cách làm việc của học giả[2].

Ithaca không chỉ đưa ra một đáp án. Nó hiển thị nhiều ứng viên có thể điền vào ô trống, kèm xác suất. Nó cũng đánh dấu trên màn hình những chữ khiến nó phán đoán như vậy. Học giả có thể nhìn thấy máy đã dựa vào đâu để trả lời. Phán đoán cuối cùng vẫn thuộc về con người. Nhóm nghiên cứu đã công khai mã nguồn và dữ liệu của Ithaca[2].

Ý nghĩa của hai mươi ứng viên

Đáp án do máy đưa ra không chỉ có một. Nó xếp các ứng viên theo thứ tự xác suất cao. Tỷ lệ ứng viên đầu tiên đúng được gọi là độ chính xác top 1. Tỷ lệ đáp án đúng nằm trong hai mươi ứng viên đầu được gọi là độ chính xác top 20[1].

Sự phân biệt này rất quan trọng. Thứ học giả cần không phải là một đáp án duy nhất. Đó là danh sách ứng viên đáng xem xét. Nếu đáp án đúng nằm trong hai mươi ứng viên, học giả có thể chọn trong đó[1]. Đôi khi họ cũng tìm thấy trong danh sách đó một từ mà bản thân chưa nghĩ tới[2].

Aeneas đi sang tiếng Latin

Cùng nhóm nghiên cứu đã mở rộng hướng đi sang bia ký Latin. Vào tháng 7 năm 2025, họ công bố trên Nature một mô hình tên là Aeneas[3]. Tên gọi lấy từ Aeneas, nhân vật chính trong câu chuyện lập quốc của La Mã[5].

Aeneas học từ hơn 176.000 bia ký Latin. Đây là dữ liệu hợp nhất từ ba cơ sở dữ liệu bia ký của Rome, Heidelberg và Klaus Slaby. Ở những chỗ bị xóa trong phạm vi mười chữ, tỷ lệ đáp án đúng nằm trong 20

ứng viên đạt 73 phần trăm. Ngay cả khi không biết có bao nhiêu chữ bị xóa, mô hình vẫn đạt 58 phần trăm[3]. Trả lời khi không biết độ dài phần bị xóa là bài toán khó cả với con người.

Mô hình cũng đoán được bia ký đến từ tỉnh nào của La Mã. Tỉnh là vùng đất địa phương do La Mã cai trị. Trong 62 lựa chọn, mô hình đoán đúng 72 phần trăm[5]. Thời điểm khắc được thu hẹp còn trung bình trong vòng 13 năm so với phạm vi do các nhà sử học xác định[3, 5].

Aeneas có chức năng mới. Nó có thể phán đoán bằng cách đưa ảnh bia ký vào cùng[3]. Nghĩa là nó không chỉ nhìn chữ đã chép lại, mà còn nhìn cả hình dạng của đá. Nó cũng tìm các bia ký khác có câu tương tự và hiển thị cạnh nhau. Những trường hợp tương tự mà học giả phải mất vài ngày mới tìm được, mô hình có thể đưa ra chỉ trong vài giây[5].

Nhóm nghiên cứu đã thử áp dụng công cụ này vào ghi chép về hoàng đế La Mã Augustus. Aeneas đưa ra đồng thời hai khoảng thời gian có thể là thời điểm văn bản được khắc. Một khoảng là từ năm 10 trước Công nguyên đến năm 1 trước Công nguyên. Khoảng còn lại là từ năm 10 sau Công nguyên đến năm 20 sau Công nguyên, với xác suất cao hơn[5]. Hai lập luận từng chia rẽ giới học thuật hiện ra nguyên dạng trên đồ thị xác suất.

Nhóm nghiên cứu đã để 23 nhà sử học dùng thử chức năng này. Các học giả trả lời rằng họ đã có ý tưởng nghiên cứu mới ở 9 trong 10 bia khắc[3, 5]. Mô hình này đã được công bố để ai cũng có thể dùng[5].

Apollo học toàn bộ tiếng Hy Lạp

Tháng 3 năm 2026, Viện Hàn lâm Khoa học Áo công bố một mô hình mới. Tên của nó là Apollo. Mô hình này đang được phát triển chung bởi công ty AI Mistral AI và công ty công nghệ thông tin Sail Reply[7]. Nếu Ithaca là công cụ xử lý một bia khắc, thì Apollo đặt mục tiêu trở thành mô hình học toàn bộ tiếng Hy Lạp[6].

Những mô hình như vậy được gọi là mô hình nền tảng. Điều này có nghĩa là một mô hình cơ bản đã được học rộng để dùng cho nhiều việc. Nó không phải là công cụ được tạo ra chỉ để làm một việc[4]. Cùng một nền tảng được dùng cho phục dựng bia khắc, tìm kiếm văn bản và đọc chữ viết[6].

Tư liệu mà Apollo học có quy mô khoảng 600 triệu từ tiếng Hy Lạp. Thêm vào đó là hàng chục nghìn bia khắc và giấy cói đã được xuất bản[6, 7]. Nhóm nghiên cứu cho biết đây là một trong những kho tư liệu tiếng Hy Lạp có quy mô lớn được tập hợp đến nay[7].

Thư viện Quốc gia Áo sở hữu nhiều giấy cói. Chỉ riêng thư viện này đã lưu giữ hàng chục nghìn tư liệu[7]. Trong đó có cả những văn bản chưa được đọc.

Người dẫn dắt nghiên cứu là Anna Dolganov thuộc Viện Khảo cổ học Áo. Bà nói rằng trên thế giới có một triệu giấy cói tiếng Hy Lạp chưa từng được ai đọc[7]. Apollo đảm nhận việc phục dựng chữ và tìm nội dung trong những tư liệu như vậy. Việc đọc chữ viết tay cũng nằm trong mục tiêu của mô hình[6, 7]. Nhóm nghiên cứu cho biết công việc này có thể hoàn tất không phải trong vài năm, mà chỉ trong vài giờ[7].

Tài liệu tham khảo

[1] Assael, Y., Sommerschield, T., & Prag, J. (2019). Restoring ancient text using deep learning: a case study on Greek epigraphy. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 6368-6375). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1668

[2] Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., & de Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. Nature, 603(7900), 280-283. https://doi.org/10.1038/s41586-022-04448-z

[3] Assael, Y., Sommerschield, T., Cooley, A., Pavlopoulos, J., Shillingford, B., Herms, B., Suresh, P., Maynard, B., Grayston, J., Wulgaert, R., Prag, J., Mullen, A., & Mohamed, S. (2025). Contextualizing ancient texts with generative neural networks. Nature, 645(8079), 141-147. https://doi.org/10.1038/s41586-025-09292-5

[4] Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I., & de Freitas, N. (2023). Machine learning for ancient languages: A survey. Computational Linguistics, 49(3), 703-747. https://doi.org/10.1162/coli_a_00481

[5] Google DeepMind. (2025, 7월 23일). Aeneas transforms how historians connect the past. deepmind.google. https://deepmind.google/blog/aeneas-transforms-how-historians-connect-the-past/ (truy cập ngày 3 tháng 9 năm 2026)

[6] Mistral AI. (2026). Austrian Academy of Sciences. mistral.ai. https://mistral.ai/customers/austrian-academy-of-sciences/ (truy cập ngày 3 tháng 9 năm 2026)

[7] Reply. (2026, 3월 19일). Austrian Academy of Sciences is developing the Ancient Greek AI "Apollo" with Mistral AI and Reply. reply.com. https://www.reply.com/en/newsroom/news/austrian-academy-of-sciences-is-developing-the-ancient-greek-ai-apollo-with-mistral-ai-and-reply (truy cập ngày 3 tháng 9 năm 2026)

11.3. Sự thật đã được làm sáng tỏ

62 phần trăm và 72 phần trăm

Nhóm nghiên cứu đã kiểm chứng năng lực của Ithaca bằng con số. Đó là một bài kiểm tra trong đó họ đưa cho mô hình các bia khắc bị cố ý che một phần chữ và yêu cầu điền vào. Bài kiểm tra tương tự cũng được giao cho các nhà kim thạch học đã qua đào tạo[1].

Khi con người tự giải một mình, tỷ lệ đáp án đầu tiên đúng là 25 phần trăm. Khi Ithaca tự giải một mình, tỷ lệ đó là 62 phần trăm. Đây là mức chênh lệch hơn gấp đôi[1].

Bài kiểm tra này còn có điều kiện thứ ba. Đó là điều kiện trong đó con người nhìn danh sách ứng viên của Ithaca và cùng giải. Khi đó, độ chính xác tăng lên 72 phần trăm[1]. Đây là con số cao hơn cả con người một mình lẫn máy một mình.

Con số này là trọng tâm của nghiên cứu. Máy tính xác suất bằng ký ức đã đọc toàn bộ 78.608 mục. Con người biết bia khắc đó nằm trong lịch sử nào. Khi máy đưa ra 20 ứng viên, con người dùng kiến thức lịch sử để sàng lọc. Khi hai yếu tố này kết hợp, độ chính xác là cao nhất[1, 2].

Năng lực đoán đúng địa điểm và thời kỳ cũng được kiểm chứng. Ithaca học bằng cách chia thế giới Hy Lạp thành 84 khu vực. Nó đoán đúng quê gốc của các bia khắc không rõ nơi khai quật với độ chính xác 71 phần trăm. Về thời điểm khắc, nó đưa ra giá trị lệch trung bình 29,3 năm so với khoảng thời gian do các nhà sử học xác định. Phương pháp dùng mắt người để đo kiểu chữ đôi khi có sai số tới 100 năm[1].

Câu trả lời là năm 421 trước Công nguyên

Ithaca cũng đưa ra câu trả lời cho một tranh luận lâu đời. Đó là vấn đề niên đại của bia khắc hiệp ước Athens đã thấy ở phần trước. Vì chữ sigma có ba nét, những bia khắc này từng được xem là có niên đại sớm hơn khoảng năm 446 trước Công nguyên[1].

