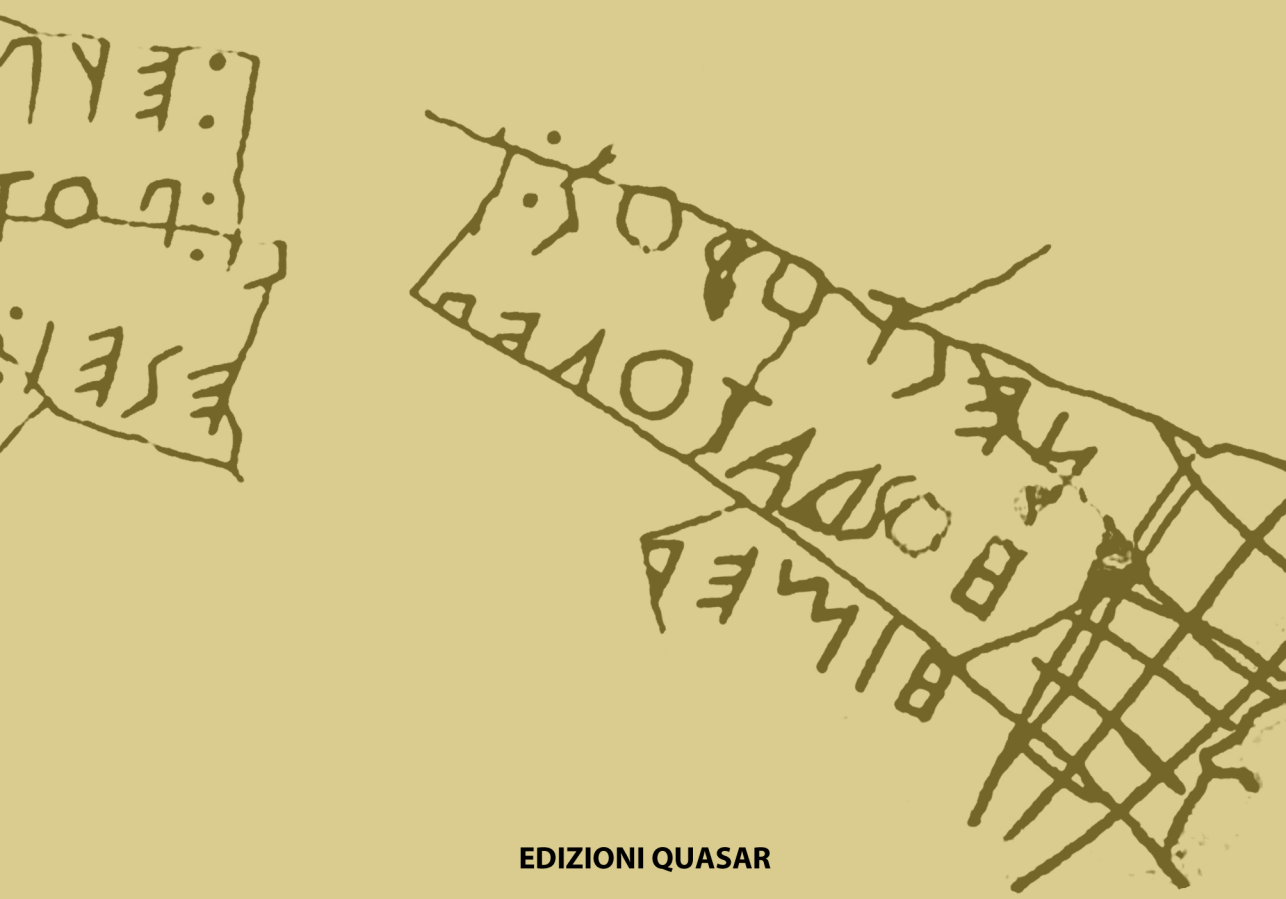
SAEG Papers 1

Seminari Avanzati di Epigrafia Greca 2025



EDIZIONI QUASAR

SAEG Papers 1

Seminari Avanzati di Epigrafia Greca **2025**

EDIZIONI QUASAR

«SAEG *Papers*. Seminari Avanzati di Epigrafia Greca»

Anno di fondazione: 2025

Maddalena Zunino, Valentina Mignosa, *Epigrafia greca e intelligenza artificiale. Il progetto* Lacunae *dell'Università di Udine*

Edizioni Quasar di Severino Tognon s.r.l.
via Ajaccio 41-43, 00198 Roma (Italia) http://www.edizioniquasar.it/

ISSN 3103-3431

Come citare l'articolo: M. Zunino, V. Mignosa, *Epigrafia greca e intelligenza artificiale. Il progetto Lacunae dell'Università di Udine*, in *SAEGPapers* 1, 2025, pp. 277–309.

Gli articoli pubblicati in SAEG *Papers* sono sottoposti a *referee* nel sistema a doppio cieco.

Epigrafia greca e intelligenza artificiale tra testo e documento. Il progetto *Lacunae* dell'Università di Udine

Maddalena Luisa Zunino, Valentina Mignosa

1. Il centro AI4CH dell'Università di Udine e il progetto *Lacunae*

Il progetto *Lacunae*, di cui presentiamo – brevemente e per la sola parte greca[1] – l'attività sin qui svolta, per concentrarci soprattutto sugli sviluppi attuali e i *desiderata* futuri, è una delle forme in cui il *Dipartimento di Studi Umanistici e del Patrimonio Culturale* (DIUM) dell'Università di Udine ha inteso contribuire al Piano Strategico 2022-25 del proprio Ateneo; nell'ambito del piano è infatti nato il centro di ricerca interdipartimentale *Artificial Intelligence for Cultural Heritage* (AI4CH),[2] dal quale il progetto è ospitato e vengono più ampiamente coltivate le nuove e interessanti occasioni di dialogo e collaborazione fra le discipline umanistiche e la scienza informatica, offerte dagli sviluppi più recenti dell'utilizzo dell'intelligenza artificiale applicata ai beni culturali.

Il progetto è dedicato all'integrazione delle lacune epigrafiche, un esito ben noto dei danni materiali che le vicissitudini attraversate nel tempo da un documento epigrafico arrecano al suo supporto, compromettendo la leggibilità di quello stesso documento, spesso irreparabilmente: proprio come nel caso di qualunque altro aspetto relativo a una testimonianza epigrafica, è

[1] Questo lavoro, a firma congiunta ed esplicita di Maddalena Luisa Zunino (§§1, 2, 5 e 7) e Valentina Mignosa (§§ 3, 4, 6 e 8), costituisce una versione ampliata del nostro intervento alla IX edizione del *SAEG* (Roma, 8-10 gennaio 2025), al quale hanno idealmente partecipato anche Silvia Zottin e Axel De Nardin che, nel solco delle attività dell'*Artificial Vision and Machine Learning Lab* del *Dipartimento di Scienze Informatiche, Matematiche e Fisiche* (DMIF) dell'Università di Udine (https://sites.google.com/view/avml-lab/home) di cui fanno parte, stanno curando la componente informatica del progetto relativa alla *Computer Vision*, su cui l'intervento si è appunto concentrato, e che, insieme a Gian Luca Foresti che li guida, cogliamo nuovamente l'occasione di ringraziare. A Roberta Fabiani e Giulio Vallarino va ancora una volta la nostra affettuosa gratitudine per averci accolto al Seminario; un sentito e doveroso ringraziamento vogliamo infine rivolgere agli anonimi Referees che hanno letto queste pagine e voluto anche offrire interessanti spunti di riflessione.

[2] La direzione del Centro (https://ai4ch.uniud.it), che vede il DMIF a fianco del DIUM, è di Gian Luca Foresti.

immediatamente chiaro che anch'esso deve essere affrontato tanto sul piano testuale quanto su quello, appunto, materiale.[3] Ed è in entrambe le direzioni che si è sin qui mosso il nostro progetto, coltivando da ultima l'ambizione di farle possibilmente confluire nella creazione di uno strumento digitale da mettere a disposizione della comunità di quanti si occupano appunto di documenti epigrafici – innanzitutto in lingua greca ma in progresso di tempo, auspicabilmente, anche in altre lingue antiche.

È ben noto – in particolare proprio ai grecisti dediti all'epigrafia – che qualunque strumento intenda 'parlare' (anche) il greco che ha preceduto quella sorta di astrazione che gli antichi hanno comunque definito κοινὴ διάλεκτος dovrà in realtà parlare una pluralità di dialetti che, pur greci, possono differenziarsi sensibilmente gli uni dagli altri, anche quando appartengano allo stesso gruppo:[4] l'ostinata assenza di aspirazione nel dialetto dorico di Creta (per questo definito psilotico severo) si oppone ad esempio alla sua presenza nell'altrettanto dorico dialetto di Thera, che esprime inoltre con il segno Ξ, per aggiungere alla varietà dialettale quella alfabetica, il suono all'inizio del nome di quel re degli dei[5] che a Rodi – il cui dialetto nuovamente dorico non fa tuttavia uso, ad esempio, del *digamma* (Ϝ) – è invece scritto Δεύς.[6]

Non soltanto, infatti, relativamente al dialetto e alla sua peculiare declinazione epicorica le produzioni epigrafiche delle differenti zone della Grecia si distinguono l'una dall'altra; esse si caratterizzano anche per l'utilizzo di una varietà altrettanto notevole di strumenti alfabetici: senza che sia in alcun modo possibile stabilire una qualche corrispondenza regolare fra un certo dialetto e un determinato alfabeto, quella varietà, ancora ripartita dai moderni per comodità

[3] L'interesse esplicito per la materialità dei documenti antichi è uno sviluppo degli studi relativamente recente: Hoogendijk, Gompel 2018; Petrovic, Petrovic, Thomas 2019; Angliker, Bultrighini 2023.