Nhóm nghiên cứu đưa các bia khắc này vào Ithaca. Ithaca phán đoán không dựa vào hình dạng chữ, mà dựa vào chính câu văn. Nó xem xét từ nào được viết cùng với từ nào. Nó cũng xem xét tên chức vụ nào xuất hiện cùng với cách diễn đạt nào[1].

Niên đại trung bình mà Ithaca đưa ra là năm 421 trước Công nguyên. Muộn hơn một thế hệ so với tiêu chuẩn cũ. Giá trị này cùng hướng với các chứng cứ mới khác. Nó chênh trung bình 5 năm so với niên đại mà các học giả gần đây đã xác định lại. Niên đại theo tiêu chuẩn cũ thì lệch tới 27 năm[1].

Máy không tự mình viết lại lịch sử. Nó đã làm cho một phía trong hai lập luận tranh cãi lâu dài có thêm trọng lượng[1, 2]. Các học giả phải xem lại thói quen ấn định niên đại chỉ bằng một kiểu chữ. Một công cụ đã đổi hướng một cuộc tranh luận kéo dài 100 năm.

Những câu văn có vẻ hợp lý do máy tạo ra

Dù có các thành quả này, vẫn còn vấn đề. Đó là tư liệu học bị lệch về một phía. Trong tư liệu bia khắc tiếng Hy Lạp, có nhiều tư liệu từ Athens. Về thời đại, tư liệu cũng tập trung vào thời cổ điển. Vì vậy, bia khắc ở các khu vực rìa có ít tư liệu sẽ kém chính xác hơn[2].

Cũng có vấn đề là khó biết vì sao máy trả lời như vậy. Bên trong mạng nơ-ron có hàng triệu con số. Nó cho thấy máy tập trung hơn vào chữ nào, nhưng không thể giải thích bằng lời vì sao đã chọn câu trả lời đó[2].

Vấn đề cũng phát sinh khi chỗ trống trở nên dài. Máy làm tốt việc điền vài chữ, nhưng khó xử lý nơi nhiều từ biến mất hoàn toàn. Trong bài kiểm tra Aeneas, khi không biết độ dài phần bị xóa, kết quả giảm từ 73 phần trăm xuống 58 phần trăm[3]. Khi đó, mô hình có thể tạo ra câu đúng ngữ pháp và có vẻ hợp lý, nhưng thực tế chưa từng tồn tại. Câu trả lời được tạo ra như vậy được gọi là ảo giác. Điều này có nghĩa là câu trả lời sai do máy bịa ra như thể là sự thật. Ảo giác nguy hiểm khi xử lý ghi chép cổ. Vì lịch sử chưa từng có có thể đông cứng lại như sự thật[2].

Một nghiên cứu công bố tháng 8 năm 2026 đã trực diện xử lý giới hạn này. Đó là thí nghiệm phục dựng các văn bản cổ bị hư hại trong nhiều mô hình và nhiều điều kiện. Nhóm nghiên cứu kết luận rằng không thể giao việc phục dựng hoàn toàn cho tự động hóa. Kết quả thay đổi lớn tùy theo vị trí cần điền trong văn bản. Những văn bản lặp lại cụm từ cố định được điền tốt. Những câu mỗi người viết khác nhau thì khó hơn. Kết quả cũng khác nhau tùy theo có biết độ dài phần bị xóa hay không. Nhóm nghiên cứu cho biết cần tạo máy thành công cụ hỗ trợ các nhà cổ văn tự học[4].

Các nhà nghiên cứu đặt ra quy tắc để giảm rủi ro này. Chữ do máy điền cũng được đặt trong ngoặc vuông. Những ứng viên có xác suất thấp được học giả loại bỏ. Các quy ước ký hiệu mà các nhà kim thạch học xưa tuân thủ cũng được áp dụng nguyên như vậy cho máy[2].

Vì vậy, câu trả lời mà Ithaca đưa ra không phải là sự kiện lịch sử. Đó là danh sách giả thuyết cần xem xét. Chỉ khi học giả đối chiếu với tư liệu khác và xác nhận, nó mới trở thành sử liệu[2]. Sử liệu là tư liệu căn cứ để làm sáng tỏ lịch sử.

Những việc xảy ra trong 7 năm

Tốc độ thay đổi của lĩnh vực này có thể so sánh bằng con số. Năm 2019, Pythia đã hạ tỷ lệ lỗi của con người từ 57,3 phần trăm xuống 30,1 phần trăm[5]. Năm 2022, Ithaca nâng độ chính xác của đáp án đầu tiên lên 62 phần trăm[1]. Năm 2025, Aeneas chuyển sang tiếng Latin và thu hẹp sai số niên đại xuống trung bình 13 năm[3]. Năm 2026, Apollo, mô hình học toàn bộ tiếng Hy Lạp, được khởi động[6].

Trong 7 năm, ngôn ngữ mục tiêu tăng lên và độ chính xác cũng tăng. Tư liệu xử lý cũng mở rộng từ chữ viết sang ảnh[3]. Trong thời gian đó, có một điều không thay đổi. Đó là nguyên tắc con người đưa ra phán đoán cuối cùng[2].

Ảnh, ngữ cảnh và bước tiếp theo

Nghiên cứu tiếp tục mở rộng. Một bài báo ra năm 2026 đã huấn luyện lại một mô hình ngôn ngữ lớn phổ biến bằng tư liệu tiếng Hy Lạp. Đây là thí nghiệm dùng mô hình Llama 3.1 do Meta công bố. Trong phục dựng giấy cói, tỷ lệ đáp án đầu tiên đúng là 73,5 phần trăm. Trong phục dựng bia khắc, tỷ lệ là 63,7 phần trăm. Tỷ lệ đáp án đúng nằm trong 20 ứng viên lần lượt là 86,0 phần trăm và 83,0 phần trăm. Điều này có nghĩa là ngay cả mô hình không được tạo riêng cho bia khắc cũng có thể làm việc này[7].

Nghiên cứu phục hồi hình ảnh chữ cũng đã xuất hiện. Mesa, công bố tháng 4 năm 2026, là phương pháp điền các nét bị xóa trong ảnh bia khắc. Nó dùng những chữ còn rõ trên cùng tấm bia làm mẫu để phục hồi hoa văn ở phần bị mòn[8]. Mô hình điền câu và mô hình điền ảnh đang phát triển song song.

Những bi kịch đã mất cũng đang thực sự trở lại. Năm 2022, một tờ giấy cói được phát hiện tại di tích Philadelphia ở Ai Cập. Đó là tờ giấy được cho là viết vào thế kỷ 3 sau Công nguyên. Năm 2024, các học giả đọc và công bố nội dung của nó. Đó là 97 dòng lời thoại trong hai bi kịch Polyidus và Ino do Euripides viết. Trước đó, hai tác phẩm này chỉ được biết đến qua các trích dẫn ngắn và tóm tắt cốt truyện[9].

Việc phục dựng này do bàn tay và con mắt của con người thực hiện. Không phải máy làm. Trên thế giới còn một triệu tờ giấy như vậy. Chỉ bằng mắt người thì không thể đọc hết chồng giấy này. Đây chính là nơi mà các mô hình như Apollo nhắm tới[6].

Khi công cụ được công bố, cách nghiên cứu cũng thay đổi. Ithaca và Aeneas đã công bố mã và dữ liệu[1, 10]. Dù không thuộc phòng thí nghiệm đại học, ai cũng có thể dùng thử trên web. Chỉ cần tải lên một ảnh bia khắc, danh sách ứng viên, bản đồ và đồ thị niên đại sẽ hiện ra[10]. Trước đây, đó là công việc so sánh mất vài năm[3].

Việc còn lại vẫn rất nhiều. Chỉ riêng giấy cói tiếng Hy Lạp đã có một triệu tư liệu chưa được đọc[6]. Tư liệu bia khắc không đồng đều theo khu vực và thời đại[2]. Những phần khuyết dài mà máy xử lý chưa tốt vẫn còn nguyên[3]. Cần giảm dần từng vấn đề này bằng tư liệu tốt hơn và nhiều kiểm chứng hơn[4].

Máy đưa ra ứng viên, con người chọn. Phương pháp đã nâng Ithaca từ 62 phần trăm lên 72 phần trăm chính là câu trả lời[1]. Để đá và giấy từng im lặng có thể lên tiếng trở lại, hai phía phải cùng có mặt.