[4] Ossia a uno dei grandi macrogruppi della classificazione canonica moderna: nordoccidentale, dorico, ionico-attico, eolico e arcado-cipriota. Sui dialetti greci in generale, quali testimoniati dai documenti epigrafici, ancora fondamentale, nonostante il tempo trascorso e i nuovi rinvenimenti, Buck 1955; per un aggiornamento si può innanzitutto fare riferimento alle voci di Giannakis 2014. Fra i progetti editoriali in corso, la serie *Paradeigmata*: Nieto Izquierdo 2025.

[5] Il segno è utilizzato per l'iniziale del nome di Zeus in due casi e per quella di un antroponimo in un caso: Inglese 2008, nrr. 1-4, 28 (= *IG* XII 3, 350 a-353, 1313) e 95; Domínguez Casado 2014, pp. 92-94, in cui si discutono anche le principali deviazioni dell'alfabeto di Thera rispetto a quello di Creta, dal quale il primo senz'altro deriva e nel quale, a sua volta, «sur le plan graphique, Z, T, TT, Δ sont les transcriptions successives du groupe **dj*» (Bile, *Dialecte crétois*, p. 203).

[6] Wachter, *Non-Attic Vase Inscr.*, pp. 222-223, DOH 3.

di classificazione in quattro grandi gruppi,[7] è in realtà così ricca – quanto alle forme assunte dalle lettere e alle loro differenti combinazioni nei singoli alfabeti – da creare l'impressione, nient'affatto lontana dal vero, che ogni singola comunità politica greca abbia il suo specifico alfabeto.[8]

Se quanto appena affermato ci ricorda nuovamente che un documento epigrafico (e in particolare, forse, un documento epigrafico greco di età arcaico-classica) è al tempo stesso suono e segno, è pur vero che le particolari declinazioni epicoriche di questo connubio si rivelano spesso fondamentali, anche a fronte di testi frammentari, per individuare con buona approssimazione l'area di provenienza di un determinato documento e/o il periodo in cui esso è stato prodotto. Al tempo stesso, tuttavia, una così accentuata differenziazione in ambito tanto dialettale quanto alfabetico (come già detto, l'uno dall'altro il più delle volte indipendenti) finisce per assottigliare, in qualche caso senza rimedio,[9] i *corpora* delle iscrizioni che possano essere in modo ragionevole, ossia scientificamente fondato, raggruppate e considerate insieme: non c'è alcun bisogno di sottolineare che in epoca arcaico-classica nessuna comunità politica greca ha una produzione epigrafica che sul piano della quantità possa davvero paragonarsi a quella ateniese, che rimane inevitabilmente la più ricca.[10] E proprio questo – il numero perlopiù esiguo di iscrizioni accomunate dallo stesso dialetto (magari in una sua manifestazione epicorica) o dallo stesso alfabeto – rappresenta la difficoltà maggiore nel momento in cui si intendano utilizzare gli strumenti informatici che, innanzitutto nel caso di quelli destinati all'elaborazione del linguaggio naturale (*Natural Language Processing* [NLP]), sono inesorabilmente avidi di dati.

Come vedremo, quella difficoltà è emersa in tutta la sua evidenza nel momento in cui il nostro progetto – dedicato principalmente, anche in

[7] Contraddistinti dai colori verde, rosso, azzurro e blu secondo la classificazione di Kirchhoff 1887, Karte.

[8] Il quale alfabeto, secondo l'approccio degli studi più recenti, sembra sempre meno il risultato accidentale della trasmissione dei segni e sempre più il frutto di una (consapevole) standardizzazione dettata dalle pratiche epigrafiche locali e dal loro particolare contesto sociale e politico: vd. ad esempio Steele 2020.

[9] È ad esempio stato il caso, nei nostri esperimenti, delle iscrizioni appartenenti al gruppo dialettale arcado-cipriota: vd. *infra*.

[10] Costituiscono un'eccezione solo apparente i grandi santuari panellenici, che per loro stessa natura ospitano, anche e soprattutto, documenti di produzione non locale, mentre è forse più promettente la produzione epigrafica dell'isola di Creta, senz'altro standardizzata sul piano delle tipologie testuali, nonostante le differenze, più e meno evidenti, fra gli alfabeti delle singole comunità locali: all'isola, non a caso, abbiamo guardato per i primi esperimenti di *Computer Vision* (vd. *infra*).

rispondenza agli interessi di chi scrive, proprio ai testi epigrafici greci di età arcaico-classica, caratterizzati appunto da una notevole varietà tanto dialettale quanto alfabetica – ha dovuto necessariamente misurarsi, innanzitutto, con i principali strumenti informatici che sono già a disposizione della comunità scientifica, anche per affinare ed eventualmente ridefinire le proprie legittime aspettative e i propri obiettivi.