Tài liệu tham khảo

[1] Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., & de Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. Nature, 603(7900), 280-283. https://doi.org/10.1038/s41586-022-04448-z

[2] Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I., & de Freitas, N. (2023). Machine learning for ancient languages: A survey. Computational Linguistics, 49(3), 703-747. https://doi.org/10.1162/coli_a_00481

[3] Assael, Y., Sommerschield, T., Cooley, A., Pavlopoulos, J., Shillingford, B., Herms, B., Suresh, P., Maynard, B., Grayston, J., Wulgaert, R., Prag, J., Mullen, A., & Mohamed, S. (2025). Contextualizing ancient texts with generative neural networks. Nature, 645(8079), 141-147. https://doi.org/10.1038/s41586-025-09292-5

[4] Zhang, S., Caraffa, E., Zuffrano, A., Modesti, M., & Colavizza, G. (2026, 8월 28일). Text restoration of ancient documents with language models (arXiv:2608.28170). arXiv. https://arxiv.org/abs/2608.28170 (truy cập ngày 3 tháng 9 năm 2026)

[5] Assael, Y., Sommerschield, T., & Prag, J. (2019). Restoring ancient text using deep learning: a case study on Greek epigraphy. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 6368-6375). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1668

[6] Reply. (2026, 3월 19일). Austrian Academy of Sciences is developing the Ancient Greek AI "Apollo" with Mistral AI and Reply. reply.com. https://www.reply.com/en/newsroom/news/austrian-academy-of-sciences-is-developing-the-ancient-greek-ai-apollo-with-mistral-ai-and-reply (truy cập ngày 3 tháng 9 năm 2026)

[7] Cullhed, E. (2026). Instruction-tuning pretrained causal language models to restore ancient Greek papyri and inscriptions. Digital Scholarship in the Humanities, 41(1), 69-76. https://doi.org/10.1093/llc/fqaf131

[8] Toulatzis, V., Theodoridou, S., & Fudos, I. (2026, 4월 19일). MESA: A training-free multi-exemplar deep framework for restoring ancient inscription textures (arXiv:2604.17390). arXiv. https://arxiv.org/abs/2604.17390 (truy cập ngày 3 tháng 9 năm 2026)

[9] O'Grady, E. (2024, 10월 11일). Unearthed papyrus contains lost scenes from Euripides' plays. The Harvard Gazette. https://news.harvard.edu/gazette/story/2024/10/unearthed-papyrus-contains-lost-scenes-from-euripides-plays/ (truy cập ngày 3 tháng 9 năm 2026)

[10] Google DeepMind. (2025, 7월 23일). Aeneas transforms how historians connect the past. deepmind.google. https://deepmind.google/blog/aeneas-transforms-how-historians-connect-the-past/ (truy cập ngày 3 tháng 9 năm 2026)

Chương 12. Thư viện ảo không bao giờ cháy: kho lưu trữ đa phổ siêu độ phân giải cao và tương lai

12.1. Hồ sơ vụ việc

Cách Thư viện Alexandria biến mất

Ở Alexandria, Ai Cập từng có thư viện lớn nhất thế giới cổ đại. Triều Ptolemaios lập thư viện này vào khoảng thế kỷ 3 trước Công nguyên. Các vị vua muốn gom sách của thế gian vào một nơi. Sách trên các tàu vào cảng bị thu lại để sao chép. Bản sao được trả cho chủ, còn bản gốc thuộc về thư viện. Ghi chép xưa kể rằng thư viện có hàng trăm nghìn cuộn sách [1].

Nhiều người nghĩ thư viện này biến mất vì một trận hỏa hoạn lớn. Nhưng các nhà nghiên cứu cho rằng nó sụp đổ dần dần. Năm 145 trước Công nguyên, nhà vua trục xuất các học giả nước ngoài. Khi học giả rời đi, những bàn tay chăm sóc sách cũng rời đi. Vào thời La Mã, các tòa nhà nhiều lần bị phá hủy. Tiền mua sách mới cũng bị cắt [1].

Chữ viết cổ vốn khó tồn tại lâu. Các thư lại thời đó thường sao chép sách có sẵn hơn là viết tác phẩm mới. Những cuộn giấy cói bị dùng nhiều sẽ cũ đi sau khoảng một trăm năm và khó đọc. Trong lòng đất khô của Ai Cập, chúng đôi khi tồn tại lâu hơn nhiều. Vì vậy, phải có người sao chép lại đúng lúc thì văn bản mới tiếp tục tồn tại. Khi việc sao chép dừng lại, cuốn sách ấy bị xóa khỏi thế gian.

Những văn bản cổ mà chúng ta đọc hôm nay chỉ là một phần nhỏ của chữ viết thời đó. Số sách chỉ còn tên mà mất phần nội dung lớn hơn nhiều. Phần lớn tác phẩm của các nhà viết bi kịch Hy Lạp cổ đại cũng biến mất như vậy. Những sách còn lại là những sách may mắn có bản sao rải rác ở nhiều nơi.

Sự phá hủy vẫn tiếp diễn

Sức mạnh phá hủy ghi chép không chỉ tồn tại trong quá khứ. Chiến tranh ngày nay vẫn làm sụp đổ thư viện và kho lưu trữ. UNESCO đang xác nhận và thống kê từng cơ sở văn hóa bị phá hủy ở Ukraine. Tính đến ngày 29 tháng 7 năm 2026, số nơi được xác nhận là 545. Trong đó có 43 bảo tàng, 22 thư viện và 5 kho lưu trữ. Hàng trăm cơ sở tôn giáo và công trình cổ cũng bị hư hại [2].

Chiến tranh không chỉ phá hủy di vật. Di vật cũng có thể bị lén đưa ra khỏi kho bảo quản khi việc quản lý dừng lại. Một di vật bị xóa mất ghi chép về nơi xuất xứ sẽ mất phần lớn giá trị nghiên cứu. Đó là vì việc một tấm đất sét nằm trong phòng nào và cùng những tấm nào chính là thông tin.

Khí hậu cũng đe dọa ghi chép. Di tích ven biển bị sóng bào mòn và sụp đổ từng chút một. Di vật trong lòng đất hư hại nhanh hơn khi mưa tăng hoặc nước ngầm dâng. Có nơi tốc độ biến mất nhanh hơn tốc độ khai quật.

Không phải mọi di vật vỡ nát đều biến mất. Nếu hình dạng còn lại, vẫn có cách dựng lại. Vì vậy, việc đo đạc trước khi tai nạn xảy ra rất quan trọng. Ở những vùng chiến tranh kéo dài, người ta đang đo đạc công trình và di vật bằng 3D.

Mỗi chất liệu có tuổi thọ khác nhau

Chất liệu lưu giữ ghi chép hư hỏng theo những cách khác nhau. Tấm đất sét Lưỡng Hà là tấm được ép từ bùn để khắc chữ. Nhiệt sinh ra khi cung điện cháy đã nung tấm đất sét cứng như đồ gốm. Những tấm đất sét không gặp lửa sẽ tan trở lại thành bùn khi nước ngấm vào.

Tấm đất sét còn có giới hạn khác. Các thư lại làm ướt tấm đã viết xong, gạt bỏ bề mặt rồi viết lại. Vì vậy, các tấm cũ thường trở về thành bùn. Đó là lý do nhiều trường hợp chỉ còn lại vài năm ngay trước khi vương quốc sụp đổ. Chỉ những tấm cuối cùng trong kho sách khi cung điện cháy mới được nung cứng.

Giấy cói, da và giấy yếu hơn. Những chất liệu này là vật chất hữu cơ từ thực vật và động vật. Ở nơi nóng ẩm, chúng không chịu nổi vài trăm năm mà mục nát. Trên bảng sáp của binh lính La Mã, lớp sáp đã biến mất hoàn toàn. Phần còn lại chỉ là những vết nông do bút sắt cào qua tới tấm gỗ.

Chất liệu cứng cũng không an toàn. Giáp cốt văn khắc trên yếm rùa và xương vai bò cũng bị ép vỡ trong lòng đất. Phần giáp cốt còn lại hôm nay đa số là mảnh nhỏ hơn lòng bàn tay. Có khi một câu bị tách thành ba mảnh và nằm trong các kho khác nhau [3].

Thư viện được lửa cứu

Cách ghi chép sống sót đôi khi rất bất ngờ. Vua Assyria Ashurbanipal đặt một kho sách lớn ở Nineveh. Ông cai trị đất nước từ khoảng năm 669 trước Công nguyên. Trong kho sách có thần thoại, bói toán, y học và văn bản hành chính [4].

Khoảng năm 612 trước Công nguyên, Nineveh bị đốt cháy. Nếu đó là sách giấy, tất cả đã biến mất khi ấy. Nhưng tấm đất sét được lửa nung nên lại cứng hơn. Các tấm này được khai quật nhiều lần từ thập niên 1840 đến thập niên 1930. Hiện Bảo tàng Anh có hơn 30.000 mảnh như vậy [4].

Nhưng chúng không được tìm thấy trong tình trạng nguyên vẹn. Khi thành phố sụp đổ, các tấm bị cố ý đập vỡ và vứt ở nhiều nơi trong cung điện. Thứ các nhà khai quật gặp không phải giá sách ngăn nắp, mà là đống mảnh vỡ rải rác. Việc phần trước và phần sau của một tác phẩm xuất hiện ở các phòng khác nhau là chuyện thường gặp [4]. Ghép lại các mảnh này là bài toán lớn của khảo cổ học tính toán ngày nay. Khảo cổ học tính toán là lĩnh vực nghiên cứu di vật cổ bằng máy tính.

Những chữ viết đã được hồi sinh

Việc chuyển sang dạng số cũng có rào cản. Di vật trên thế giới nhiều hơn rất xa sức người. Đống mảnh vỡ Nineveh đã nêu là một ví dụ [4]. Chụp từng hiện vật rồi gắn cả chữ đọc được vào đó mất rất nhiều thời gian. Vì vậy, máy móc phải giúp tăng tốc độ của con người.

Cách cứu di vật đang biến mất chia thành hai hướng. Một là bảo tồn để giữ chính đồ vật lâu hơn. Hướng kia là đọc bên trong mà không chạm vào đồ vật rồi chuyển vào máy tính. Cách thứ hai đã nhiều lần đạt kết quả.

Cuộn sách tìm thấy ở En-Gedi, Israel đã cháy thành than. Nếu cố mở bằng tay, nó sẽ vỡ vụn ngay. Nhóm nghiên cứu nhìn vào bên trong bằng chụp CT, cách nhìn bên trong mà không cắt vật thể. Máy tính lần theo các lớp da cuộn bằng toán học rồi trải chúng thành mặt phẳng. Trên đó, mực có pha thành phần kim loại hiện ra và phần đầu sách Lê-vi được đọc [5].

Các cuộn giấy Herculaneum bị chôn dưới tro núi lửa Vesuvius cũng đã mở ra một con đường. Ngày 25 tháng 6 năm 2026, nhóm nghiên cứu cho biết họ đã đọc được một cuộn từ đầu đến cuối. Tên cuộn này là PHerc. 1667. Bên trong có một câu chuyện đạo đức bằng tiếng Hy Lạp dài 22 cột [6]. Mười ba cuộn còn niêm phong được chọn làm thử thách đã có giải thưởng mới. Hạn cuối là ngày 25 tháng 6 năm 2027 [7].