2. Intelligenza artificiale e testi antichi: lo stato dell'arte

Che lo studio delle lingue antiche stia in misura crescente esplorando e sfruttando le potenzialità dell'applicazione delle tecniche proprie dell'apprendimento automatico (*Machine Learning* [ML]) e della visione artificiale (*Computer Vision* [CV]) è dimostrato anche dal fatto che, ai principali strumenti a disposizione della comunità scientifica già censiti da un'accurata e recente ricognizione dedicata proprio a questo tema,[11] ne vanno ormai aggiunti almeno due. L'uno è nato anche per suggerire, a professionisti e ricercatori, «unexpected connections between texts from different fields, powerfully supporting interdisciplinary studies and the discovery of new ideas»,[12] mentre il secondo (a sua volta dedicato all'epigrafia latina) è gemello di uno già da tempo noto proprio agli studiosi di epigrafia greca;[13] è invece ancora privo di un'interfaccia utente *Logion*, lo strumento sviluppato dall'Università di Princeton allo scopo di aiutare «the restoration and elucidation of premodern Greek texts».[14]

Gli uni e l'altro – sia pure in modi diversi e diversamente evidenti – confermano che la parte del leone nel rapporto tra lo studio dei documenti antichi e le tecniche digitali è giocata appunto dalle caratteristiche più propriamente

[11] Sommerschield *et alii* 2023.

[12] «Humanitext Antiqua is an innovative conversational platform developed to explore the vast world of Western Classics using cutting-edge AI. At its core, it combines Large Language Models (LLMs) with a trusted academic database, operating on Retrieval-Augmented Generation (RAG) technology»: così dalla pagina di presentazione della *chatbot* (https://humanitext.ai/apps/antiqua), che è parte del progetto *Humanitext. Shaping the Future of Humanities with AI. A Textual Research & Analysis Platform — designed to accelerate your research workflow* (https://humanitext.ai), curato da N. Iwata (Università di Nagoya), I. Tanaka (Università J.F. Oberlin, Tokyo) e J. Ogawa.

[13] Fratello di *Ithaca*, di cui non mancheremo di parlare, anche *Aeneas* è dedicato all'integrazione delle lacune dei testi epigrafici (latini), nonché alla loro datazione e collocazione geografica (Assael, Sommerschield, Cooley *et alii* 2025). Entrambi gli strumenti sono ora raggiungibili dalla pagina web *Predicting the Past. Contextualising, restoring, and attributing ancient texts* (https://predictingthepast.com).

[14] https://logionproject.squarespace.com. Per la prova di fattibilità del modello (comunque già utilizzabile) è stata scelta l'opera dell'autore bizantino Michele Psello.

testuali e linguistiche dei primi, siano esse riferite tanto agli autori quanto alle opere: che si tratti di ricostruire la storia della trasmissione di un testo, di identificare l'autore di un'opera di dubbia attribuzione, di condurre appunto ricerche intertestuali, stabilendo relazioni fra autori e opere o, più ampiamente, individuando temi e motivi ricorrenti, gli strumenti utilizzati sono dunque quelli propri del NLP.[15] Nel loro progressivo affinarsi, l'introduzione di un modello di linguaggio basato su un'architettura trasformativa come *BERT* ha costituito una svolta, in grado di incrementare significativamente i risultati ed elevare le aspettative, anche nel campo di nostro interesse dell'integrazione testuale.[16]

Di fatto, è uno dei compiti che *BERT* è in grado di portare a termine: testato per la prima volta, quanto alle lingue antiche, su quella latina, il modello è in grado di suggerire termini appropriati in base al contesto della frase, rivelandosi utile nei casi di varianti fra cui scegliere o guasti della tradizione. Tale abilità è stata tuttavia affinata allenando il modello con tutti, o quasi, i testi in latino gratuitamente disponibili in formato digitale, da quelli contenuti nella *Perseus Digital Library*[17] alle pagine di *Vicipaedia*;[18] questo bisogno quasi insaziabile di dati, che è alla base della buona performance di *LatinBERT*[19], spiega perché il modello 'padre' abbia avuto successo anche quando applicato alla ricchissima documentazione annalistica prodotta dalla dinastia coreana Joseon (1392-1897),[20] ma spiega anche perché esso non abbia rappresentato un'opzione per chi ha voluto confrontarsi, ad esempio, con le lacune dei testi greci del II millennio, in scrittura Lineare B e dialetto miceneo.[21]

---

[15] Quanto alle tecniche di CV, vd. *infra*.

[16] Devlin *et alii* 2019. Vd. anche *infra* e nt. 19.

[17] https://www.perseus.tufts.edu/hopper.

[18] https://la.wikipedia.org/wiki/Vicipaedia:Pagina_prima.