Thư viện ảo không phải là kho chất đầy ảnh. Hình dạng di vật, thành phần mực và văn bản đã đọc phải được buộc lại với nhau ở một nơi. Khi phương pháp mới xuất hiện sau này, dữ liệu cũ mới có thể được tính toán lại. Ngay cả sau khi bản gốc biến mất, việc kiểm chứng vẫn cần gói dữ liệu này.

Chụp ảnh và đọc giải mã cần tiền và con người. Việc quyết định chụp di vật nào trước cũng không dễ. Di vật ở vùng nguy hiểm khó được chụp ngay từ đầu. Vì vậy, mỗi nước đều đặt thứ tự rồi làm theo.

Những vụ việc này để lại một câu hỏi. Nếu đồ vật cuối cùng vỡ nát, chữ bên trong có thể còn lại không. Nơi tìm câu trả lời chính là thư viện ảo.

Tài liệu tham khảo

[1] Mark, J. J. (2023, 7월 25일). Library of Alexandria. World History Encyclopedia. https://www.worldhistory.org/Library_of_Alexandria/ (truy cập ngày 3 tháng 9 năm 2026)

[2] UNESCO. (2026). Damaged cultural sites in Ukraine verified by UNESCO. UNESCO. https://www.unesco.org/en/ukraine-war/damaged-cultural-sites (truy cập ngày 3 tháng 9 năm 2026)

[3] Ma, Y. R. (2026). Towards Computational Chinese Paleography. arXiv:2601.06753. https://arxiv.org/abs/2601.06753 (truy cập ngày 3 tháng 9 năm 2026)

[4] The British Museum. What was Ashurbanipal's Library? The British Museum. https://www.britishmuseum.org/research/projects/what-was-ashurbanipals-library (truy cập ngày 3 tháng 9 năm 2026)

[5] Seales, W. B., Parker, C. S., Segal, M., Tov, E., Shor, P., & Porath, Y. (2016). From damage to discovery via virtual unwrapping: Reading the scroll from En-Gedi. Science Advances, 2(9), e1601247. https://doi.org/10.1126/sciadv.1601247

[6] Vesuvius Challenge. (2026, 6월 25일). An entire Herculaneum scroll has been read for the first time. scrollprize.org. https://scrollprize.org/firstscroll (truy cập ngày 3 tháng 9 năm 2026)

[7] Vesuvius Challenge. (2026). Open Prizes. scrollprize.org. https://scrollprize.org/prizes (truy cập ngày 3 tháng 9 năm 2026)

12.2. Cuộc truy vết của đội điều tra khoa học: đồ thị tri thức và phân tích phổ ánh sáng

Tấm bản đồ lớn nối các di vật

Trong phòng nghiên cứu xưa, hình dạng và chữ viết của di vật được xử lý riêng. Hình dạng do nhà khoa học bảo tồn đo đạc, còn chữ do nhà ngữ văn đọc. Hai kết quả đi vào hai ngăn kéo khác nhau. Vì vậy, dù nghiên cứu cùng một di vật hai lần, họ vẫn không ghép được câu trả lời.

Khảo cổ học tính toán đặt cả hai lên một bản đồ. Tên của bản đồ đó là đồ thị tri thức. Đồ thị tri thức là hình vẽ thông tin bằng điểm và đường. Điểm là sự vật, còn đường là quan hệ giữa các sự vật.

Trong bản đồ này, một tấm đất sét trở thành một điểm. Người thư lại viết tấm đó cũng là một điểm, và thành phố nơi tấm đó được tìm thấy cũng là một điểm. Thành phần mực và thông tin hình dạng 3D cũng lần lượt trở thành điểm. Tên người và tên đồ vật ghi trên tấm cũng được rút ra thành điểm. Máy tính học bằng cách tìm các đường nối những điểm này.

Khi bản đồ được tạo ra, hình dạng câu hỏi thay đổi. Trước đây, người ta đặt một tấm trước mặt rồi hỏi đó là chữ gì. Giờ đây, có thể hỏi những văn bản khác dùng loại mực này nằm ở đâu. Cũng có thể tìm ngay những tấm do cùng một thư lại viết được tìm thấy ở thành phố nào.

Tháng 2 năm 2026, nhóm nghiên cứu của Đại học Jagiellonian ở Ba Lan công bố một bản đồ như vậy. Bản đồ này chứa khoảng 3,2 triệu dòng quan hệ giữa con người, sự kiện và đồ vật. Nhóm nghiên cứu làm cho máy gợi ý quan hệ và giải thích cả lý do. Các chuyên gia đọc phần giải thích đó và kiểm tra xem có dùng được không [1].

Lớp đo hình dạng

Lớp đầu tiên của bản đồ là hình dạng bên ngoài của di vật. Bia đá và tấm đất sét cổ bị mài mòn bề mặt, làm rãnh chữ nông đi. Bằng mắt thường khó biết rãnh còn hay không. Vì vậy, các nhà nghiên cứu quét bề mặt bằng máy quét 3D.

Máy quét biến bề mặt thành hàng triệu điểm rồi lưu lại. Máy tính nối các điểm đó thành lưới tam giác. Trên lưới này, có thể chiếu ánh sáng ảo từ nhiều hướng. Khi đổi hướng ánh sáng, cả rãnh nông cũng tạo bóng và chữ hiện ra.

Máy cũng học hình dạng mặt cắt của rãnh. Vết do con người khắc bằng công cụ có mặt cắt đều và hướng ổn định. Vết nứt tự nhiên thì không như vậy. Mô hình học được khác biệt này sẽ tách chữ khỏi vết xước.

Cũng bằng cách này, mặt cắt bị vỡ được đo đạc. Máy đối chiếu xem độ cong mặt cắt của hai mảnh có khớp vào nhau không. Phép tính này thay cho việc con người cầm trực tiếp các mảnh để ghép thử. Nó cũng có thể ghép với mảnh ở bảo tàng nước khác mà tay không thể chạm tới.

Với cuộn sách đang cuộn, một phép tính khác được dùng. Trong cuộn En-Gedi, nhóm nghiên cứu tách từng lớp da bên trong máy tính. Sau đó, họ trải từng lớp thành mặt phẳng để dựng chữ lên [2].

Lớp đọc mực bằng ánh sáng

Lớp thứ hai của bản đồ là mực. Mỗi vật liệu mực tiếp nhận ánh sáng khác nhau. Chụp ảnh đa phổ là phương pháp chụp tách không chỉ ánh sáng nhìn thấy, mà cả tia hồng ngoại và tia cực tím. Ở một bước sóng nào đó, nền tối đi còn mực sáng lên.

Nhờ khác biệt này, chữ đã bị xóa cũng được hồi sinh. Người xưa xóa chữ rồi viết lại để tiết kiệm da thuộc đắt tiền. Trên bề mặt đã xóa vẫn còn lại một ít thành phần mực. Khi chụp bằng nhiều bước sóng rồi chồng lên nhau, dấu vết đó hiện lên từ bên dưới. Văn bản được viết hai lần như vậy gọi là palimpsest.

Tu viện Thánh Catherine trên núi Sinai, Ai Cập có nhiều văn bản như vậy. Các nhà nghiên cứu Mỹ và Hy Lạp đã chụp đa phổ 74 cuốn tại đây. Số mặt được chụp lên tới 6.800 trang. Bên dưới đó, 305 văn bản đã bị xóa lại hiện ra. Những văn bản này được viết bằng 10 ngôn ngữ, trong đó có tiếng Hy Lạp và tiếng Syriac [3].

Cũng có phương pháp đo chính thành phần. Quang phổ Raman chiếu ánh sáng vào mực và đọc hoa văn của ánh sáng bật trở lại. Hoa văn khác nhau theo từng chất nên đóng vai trò như dấu vân tay. AI nhìn hoa văn này để phân biệt mực làm từ muội than với mực có thành phần sắt. Công thức pha mực khác nhau đôi chút theo từng xưởng, nên trở thành manh mối đoán nơi xuất xứ.

Với giấy cói Herculaneum, ánh sáng mạnh hơn được dùng. Nhóm nghiên cứu chụp bên trong cuộn sách bằng máy gia tốc synchrotron tạo X-ray rất mạnh. Giấy cói bị cháy và mực carbon có mật độ gần giống nhau nên bình thường không phân biệt được. Mô hình học sâu học khác biệt rất nhỏ trong kết cấu và tìm ra nơi từng có chữ [4].

Lớp nối văn bản với văn bản

Lớp thứ ba của bản đồ là quan hệ giữa các văn bản. Sách cổ lan truyền qua việc sao chép. Mỗi lần sao chép lại có một lỗi nhỏ chen vào. Các học giả lần theo những lỗi này để vẽ phả hệ của các bản sao.

Máy tính mượn cách tính từng dùng trong sinh học cho công việc này. Nó giống cách so sánh chuỗi gien để vẽ cây phả hệ của sinh vật. Máy đếm khác biệt giữa các văn bản và sắp xếp bản sao nào sinh ra từ bản sao nào. Phả hệ vẽ như vậy giúp hình dung diện mạo của bản gốc đã mất.

Kiểu chữ cũng là manh mối nối quan hệ. Mỗi thư lại có lực nhấn bút và góc nét khác nhau. AI biến thói quen này thành số rồi lưu lại. Khi gom các văn bản có các số gần nhau, mạng lưới quan hệ của thư lại xuất hiện.

Những kho lưu trữ đang vận hành

Có nhiều kho lưu trữ thực sự vận hành bằng cách gom các lớp này lại. Trismegistos là dự án gán số cho văn bản của thế giới cổ đại. Nơi này đã buộc khoảng 960.000 tư liệu bằng hơn 100 ngôn ngữ vào một hệ thống. Vấn đề cùng một di vật bị mỗi nước gọi bằng một tên khác nhau được giải quyết bằng mã số này [5].