[19] Bamman, Burns 2020. Quanto ad *Ancient GreekBERT*, di cui riparleremo: Singh, Rutten, Lefever 2021. Sebbene ulteriori modelli, comunque derivati da *BERT*, siano stati successivamente preparati e affinati per supportare il lavoro dei filologi classici (ad esempio, nel campo degli studi dedicati all'intertestualità: Riemenschneider, Frank 2023a), soltanto di *GREτA* e *PHILTA* è stata messa alla prova, con risultati non particolarmente incoraggianti, la capacità predittiva (di indovinare il termine che in una frase sia stato mascherato), relativamente a un *dataset* preparato *ad hoc* e decisamente specifico: una lista di rapporti di parentela fra dei ed eroi tratta dalla *Teogonia* di Esiodo, in base alla quale il modello è stato appunto chiamato a completare frasi come Τηλέμαχος ὁ τοῦ <…> παῖς (Riemenschneider, Frank 2023b). Tutto considerato, continuare a fare riferimento anche ad *Ancient GreekBERT* per testarne la capacità di colmare le lacune epigrafiche in testi non standardizzati è parso opportuno e ha restituito risultati del tutto comparabili a quelli di altri modelli: vd. *infra*.

[20] Kang *et alii* 2021.

[21] Papavassileiou, Kosmopoulos, Owens 2023.

In questo caso, il *dataset* era decisamente ridotto (le tavolette della serie D di Cnosso, ca. 1100 esemplari), sebbene la standardizzazione dei testi scelti abbia permesso la creazione di dati sintetici, che ne hanno aumentato il numero; all'origine dei suggerimenti – che, per lacune che interessano soprattutto i margini delle tavolette e sono pertanto di lunghezza quasi mai precisabile, vengono forniti a ogni interrogazione una singola sillaba alla volta, sulla base di quella adiacente (precedente o successiva che sia) – è una rete neurale ricorrente (*Recurrent Neural Network* [RNN]) e bidirezionale, ossia una forma di apprendimento profondo (*Deep Learning*) precedente la tecnologia utilizzata da *BERT*.

L'indifferenza di quest'ultimo (e dei suoi derivati) per la lunghezza, espressa in caratteri, dei suoi suggerimenti non costituisce in alcun modo un problema finché il modello, nato con questo scopo, viene applicato ai testi della letteratura – per quanto ci riguarda, latina o greca che sia –, che di certo non ci sono giunti nella loro versione (scritta) originale; può tuttavia rivelarsi d'ostacolo, come brevemente vedremo anche in seguito, quando quello stesso modello venga piuttosto messo alla prova delle lacune epigrafiche, che nella maggioranza dei casi sono spazi di dimensione nota, da riempire dunque con esattezza.[22] Né la possibilità di procedere, per dir così, empiricamente un segno per volta può davvero ritenersi la soluzione più appropriata per testi in cui i segni esprimono singole lettere e non sillabe e che possono essere afflitti da lacune anche di notevole lunghezza.

Il problema sembra essere stato affrontato e risolto da due architetture trasformative, dedicate l'una alle iscrizioni greche, l'altra alle tavolette in scrittura cuneiforme e lingua accadica, che sono in grado di suggerire integrazioni per un numero prestabilito di caratteri, anche decisamente elevato. Lo strumento elaborato da K. Lazar, un modello multilingue affinato specificatamente per quella accadica, costruisce infatti i suoi suggerimenti sillabogramma per sillabogramma, sino a raggiungere il numero di volta in volta già 'comunicato' all'algoritmo, in quanto predeterminato da chi ha proceduto alla traslitterazione dei testi utilizzati per l'esperimento.[23] In modo non troppo diverso, da questo punto di vista, procede a sua volta *Ithaca*, un'architettura trasformativa che rappresenta l'evoluzione del RNN già denominato *Pythia* (il primo ad applicare

[22] Laddove con esattezza si intende l'ineludibile rispetto, nella formulazione della proposta o delle proposte di integrazione, dello spazio disponibile – ovviamente, nei casi in cui questo sia noto.

[23] Lazar *et alii* 2021.

tecniche di apprendimento profondo al problema dell'integrazione dei testi) e che, in grado anche di collocare un documento epigrafico in un'area geografica e in un periodo cronologico, è stato ora affiancato, come accennato all'inizio, dal gemello *Aeneas*, dedicato invece alle iscrizioni in lingua latina.[24]