Mạng lưới Pelagios là cộng đồng kết nối các tư liệu rải rác với nhau. Họ đặt ra một cách chung để nối địa điểm, con người và thời gian. Họ làm cho địa danh cổ dù thay đổi theo từng thời kỳ vẫn được buộc vào một tọa độ [6]. Có Thư viện Babylon điện tử do Munich vận hành. Nơi này gom các mảnh chữ hình nêm trên khắp thế giới vào một cơ sở dữ liệu. Các học giả tìm kiếm mảnh tại đây để khôi phục những câu đã mất [7].

Số hiệu chung trông có vẻ nhàm chán, nhưng có sức mạnh lớn. Chỉ khi số hiệu giống nhau, tư liệu của các cơ quan khác nhau mới nối thành một dòng [5]. Thường có trường hợp ảnh nằm ở bảo tàng, văn bản đã đọc ra nằm ở đại học, còn ghi chép khai quật nằm ở viện nghiên cứu. Khi số hiệu buộc ba phần này lại với nhau, thông tin rời rạc gom thành câu chuyện của một hiện vật.

Cách công bố tư liệu cũng đã thay đổi. Trước đây, chỉ sau khi bài báo được xuất bản, người ta mới có thể xem ảnh và văn bản đã đọc ra. Gần đây, các nhóm nghiên cứu công bố cả dữ liệu quét và chương trình. Nhờ vậy, nhà nghiên cứu ở những nước không có thiết bị cũng có thể tính toán lại trên cùng tư liệu [8].

Những bài toán còn lại

Vesuvius Challenge đã tổng hợp và công bố các vấn đề còn lại vào ngày 10 tháng 7 năm 2026. Ở những chỗ bị ép và nát, các lớp của cuộn giấy không tách khỏi nhau. Máy tính khi lần theo một lớp đôi khi cũng nhảy nhầm sang lớp bên cạnh. Mực đã thấm sâu bao nhiêu từ bề mặt vẫn chưa được biết chắc. Mô hình học từ một cuộn giấy lại mất sức ở cuộn giấy khác. Dữ liệu quá lớn nên máy tính thông thường khó xử lý. Mốc tiếp theo mà nhóm nghiên cứu đặt ra là tháng 6 năm 2027 [8].

Các lĩnh vực khác xử lý chữ viết cổ cũng gặp cùng bức tường. Một bài báo năm 2026 tổng kết nghiên cứu cổ tự Trung Quốc đã nêu thiếu dữ liệu là một khó khăn lớn. Bài báo cũng khuyến nghị hướng dùng AI để hỗ trợ phán đoán của con người, thay vì thay thế con người [9].

Tài liệu tham khảo

[1] Kutt, K., Sroka, E., Ishchuk, O., & do Valle Miranda, L. (2026). A three-stage neuro-symbolic recommendation pipeline for cultural heritage knowledge graphs. arXiv:2602.19711. https://arxiv.org/abs/2602.19711 (truy cập ngày 3 tháng 9 năm 2026)

[2] Seales, W. B., Parker, C. S., Segal, M., Tov, E., Shor, P., & Porath, Y. (2016). From damage to discovery via virtual unwrapping: Reading the scroll from En-Gedi. Science Advances, 2(9), e1601247. https://doi.org/10.1126/sciadv.1601247

[3] Early Manuscripts Electronic Library. Sinai Palimpsests Project. EMEL. https://www.emelibrary.org/sinai-palimpsests-project/ (truy cập ngày 3 tháng 9 năm 2026)

[4] Vesuvius Challenge. (2026, 6월 25일). An entire Herculaneum scroll has been read for the first time. scrollprize.org. https://scrollprize.org/firstscroll (truy cập ngày 3 tháng 9 năm 2026)

[5] KU Leuven. Trismegistos portal. Trismegistos+. https://tmplus.kuleuven.be/trismegistos-portal (truy cập ngày 3 tháng 9 năm 2026)

[6] Pelagios Network. Pelagios Network. https://pelagios.org/ (truy cập ngày 3 tháng 9 năm 2026)

[7] Ludwig-Maximilians-Universität München & Bayerische Akademie der Wissenschaften. electronic Babylonian Library. https://www.ebl.badw.de/ (truy cập ngày 3 tháng 9 năm 2026)

[8] Vesuvius Challenge. (2026, 7월 10일). Open problems: Why reading every Herculaneum scroll is still a challenge. scrollprize.org. https://scrollprize.org/2026_open_problems (truy cập ngày 3 tháng 9 năm 2026)

[9] Ma, Y. R. (2026). Towards Computational Chinese Paleography. arXiv:2601.06753. https://arxiv.org/abs/2601.06753 (truy cập ngày 3 tháng 9 năm 2026)

12.3. Lời kết

Nhìn lại mười một sự kiện

Cuốn sách này đã đi theo mười một sự kiện. Điểm bắt đầu là các cuộn giấy Herculaneum bị vùi trong tro núi lửa. Máy tính đã mở ảo khối than chạm vào là vỡ. Giấy cói bị cháy và mực carbon nhìn bằng X-ray trông gần như giống nhau. Máy học những khác biệt rất nhỏ trong vân còn lại trên bề mặt để tìm ra chữ [1].

Cuộn sách Lê-vi ở En-Gedi cũng lấy lại được câu đầu tiên bằng mở cuộn ảo. Ở trường hợp này, mực có thành phần kim loại hiện rõ trên X-ray [2]. Trong Cuộn Isaiah, kiểu chữ là manh mối. AI đo những thói quen rất nhỏ trong nét chữ và cho thấy người chép không chỉ có một người. Đó là phần mà mắt người chỉ thấy như cùng một nét chữ [3].

Từ sự kiện thứ tư trở đi là công việc ghép mảnh. Các mảnh bảng đất sét và da thuộc rải rác được ghép lại theo hình dạng mặt cắt. Cách tính toán đó cũng được dùng để nhận ra hiện vật giả. Những dòng bị thiếu trong Sử thi Gilgamesh được lấp đầy bằng cách nối các mảnh vỡ.

Những biên nhận của Mesopotamia đã bắt đầu cất tiếng nhờ dịch máy. Sổ ngũ cốc và thư từ của hàng nghìn năm trước được chuyển sang ngôn ngữ ngày nay. Giáp cốt văn trên yếm rùa được ghép lại bằng quét 3D và mô hình chuyên dụng [4]. Thẻ gỗ của Silla cũng được mô hình nhận dạng chữ đọc ra. Tài liệu chữ thảo của Nhật Bản cũng được giải theo cách tương tự [5].

Chúng ta cũng đã thấy những cánh cửa chưa mở. Chữ tuyến tính A, chữ viết Ấn Hà và rongorongo vẫn chưa có lời giải. Máy đo tính quy luật trong cách sắp xếp chữ để xem đó có phải là ngôn ngữ hay không. Việc đọc nghĩa còn lại cho thế hệ sau.

Sự kiện cuối cùng là tự động hoàn thành văn khắc cổ đại. Ithaca là mô hình điền những chữ bị thiếu trong văn khắc Hy Lạp. Khi điền chữ, mô hình này cũng xác định nơi và thời điểm văn khắc được tạo ra [6]. Aeneas xuất hiện sau đó xử lý văn khắc Latin. Aeneas xem cả chữ được chép lại và ảnh văn khắc [7].

Mười một sự kiện có một điểm chung. Không có vấn đề nào được giải bởi máy một mình. Nhà vật lý phụ trách chụp ảnh, nhà khoa học máy tính phụ trách tính toán. Học giả đọc ngôn ngữ cổ kiểm tra kết quả và gắn nghĩa cho nó. Khi ba lĩnh vực gặp nhau trước cùng một hiện vật, một dòng chữ mới sống lại.

Những việc máy vẫn chưa làm được

Trong suốt hành trình đến đây, điểm yếu của máy cũng lộ ra. Bức tường đầu tiên là thiếu dữ liệu. Chữ tuyến tính A dù cộng tất cả ký hiệu còn lại cũng chỉ khoảng 7.400 ký hiệu. Với lượng dữ liệu như vậy, máy khó học được quy luật [8]. Các mô hình ngôn ngữ ngày nay lớn lên bằng cách đọc hàng tỷ từ. Tư liệu của tiếng cổ đã biến mất còn kém xa quy mô đó.

Ở nơi dữ liệu ít, nếu cố đưa ra đáp án, máy sẽ bịa ra từ không tồn tại. Khoảng trống càng dài, độ tin cậy của đáp án càng giảm. Có thể xuất hiện câu nghe hợp văn cảnh nhưng chưa từng có trong lịch sử. Khi câu như vậy đã được đưa vào một bài báo, nghiên cứu sau đó có thể nhận nó là sự thật.

Khi dữ liệu học lệch về một phía, đáp án cũng bị nghiêng. Nếu văn khắc của một vùng nào đó có nhiều, máy có thể đọc cả văn khắc vùng khác theo phong cách của vùng ấy. Máy nhiều khi không thể tự giải thích vì sao nó trả lời như vậy. Vì thế, nghiên cứu gần đây đang chuyển sang hướng đưa ra cả căn cứ cùng với kết quả [9].

Nhóm nghiên cứu đọc cuộn giấy cũng công bố danh sách những vấn đề còn lại. Ở những chỗ bị ép và nát, các lớp không tách khỏi nhau. Mô hình học từ một cuộn giấy đôi khi mất sức ở cuộn giấy khác. Việc nói rõ điều gì chưa làm được sẽ tạo ra bước tiếp theo [10].

Cũng cần có quy tắc xử lý chữ do máy đưa ra. Chữ đọc chắc chắn và chữ suy đoán phải được đánh dấu khác nhau. Cần ghi kèm mô hình nào đã đọc như vậy bằng tư liệu nào. Như thế, về sau nhà nghiên cứu khác mới có thể tính toán lại cùng cách đó.