È senz'altro opportuno sottolineare che nell'uno e nell'altro caso – accadico e greco, ora affiancato dal latino – si tratta di soluzioni che, come già detto, ancora una volta affrontano il problema della lacuna sul piano puramente testuale: se i segni della scrittura cuneiforme non vengono proposti all'intelligenza artificiale, se non nella forma delle sillabe (e parole) in cui li ha interpretati e traslitterati l'essere umano,[25] che ha inoltre stabilito, come già detto, quante di quelle stesse sillabe sono contenute dalla lacuna, è l'utente che utilizza *Ithaca* (e ora anche *Aeneas*) a stabilire di quanti caratteri (alfabetici) deve essere costituito il suggerimento elaborato dall'algoritmo, tenendo inoltre conto, in quel calcolo, anche degli spazi che separano una parola dall'altra, quando ritenga che la lacuna sia grande abbastanza da contenere una parola intera o più parole; in altri termini, per limitarci all'ambito greco di nostro interesse, quella che è forse la caratteristica materiale più peculiare e tipica di un testo epigrafico, ossia l'essere redatto in *scriptio continua*, deve essere già stata affrontata e, per dir così, neutralizzata dall'utente, nel momento in cui interroga lo strumento.[26] Ha di conseguenza voluto ricreare una situazione più realistica E. Cullhed che, nel mettere a punto il modello *Llama* 3.1 8B Instruct di Meta, per il ripristino dei caratteri perduti nelle iscrizioni greche e nei papiri documentari, lo ha allenato a contare soltanto le lettere mancanti (sinora per un massimo di 10), il cui numero è anche in questo caso calcolato dall'utente. Tuttavia, all'algoritmo viene nuovamente sottoposto un testo già interpretato, in cui le singole parole che lo compongono sono state individuate e separate tra loro dagli spazi bianchi; frutto di analoga interpretazione è a sua volta la risposta dell'algoritmo.[27]

[24] Vd. *supra* nt. 13 e: Assael, Sommerschield, Prag 2019 (*Pythia*); Assael, Sommerschield, Shillingford *et alii* 2022 (*Ithaca*).

[25] Così è, del resto, anche nel caso dell'algoritmo dedicato all'integrazione dei testi micenei (vd. *supra* nt. 21).

[26] Il problema del calcolo da parte dell'utente sembra essere meno evidente nel caso delle tavolette in cuneiforme, che sono prodotti di scribi professionisti con impaginato regolare o quasi; tra le iscrizioni greche, si ha tuttavia qualcosa di simile soltanto nel caso di quelle a impaginato stoichedico o quasi stoichedico (vd. anche *infra*).

[27] Cullhed 2025 (il modello è ancora in fase sperimentale).

Se al momento sembrano perciò confrontarsi direttamente con il problema della *scriptio continua* gli studiosi che si occupano delle lingue (orientali) che a tutt'oggi la utilizzano,[28] *BERT*, o meglio *Ancient GreekBERT*, e *Ithaca* costituiscono gli unici strumenti sinora realizzati (anche) per offrire suggerimenti testuali appropriati in greco antico e hanno di conseguenza rappresentato il punto di partenza pressoché obbligato del nostro progetto e dei suoi primi esperimenti, che hanno dunque rivolto la loro attenzione, ancora una volta, all'aspetto testuale e linguistico di una lacuna epigrafica. A quegli strumenti abbiamo inoltre voluto affiancare, come vedremo, anche uno di quei modelli linguistici di grandi dimensioni (*Large Language Model* [LLM]) che, sotto forma di *ChatGPT*, *Gemini* o *CoPilot* o altri, sempre più frequentemente si trova a interrogare chi è dietro lo schermo di un computer, ormai anche nel caso in cui, come già accennato all'inizio di questo paragrafo, la conversazione fra uomo e macchina riguardi i testi delle letterature antiche greca e latina.[29]

3. I testi greci di epoca arcaico-classica alla prova di *Ithaca* e *BERT*: esperimenti e risultati

Come già accennato, la varietà dialettale che caratterizza le iscrizioni greche di epoca arcaico-classica – varietà tanto ampia quanto fortemente localizzata – pone sfide specifiche nel momento in cui si intendono applicare ad essa strumenti nati per l'elaborazione automatica del linguaggio.

I modelli NLP attualmente disponibili, pur se raffinati, si basano infatti su un principio tanto semplice quanto ineludibile: per imparare a completare un testo frammentario, il sistema ha bisogno di un numero significativo di esempi da cui apprendere. Nel caso dei documenti del periodo arcaico, però, non disponiamo di dati così copiosi.

Al fine di esplorare questa tensione tra affidabilità del modello e addestramento su un campione limitato di testi, abbiamo predisposto un primo ciclo di esperimenti su testi epigrafici greci di epoca arcaico-classica utilizzando due strumenti di riferimento: *Ancient GreekBERT*, modello trasformativo adattato al greco antico, originariamente su base letteraria (https://huggingface.co/pranaydeeps/Ancient-Greek-BERT), e *Ithaca*, un'architettura informatica

[28] Widiarti, Pulungan 2020; Butskhrikidze 2021; Fujii *et alii* 2023.

[29] Vd. *supra* e nt. 12.

sviluppata specificamente per testi epigrafici greci, in grado di proporre integrazioni testuali, ipotesi sulla cronologia e sulla provenienza geografica di un testo (https://predictingthepast.com/ithaca).[30]

Per i nostri esperimenti abbiamo selezionato le iscrizioni arcaiche greche presenti nel *corpus* digitale PHI (*Packard Humanities Institute*: https://inscriptions.packhum.org), basandoci sul *dataset* preparato per *Ithaca*.[31] Abbiamo selezionato le iscrizioni databili tra il 750 e il 450 a.C., ottenendo una selezione iniziale di 3.281 documenti. Successivamente, sono state escluse le iscrizioni duplicate, quelle troppo brevi, composte da meno di 50 caratteri, e quelle troppo lunghe, di oltre 768 caratteri, soglie che corrispondono ai limiti richiesti dal modello *Ithaca*. Il risultato finale è un *corpus* composto da 930 testi, per un totale di circa 18.000 parole.