Bản gốc vẫn là bản gốc

Dù phục dựng số tinh vi đến đâu, nó cũng không thể thay thế bản gốc. Giới học thuật xem dữ liệu quét không phải là bản gốc, mà là bản sao số. Việc chụp ảnh có lựa chọn của con người. Vì con người quyết định dùng ánh sáng nào và chụp từ góc nào. Khi lựa chọn đó thay đổi, hình ảnh kết quả cũng thay đổi.

Trong hiện vật thật vẫn còn thông tin mà máy quét không ghi lại được. Cấu trúc hạt đất trên bề mặt bảng đất sét là một ví dụ. Cách sắp xếp các lỗ nhỏ trên yếm rùa cũng chỉ có thể thấy ở hiện vật thật. Mực thấm vào trong giấy cói sâu bao nhiêu cũng chỉ đo được trên hiện vật thật [10].

Các kỹ thuật đo mới sẽ tiếp tục xuất hiện. Với hiện vật đã bị bỏ đi hôm nay, ta không thể dùng kỹ thuật đó nữa. Vì vậy cần tránh cắt hiện vật hoặc cố mở nó bằng sức. Mục đích của thư viện ảo không phải là vứt bỏ bản gốc. Mục đích là làm cho người ta đọc được mà ít chạm vào bản gốc hơn [2]. Khi số lần chạm vào hiện vật giảm, tuổi thọ của bản gốc tăng lên.

Tương lai cùng đọc

AI không phải là công cụ đẩy nhà sử học ra ngoài. Nghiên cứu Ithaca đã cho thấy điều này bằng con số. Khi Ithaca tự phục dựng chữ bị thiếu, độ chính xác là 62 phần trăm. Khi các nhà sử học làm một mình, con số là 25 phần trăm. Khi hai bên cùng làm, con số tăng lên 72 phần trăm [6].

Những con số này nói lên hai điều cùng lúc. Trong thí nghiệm này, máy một mình tốt hơn con người. Và khi làm cùng con người, kết quả còn cao hơn. Chỉ một bên thì không thể đi xa đến mức đó.

Máy nhanh chóng đưa ra nhiều ứng viên. Con người xét xem các ứng viên ấy có thể tồn tại trong thời đại đó hay không. Giới học thuật gọi cách con người và máy thay phiên kiểm tra như vậy là hợp tác có con người trong vòng lặp. Để vòng lặp này vận hành, phải tiếp tục có người biết đọc chữ cổ. Nếu không có người đọc, chữ mà máy rút ra sẽ trở thành một câu đố khác.

Việc dạy học cũng quan trọng như nghiên cứu. Cách đọc ngôn ngữ cổ đã được truyền từ người sang người. Nếu mắt xích đó đứt, sẽ không còn ai kiểm chứng chữ do máy rút ra. Thế hệ mới phải học cả chữ cổ và máy tính.

Vòng lặp này không chỉ thuộc về chuyên gia. Ở Nhật Bản đã có ứng dụng miễn phí Miwo đọc chữ thảo. Khi chụp ảnh, chữ ngày nay hiện lên trên màn hình. Người lần đầu thấy tài liệu cổ cũng có thể đọc dòng đầu tiên [5]. Việc cầm điện thoại thông minh trước tủ kính bảo tàng đôi khi trở thành điểm bắt đầu của nghiên cứu.

Việc phải làm phía trước vẫn còn nhiều. Ở Herculaneum vẫn còn các cuộn giấy bị niêm phong [11]. Ở Ukraine, thư viện và kho lưu trữ vẫn đang bị phá hủy [12]. Chữ cổ trên thế giới vẫn biến mất nhanh hơn được cứu lại.

Người đọc cuốn sách này cũng có việc có thể làm. Dữ liệu quét công khai thì bất cứ ai cũng có thể tải xuống để xem [10]. Học sinh cũng tham gia các cuộc thi ghép mảnh [11]. Ở nhiều nơi, công dân cùng tham gia chép lại tài liệu cổ. Mỗi người phụ trách một dòng thì dữ liệu cũng tích tụ lại.

Giữ gìn ghi chép không chỉ là chuyện xưa. Tư liệu chúng ta tạo ra hôm nay một ngày nào đó cũng trở thành tư liệu cổ. Nếu định dạng tệp thay đổi, tài liệu hiện nay cũng có thể không đọc được. Vì vậy, các thư viện tiếp tục chuyển tư liệu sang định dạng mới. Việc dựng thư viện ảo luôn đi cùng bàn tay quản lý.

Dù vậy, hướng đi đã được xác định. Cuộn giấy bị vùi dưới tro núi lửa 2.000 năm trước đã được đọc lại [1]. Phương pháp thu được từ một cuộn giấy trở thành điểm xuất phát cho nghiên cứu cuộn giấy tiếp theo [11]. Khi công bố dữ liệu và mã, nhà nghiên cứu ở nước khác cũng có thể đi cùng con đường [10]. Dù hiện vật một ngày nào đó trở về với đất, câu chữ bên trong vẫn còn. Thư viện không cháy đang được dựng lên từng dòng như vậy.

Tài liệu tham khảo

[1] Vesuvius Challenge. (2026, 6월 25일). An entire Herculaneum scroll has been read for the first time. scrollprize.org. https://scrollprize.org/firstscroll (truy cập ngày 3 tháng 9 năm 2026)

[2] Seales, W. B., Parker, C. S., Segal, M., Tov, E., Shor, P., & Porath, Y. (2016). From damage to discovery via virtual unwrapping: Reading the scroll from En-Gedi. Science Advances, 2(9), e1601247. https://doi.org/10.1126/sciadv.1601247

[3] Popović, M., Dhali, M. A., & Schomaker, L. (2021). Artificial intelligence based writer identification generates new evidence for the unknown scribes of the Dead Sea Scrolls exemplified by the Great Isaiah Scroll (1QIsaa). PLOS ONE, 16(4), e0249769. https://doi.org/10.1371/journal.pone.0249769

[4] Ma, Y. R. (2026). Towards Computational Chinese Paleography. arXiv:2601.06753. https://arxiv.org/abs/2601.06753 (truy cập ngày 3 tháng 9 năm 2026)

[5] ROIS-DS Center for Open Data in the Humanities. みを(miwo): AIくずし字認識アプリ. CODH. https://codh.rois.ac.jp/miwo/ (truy cập ngày 3 tháng 9 năm 2026)

[6] Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., & de Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. Nature, 603(7900), 280-283. https://doi.org/10.1038/s41586-022-04448-z

[7] Assael, Y., Sommerschield, T., Cooley, A. E., Shillingford, B., Pavlopoulos, J., Suresh, P., Herms, B., Grayston, J., Maynard, B., Dietrich, N., Wulgaert, R., Prag, J., Mullen, A., & Mohamed, S. (2025). Contextualizing ancient texts with generative neural networks. Nature. https://doi.org/10.1038/s41586-025-09292-5

[8] Braović, M., Krstinić, D., Štula, M., & Ivanda, A. (2024). A systematic review of computational approaches to deciphering Bronze Age Aegean and Cypriot scripts. Computational Linguistics, 50(2), 725-779. https://doi.org/10.1162/coli_a_00514

[9] Kutt, K., Sroka, E., Ishchuk, O., & do Valle Miranda, L. (2026). A three-stage neuro-symbolic recommendation pipeline for cultural heritage knowledge graphs. arXiv:2602.19711. https://arxiv.org/abs/2602.19711 (truy cập ngày 3 tháng 9 năm 2026)

[10] Vesuvius Challenge. (2026, 7월 10일). Open problems: Why reading every Herculaneum scroll is still a challenge. scrollprize.org. https://scrollprize.org/2026_open_problems (truy cập ngày 3 tháng 9 năm 2026)

[11] Vesuvius Challenge. (2026). Open Prizes. scrollprize.org. https://scrollprize.org/prizes (truy cập ngày 3 tháng 9 năm 2026)

[12] UNESCO. (2026). Damaged cultural sites in Ukraine verified by UNESCO. UNESCO. https://www.unesco.org/en/ukraine-war/damaged-cultural-sites (truy cập ngày 3 tháng 9 năm 2026)

Phụ lục. Nhật ký khảo cổ học AI của riêng tôi: cổng khám phá toàn cầu cho công chúng

A.1. Kho dữ liệu chính thức của Vesuvius Challenge

Có thể xem gì

Vesuvius Challenge là một cuộc thi quốc tế. Cuộc thi này tranh tài trong việc đọc các cuộn giấy bị cháy bằng máy tính. Nơi tìm thấy các cuộn giấy là Herculaneum. Đó là một thành phố La Mã cổ nằm cạnh núi lửa Vesuvius. Các cuộn giấy được làm bằng giấy cói. Giấy cói là loại giấy xưa được làm bằng cách ép một loài cỏ. Địa chỉ chính thức là scrollprize.org[1]. Trang này công khai dữ liệu dùng trong cuộc thi cho mọi người[2]. Không cần đăng ký thành viên và cũng không tốn tiền[2][5].

Ở phía trên màn hình đầu tiên có một hàng menu. Prizes, Problems, Tutorials & In-depth, Data đứng trước. Sau đó là Milestones, Community, Press, Donate. Tính đến tháng 9 năm 2026, 45 cuộn giấy và mảnh giấy cói đã được quét. Trong số đó, một cuộn đã được đọc từ đầu đến cuối[1].