All'interno del *dataset* così selezionato, abbiamo successivamente inserito alcune informazioni aggiuntive (metadati), relative al dialetto di ciascuna iscrizione. La prima operazione ha riguardato la catalogazione di ogni testo secondo la sua specifica appartenenza dialettale, cercando di raggiungere il massimo livello di dettaglio possibile. Questa classificazione così particolareggiata ha però evidenziato immediatamente un limite strutturale: per molte varietà dialettali, il numero di iscrizioni disponibili risultava troppo esiguo per consentire un utilizzo statisticamente significativo nei test con i diversi algoritmi. Per ovviare a questo problema, si è quindi optato per una classificazione su base macro-dialettale, che permettesse di aggregare i dati in gruppi più ampi e comparabili (i cinque gruppi del secondo livello della tabella di fig. 1). È stato escluso dal test il gruppo arcado-cipriota perché il numero di documenti, anche considerando l'intero macrogruppo, è inferiore alla quota minima funzionale per l'algoritmo.

I testi del *dataset* sono stati 'normalizzati' e ripuliti secondo un *iter* che riprende con alcune modifiche il lavoro di normalizzazione eseguito per la creazione del *dataset* di *Ithaca*. Questa operazione consiste in un insieme di passaggi

[30] Il *team* informatico che ci ha supportato e ancora ci affianca per gli esperimenti nell'ambito del NLP è composto da Alessandro Locaputo, Andrea Brunello, Nicola Saccomanno e Giuseppe Serra, che ringraziamo per la collaborazione e il lavoro di ricerca; con loro, stiamo attendendo alla stesura del lavoro destinato a illustrare nel dettaglio i risultati degli esperimenti sin qui svolti e ancora in corso di svolgimento.

[31] Vd. *supra* nt. 24.

DIALETTI GRECI	GRECO OCCIDENTALE	GRUPPO DORICO	KOINE SICILIANA
			CRETESE
			TEREO E CIRENEO
			COO
			RODIO
			ARGOLICO ORIENTALE (GOLFO SARONICO)
			ARGOLICO OCCIDENTALE
			CORINZIO
			MEGARESE
			MESSENICO
			ACHEO
			LACONICO ed ERACLEOTA
		GRUPPO NORDOCCIDENTALE	KOINE NORDOCCIDENTALE
			ELEO
			LOCRESE
			FOCESE
	GRECO ORIENTALE	GRUPPO EOLICO	BEOTICO
			TESSALICO
			LESBIO
		GRUPPO ARCADO-CIPRIOTA	PANFILIO
			CIPRIOTA
			ARCADICO
		GRUPPO IONICO-ATTICO	IONICO OCCIDENTALE o EUBOICO
			IONICO INSULARE
			IONICO CENTRALE
			IONICO ORIENTALE
			ATTICO

1 - Tassonomia dei dialetti. Livelli di suddivisione dal più dettagliato (dx) ai raggruppamenti macro-dialettali (sx).

pensati per rendere le iscrizioni raccolte in PHI più leggibili da parte degli algoritmi linguistici, attraverso la rimozione o la sostituzione sistematica di elementi considerati 'rumorosi'[32] o non standardizzati. Tra questi figurano, nel nostro caso, gli spiriti, gli accenti, la punteggiatura, la lettera ϝ (*digamma*), perché poco attestata e ambigua dato l'esito fonetico che può avere in epoca classica, e alcune convenzioni grafiche moderne come l'uso della lettera *h* per rappresentare il segno di aspirazione. Ulteriori interventi hanno riguardato l'uniformazione delle lettere, tutte minuscole, la rimozione di numerazioni editoriali, simboli non alfabetici, numeri greci, abbreviazioni e la standardizzazione di alcune lettere in funzione delle convenzioni epigrafiche. Questa operazione è fondamentale per poter sottoporre i testi a modelli di linguaggio, che hanno bisogno di coerenza nella struttura dei dati per offrire risultati validi. Solo una volta concluso questo processo è stato possibile procedere con l'analisi e la valutazione delle prestazioni dei modelli applicati.

Una volta preparato e standardizzato il nostro *dataset* abbiamo sottoposto il materiale a una duplice sperimentazione: prima con *Ithaca*, poi con *Ancient GreekBERT*. L'esperimento, in entrambi i casi, ha comportato il mascheramento di singole lettere/parole all'interno dei testi, simulando la presenza di una lacuna. I modelli sono stati chiamati a 'indovinare' le lettere/parole mancanti – non sulla base di una probabilità generica, ma in funzione del contesto linguistico circostante la lacuna stessa.