Ngày 25 tháng 6 năm 2026, nhóm cuộc thi đã thông báo tin về cuộn giấy đó. Đó là cuộn giấy được gọi là PHerc. 1667[3]. Đây là trường hợp đầu tiên đọc hết một cuộn giấy mà không mở nó ra[3][4]. Chiều dài giấy cói được trải phẳng khoảng 1,4 mét. Sách cổ được viết bằng cách chia chữ thành các cột dọc. Mỗi cột như vậy được gọi là một cột văn bản. Trong cuộn giấy này, khoảng hai mươi hai cột viết bằng chữ Hy Lạp đã xuất hiện. Nội dung là bài viết về đạo đức của trường phái Khắc kỷ. Bài viết nói về bản tính con người, xung động và sự trưởng thành đạo đức. Người ta xem đây là văn bản khoảng thế kỷ 2 trước Công nguyên[3].

Không phải toàn bộ cuộn giấy còn nguyên vẹn. Phần còn lại hiện nay là lõi bên trong của cuộn giấy. Các lớp bên ngoài đã vỡ khi người ta cố mở bằng tay vào thế kỷ 19 và thập niên 1980. Ở những chỗ bề mặt bong ra, chữ cũng biến mất theo. Vì vậy văn bản đã đọc được có nhiều khoảng trống rải rác[3].

Việc quét được thực hiện tại Grenoble, Pháp. Đó là cơ sở bức xạ synchrotron châu Âu ESRF. Đây là một thiết bị nghiên cứu lớn tạo ra tia X rất sáng. Trong đó, nhóm đã dùng beamline BM18[4]. Beamline là đường thí nghiệm dẫn tia X đến và chiếu vào vật thể.

Trình tự làm thử tại nhà

Trước tiên, vào scrollprize.org và nhấn nút Get Started. Màn hình này hướng dẫn ba bước cho người mới đến lần đầu[5].

Bước đầu tiên là tìm mực. Người dùng tô phần mực nhìn thấy trên màn hình. Khi đó mô hình học từ dấu đánh dấu đó và dự đoán phần còn lại[5].

Bước thứ hai là mở cuộn ảo. Đây là việc thử mở cuộn giấy bằng máy tính. Ta lần theo một lớp giấy cói trong ảnh lát cắt CT. Ảnh lát cắt CT là ảnh chụp như thể cắt mỏng phần bên trong để xem. Ta trải phẳng lớp đó rồi kiểm tra kết quả[5].

Bước thứ ba là học toàn bộ quy trình. Ta thử nối tiếp việc quét, mở cuộn và đọc ở quy mô nhỏ[5].

Cần biết rằng trình tự trải nghiệm và trình tự phục hồi thực tế khác nhau. Trải nghiệm bắt đầu từ việc dễ là tìm mực. Phục hồi thực tế bắt đầu bằng quét. Sau đó mở cuộn giấy bằng phương pháp ảo. Việc đọc mực là bước cuối cùng[5].

Nếu muốn tải dữ liệu trực tiếp, hãy vào scrollprize.org/data. Dữ liệu nằm trong kho công khai s3://vesuvius-challenge-open-data/. Cũng có thể mở bằng trình duyệt web. Tệp ví dụ chỉ cách dùng nằm trên GitHub[2].

Người tham gia cũng có giải thưởng. Tính đến tháng 9 năm 2026, tổng tiền thưởng còn lại là 2,14 triệu đô la. Giải Grand Prize 2027 có quy mô 1 triệu đô la. Người đứng thứ nhất nhận 800.000 đô la. Hạn chót là ngày 25 tháng 6 năm 2027. Những cuộn giấy chưa tìm được chữ cũng có tiền thưởng. Chỉ cần đọc được mười chữ trong diện tích 4 xentimét vuông. Khi đó, mỗi cuộn giấy được thưởng 50.000 đô la. Cũng có Progress Prize trao hằng tháng. Bài nộp xuất sắc nhất được chọn trong tháng sẽ nhận 20.000 đô la[6].

Điểm cần chú ý

Dữ liệu được công khai theo điều kiện CC BY-NC 4.0. Không được dùng cho mục đích kiếm tiền. Khi dùng phải ghi nguồn. Một số tư liệu cũ có bài báo riêng để trích dẫn. Đó là các cuộn giấy số 1 đến số 4 và các mảnh số 1 đến số 6[2]. Khi đó cần ghi bài báo năm 2023 của Parsons và những người khác[2][7].

Cũng cần biết trước rằng kích thước tệp rất lớn. Toàn bộ dữ liệu quét ba chiều lên đến vài terabyte[2]. Để chạy học sâu cần có card đồ họa mạnh. Nếu máy tính ở nhà yếu, có thể bắt đầu bằng việc xem lướt dữ liệu bằng mắt.

Khi tham gia tranh giải, có điều kiện đi kèm. Phải gia nhập phòng Discord của cuộc thi. Mã nguồn phải được công khai theo điều kiện tự do như MIT. Không được công bố kết quả ra bên ngoài trước khi có thông báo trao giải[6].

Tài liệu tham khảo

[1] Vesuvius Challenge. (2026). Vesuvius Challenge: Reading the Herculaneum Scrolls with machine learning. scrollprize.org. https://scrollprize.org/ (truy cập ngày 3 tháng 9 năm 2026)

[2] Vesuvius Challenge. (2026). Data. scrollprize.org. https://scrollprize.org/data (truy cập ngày 3 tháng 9 năm 2026)

[3] Vesuvius Challenge. (2026, 6월 25일). An entire Herculaneum scroll has been read for the first time. scrollprize.org. https://scrollprize.org/firstscroll (truy cập ngày 3 tháng 9 năm 2026)

[4] Angelotti, G., Parsons, S., Nicolardi, F., et al. (2026). Complete virtual unwrapping and reading of a rolled Herculaneum papyrus. arXiv:2606.29085. https://arxiv.org/abs/2606.29085 (truy cập ngày 3 tháng 9 năm 2026)

[5] Vesuvius Challenge. (2026). Get Started. scrollprize.org. https://scrollprize.org/get_started (truy cập ngày 3 tháng 9 năm 2026)

[6] Vesuvius Challenge. (2026). Open Prizes. scrollprize.org. https://scrollprize.org/prizes (truy cập ngày 3 tháng 9 năm 2026)

[7] Parsons, S., Parker, C. S., Chapman, C., Hayashida, M., & Seales, W. B. (2023). EduceLab-Scrolls: Verifiable recovery of text from Herculaneum papyri using X-ray CT. arXiv:2304.02084. https://doi.org/10.48550/arXiv.2304.02084

A.2. Cổng eBL về văn học Babylon điện tử

Cấu trúc của cổng

eBL là viết tắt của electronic Babylonian Library. Có thể hiểu là Thư viện Babylon điện tử. Người Lưỡng Hà cổ viết bằng chữ hình nêm. Chữ hình nêm là loại chữ có hình như chiếc đinh, được ấn khắc lên bảng đất sét ướt. Cổng này tập hợp các mảnh bảng đất sét ấy vào một nơi. Đối tượng được tập hợp là các mảnh ghi văn học Babylon. Đây không phải nơi chứa mọi tư liệu chữ hình nêm trên thế giới[1].

Cổng đặt tại Đại học Ludwig Maximilian München ở Đức. Giáo sư Enrique Jiménez lãnh đạo dự án. Dự án bắt đầu vào tháng 4 năm 2018. Dự án được vận hành bằng khoản hỗ trợ từ Giải Sofia Kovalevskaya của Quỹ Alexander von Humboldt[2].

Địa chỉ là www.ebl.lmu.de. Có thể vào mà không cần đăng nhập và không mất phí. Bên trái màn hình có tám mục. eBL Project, Library, Corpus, Signs đứng trước. Sau đó là Akkadian Dictionary và Bibliography. Cuối cùng là News và Archaeology[1].

Library là nơi tập hợp từng mảnh tư liệu. Trước đây nó được gọi là Fragmentarium. Các nhà nghiên cứu eBL đã tìm ra khoảng 1.200 trường hợp ghép nối mảnh tại đây[1]. Corpus hiển thị văn bản theo từng tác phẩm. Sử thi Gilgamesh được chỉnh lý thành khoảng 3.150 dòng. Thần thoại sáng thế Enuma Elish có 1.096 dòng[3]. Bài báo dữ liệu xuất bản năm 2024 cho biết quy mô của cổng. Cổng này chứa khoảng 25.000 bảng đất sét. Bản chuyển tự có hơn 350.000 dòng[4]. Bản chuyển tự là văn bản chép chữ hình nêm sang chữ Latin.

Có thể làm gì trên màn hình

Khi nhấn Library, một ô tìm kiếm xuất hiện. Có thể tìm theo tên, năm và nhan đề. Cũng có tìm kiếm theo từ và tìm kiếm nâng cao. Nếu không biết tìm gì, hãy nhấn nút I'm feeling lucky. Khi đó một mảnh bất kỳ sẽ mở ra[5].

Khi mở một mảnh, màn hình được chia thành nhiều ô. Phía trên ghi cơ quan lưu giữ, kích thước và nơi phát hiện. Kiểu chữ và loại văn bản cũng hiện cùng. Ví dụ, K.5743 là hiện vật thuộc bộ sưu tập Kuyunjik của Bảo tàng Anh. Kuyunjik là ngọn đồi nơi từng có thành phố cổ Nineveh. Đây là tên gọi chung cho các di vật tìm được từ ngọn đồi đó. Kích thước là 3,81 xentimét nhân 2,54 xentimét. Kiểu chữ là kiểu Tân Assyria[6]. Đó là dạng chữ được dùng thời Đế quốc Tân Assyria.

Ở giữa, bản chuyển tự hiện cùng số dòng. Phía dưới có ảnh bảng đất sét. Trang cũng ghi mảnh đó được ghép với mảnh nào[6].

Trang cũng cho biết cách dùng khi trích dẫn văn bản. Ghi số mảnh và eBL edition. Sau đó thêm địa chỉ và ngày truy cập[1].

Tin mới và điểm cần chú ý

Có thể xem bản tin trong mục News. Số 22, ngày 28 tháng 5 năm 2026, là số mới nhất. Trong số này, cổng đã đưa ra bố cục màn hình mới. Số cơ quan lưu giữ cũng tăng lên ở nhiều nơi. Bảo tàng Chữ viết Thế giới Quốc gia ở Incheon cũng là một trong số đó. Hơn 3.100 ảnh của Bảo tàng Anh mới được đưa lên. Hơn 1.000 ảnh của Bảo tàng Peabody Yale cũng được đưa lên. Chức năng tự đăng ký từ mới vào từ điển cũng xuất hiện[7].

Có điều cần biết trước khi dùng. Bản chuyển tự trong Library không phải bản hoàn chỉnh mà là bản tạm thời. Nhóm nghiên cứu vẫn tiếp tục sửa. Văn bản tác phẩm được chỉnh lý trong Corpus cũng vậy. Đó là kết quả trung gian được ghép từ các mảnh đã thu thập đến nay. Nếu mảnh mới được ghép vào, số dòng và nội dung sẽ thay đổi. Vì vậy khi dùng trong bài tập ở trường hoặc bài viết, cần ghi kèm ngày truy cập[1].

Bản chuyển tự được công khai theo điều kiện CC BY-NC-SA 4.0. Ảnh có điều kiện khác. Nếu muốn dùng lại ảnh, phải xin phép riêng[1]. Bảo tàng sở hữu bảng đất sét là bên cho phép[1][4]. Trước hết cần đọc hướng dẫn gắn với chú thích ảnh ở phía dưới màn hình[1].

Tài liệu tham khảo

[1] electronic Babylonian Library. (2026). About: Library. ebl.lmu.de. https://www.ebl.lmu.de/about/library (truy cập ngày 3 tháng 9 năm 2026)

[2] Jiménez, E. (2023, 5월 25일). AI is helping us read ancient Mesopotamian literature. The Conversation. https://theconversation.com/ai-is-helping-us-read-ancient-mesopotamian-literature-205091 (truy cập ngày 3 tháng 9 năm 2026)

[3] electronic Babylonian Library. (2026). Corpus. ebl.lmu.de. https://www.ebl.lmu.de/corpus (truy cập ngày 3 tháng 9 năm 2026)

[4] Cobanoglu, Y., Laasonen, J., Simonjetz, F., Khait, I., Cohen, S., Földi, Z., Hätinen, A., Heinrich, A., Mitto, T., Rozzi, G., Sáenz, L., & Jiménez, E. (2024). Transliterated cuneiform tablets of the electronic Babylonian Library platform. Journal of Open Humanities Data, 10(1), 19. https://doi.org/10.5334/johd.148

[5] electronic Babylonian Library. (2026). Library. ebl.lmu.de. https://www.ebl.lmu.de/library (truy cập ngày 3 tháng 9 năm 2026)

[6] electronic Babylonian Library. (2026). K.5743, eBL edition. ebl.lmu.de. https://www.ebl.lmu.de/library/K.5743 (truy cập ngày 3 tháng 9 năm 2026)

[7] electronic Babylonian Library. (2026, 5월 28일). eBL Newsletter 22. ebl.lmu.de. https://www.ebl.lmu.de/news (truy cập ngày 3 tháng 9 năm 2026)

A.3. Hướng dẫn sử dụng Miwo

Đây là ứng dụng gì

Trong sách và văn bản cổ của Nhật Bản, chữ viết thảo được dùng nhiều. Loại chữ này được gọi là kuzushiji. Nó thường thấy trong sách in thời Edo và văn bản cổ[3]. Ngày nay, phần lớn người Nhật không đọc được loại chữ này.

Miwo là ứng dụng điện thoại thông minh chụp loại chữ thảo này bằng ảnh rồi chuyển thành chữ hiện nay. Trung tâm CODH, một trung tâm dùng chung dữ liệu mở nhân văn của Nhật Bản, đã tạo ra ứng dụng này[1]. Ứng dụng miễn phí[2]. Tên ứng dụng được lấy từ Miotsukushi, quyển thứ mười bốn của Truyện Genji. Miotsukushi là cột mốc chỉ đường thủy. Tên này có nghĩa là trở thành người chỉ đường trong biển văn bản cổ[3].

Ứng dụng ra mắt lần đầu vào ngày 30 tháng 8 năm 2021. Ngày 26 tháng 10 năm 2022, ứng dụng được nâng lên phiên bản 1.1. Khi đó, mô hình nhận dạng mới tên là Ruri được đưa vào. Ruri tìm chữ tốt ngay cả khi nền chữ có hoa văn hoặc màu sắc[1]. Việc huấn luyện dùng Classical Japanese Kuzushiji Dataset. Đây là bộ ảnh chữ cổ được tạo để phục vụ học máy. Dữ liệu này có hơn một triệu ký tự. Ứng dụng này nhận Giải Good Design năm 2022[3]. Tính đến tháng 9 năm 2026, số ảnh mà ứng dụng này đã nhận dạng vượt quá 3,5 triệu ảnh[1].

Trình tự cài đặt và sử dụng

Trước tiên, mở chợ ứng dụng. iPhone và iPad dùng App Store. Thiết bị Android dùng Google Play. Nhập みを hoặc miwo vào ô tìm kiếm[1]. Theo App Store, dung lượng ứng dụng khoảng 52 megabyte. iPhone và iPad cần iOS 11.0 trở lên[2].

Khi mở ứng dụng, màn hình chụp ảnh xuất hiện. Đưa văn bản cổ vào màn hình và nhấn nút chụp. Cũng có thể gọi ảnh đã chụp sẵn. Sau đó nhấn nút nhận dạng, AI sẽ đọc chữ[1].

Màn hình kết quả có nhiều chức năng. Trên màn hình có một thanh điều chỉnh. Khi kéo thanh này sang trái hoặc sang phải, hai màn hình chồng lên nhau. Có thể so sánh ảnh gốc và chữ đã đọc được đặt cạnh nhau. Cũng có chức năng nhấn vào chữ kana tiếng Nhật. Chức năng này cho biết chữ đó vốn xuất phát từ chữ Hán nào. Văn bản đã đọc có thể được chuyển sang ứng dụng khác bằng nút sao chép[1]. Cũng có thể lưu kết quả rồi mở lại sau[1][3].

Cần nhớ rằng internet phải được bật. Ảnh được gửi đến máy chủ CODH để nhận dạng. Ảnh đã gửi không được lưu lại trên máy chủ[1].

Tư liệu đọc tốt và tư liệu đọc không tốt

Nếu biết giới hạn trước khi dùng ứng dụng, sẽ ít thất vọng hơn. AI được huấn luyện chủ yếu bằng sách in thời Edo. Sách thời đó phần lớn là bản in khắc gỗ. Đó là sách được làm bằng cách khắc chữ lên ván gỗ rồi in lên giấy. Vì vậy ứng dụng đọc tốt sách in thời kỳ đó. Tư liệu thời kỳ khác hoặc văn bản viết tay có độ chính xác thấp hơn[3].

Tình trạng giấy cũng làm thay đổi kết quả. Những chỗ bị ố hoặc bị côn trùng ăn rất khó đọc. Hoa văn trên giấy, ánh sáng và bóng cũng gây cản trở[3]. Chụp dưới ánh sáng sáng và đều thì tốt hơn.

Cũng có đối tượng không đọc được. Chữ thảo khắc trên bia đá hoặc bảng hiệu hiện chưa được hỗ trợ[3].

Nhóm phát triển cho biết kết quả nhận dạng có thể có chỗ sai. Họ khẳng định nhận dạng chữ thảo bằng AI không thể trở nên hoàn hảo. Vì vậy, tư liệu quan trọng phải được con người kiểm tra lại[3].

Việc chép chữ cổ sang chữ hiện nay được gọi là phiên khắc. Có một nơi riêng để nhiều người cùng chia việc này. Đó là trang phiên khắc có sự tham gia của công dân tên là Minna de Honkoku. Dịch sang tiếng Việt là Phiên khắc của mọi người. Trang này do một tổ chức khác với nơi tạo ra Miwo vận hành. Tính chất của nó giống việc con người chỉnh lại văn bản mà ứng dụng đã đọc. Địa chỉ là honkoku.org. Đến nay, hơn 10.800 người đã tham gia. Số ký tự được chép lại đã vượt quá 48 triệu. Trang được làm mới lần thứ hai vào tháng 8 năm 2025. Phiên bản mới đang vận hành thử. Bảo tàng Quốc gia về Lịch sử và Dân tộc học Nhật Bản tham gia. Viện Nghiên cứu Động đất của Đại học Tokyo và nhóm nghiên cứu Đại học Kyoto cũng cùng vận hành[4].

Tài liệu tham khảo

[1] ROIS-DS 人文学オープンデータ共同利用センター. (2026). AIくずし字認識アプリ「みを（miwo）」. CODH. https://codh.rois.ac.jp/miwo/ (truy cập ngày 3 tháng 9 năm 2026)

[2] Clanuwat, T. (2022, 10월 27일). miwo (Version 1.1) [Mobile app]. App Store. https://apps.apple.com/jp/app/miwo/id1581794085 (truy cập ngày 3 tháng 9 năm 2026)

[3] ROIS-DS 人文学オープンデータ共同利用センター. (2026). みを（miwo）とは？. CODH. https://codh.rois.ac.jp/miwo/about/ (truy cập ngày 3 tháng 9 năm 2026)

[4] みんなで翻刻. (2026). みんなで翻刻 Minna de Honkoku. honkoku.org. https://honkoku.org/ (truy cập ngày 3 tháng 9 năm 2026)

Ủng hộ cuốn sách Thư viện AI này

Nếu cuốn sách này hữu ích với bạn, bạn có thể ủng hộ các bản dịch tiếp theo, các ấn bản công khai và những dự án tri thức mở về trí tuệ nhân tạo qua PayPal.

PayPal

https://paypal.me/kimkjkj

Quét mã QR hoặc mở đường dẫn ở trên.

Luật sư Kim Kyung-jin