È opportuno ricordare nuovamente che *Ithaca*, sviluppato con l'obiettivo specifico di ricostruire testi epigrafici frammentari, si basa su un meccanismo che richiede all'utente di indicare *a priori* la lunghezza dell'integrazione desiderata, in termini di caratteri (inclusi gli spazi tra le parole). Questa caratteristica può risultare utile quando si lavora su epigrafi le cui lacune contengono un numero fisso e certo di lettere, ovvero su documenti a impaginazione

[32] In ambito NLP, per 'rumore' si intendono tutte quelle componenti di un testo che possono interferire con l'elaborazione automatica del linguaggio da parte di un algoritmo. Si tratta spesso di elementi non standardizzati o non previsti dal modello, come segni editoriali, caratteri non alfabetici, errori di trascrizione, varianti ortografiche, simboli rari o ridondanze grafiche. Anche accenti, segni di aspirazione e lettere di utilizzo (geograficamente o cronologicamente) limitato come il *digamma*, se non attesi dal sistema, possono essere considerati 'rumorosi'. In ambiti specifici, tuttavia, ciò che per l'algoritmo costituisce 'rumore' può rappresentare un'informazione significativa: elementi grafici come lo stesso *digamma* o il segno di aspirazione possono infatti fungere da marcatori cronologici o geografici.

stoichedica;[33] risulta invece limitante e solleva problemi metodologici quando la larghezza media delle lettere o lo spazio tra una lettera e l'altra è variabile, come spesso accade nei testi oggetto del nostro interesse.

I risultati ottenuti sul *corpus* arcaico con *Ithaca* sono stati significativamente meno soddisfacenti rispetto a quelli riportati dagli stessi sviluppatori del modello su tutto il *corpus* PHI. Una possibile spiegazione di questo risultato risiede nella distribuzione cronologica delle iscrizioni dell'intero *dataset* PHI, utilizzato per l'addestramento di *Ithaca*, caratterizzato soprattutto da documenti di epoca ellenistica e romana (con un picco attorno al 160 d.C.). Le iscrizioni da noi impiegate, tutte databili tra il 750 e il 450 a.C., risultano dunque ampiamente minoritarie nel materiale su cui il modello è stato formato. È quindi comprensibile che le sue prestazioni sui testi arcaici siano meno efficaci, proprio perché si tratta di un'epoca, e dunque di una varietà linguistica e dialettale, che l'algoritmo ha incontrato raramente durante il proprio processo di apprendimento.

Successivamente, abbiamo testato *Ancient GreekBERT*.[34] A differenza di *Ithaca*, che si concentra sull'integrazione di porzioni mancanti di testo in modo 'guidato', *BERT* si comporta come un lettore 'predittivo': suggerisce parole o gruppi di lettere in base all'ambiente linguistico che le circonda. In questo caso, i risultati non sono stati uniformi: il primo esperimento ha restituito esiti nettamente insoddisfacenti. Le prestazioni migliorano quando il modello viene riaddestrato sull'intero dataset di *Ithaca* e risultano decisamente comparabili a quelle dello stesso *Ithaca* nel caso di un addestramento limitato ai soli testi arcaici.[35]

Complessivamente, i risultati si sono rivelati piuttosto deludenti: entrambi i modelli hanno mostrato una capacità limitata di predire correttamente i caratteri mancanti, soprattutto in quelle iscrizioni che si discostano dai registri e dai lessici più standardizzati della prosa letteraria o del dialetto ionico-attico di età classica. Questi test hanno messo in evidenza quanto la qualità e, soprattutto, l'omogeneità del *dataset* incidano sulle prestazioni degli strumenti, mostrando

[33] Tuttavia, anche nel caso di testi stoichedici, più aumenta l'ampiezza della lacuna e il numero delle parole da essa ospitate, più la questione degli spazi bianchi si fa problematica: vd. *supra* per la soluzione adottata da E. Cullhed, il quale esclude gli spazi bianchi dal calcolo.

[34] Vd. *supra* e nt. 19. Acronimo di *Bidirectional Encoder Representations from Transformers*, *BERT* è un modello di linguistica computazionale sviluppato da Google, capace di analizzare testi in modo bidirezionale, cioè tenendo conto sia del contesto che precede sia di quello che segue ogni parola o carattere. Nella sua versione per il greco antico (*Ancient GreekBERT*), è stato addestrato sui testi letterari.

[35] Questa operazione, nota come *fine-tuning*, consiste nello 'specializzare' un modello già esistente facendolo lavorare su un nuovo insieme di testi, per adattarlo meglio a un contesto specifico.

come modelli allenati su *corpora* in cui prevale numericamente una determinata *facies* linguistica e dialettale – in particolare, quella del greco ionico-attico di età classica ed ellenistica – fatichino a adattarsi alla maggiore varietà grafica e linguistica dei testi del periodo arcaico.

Pur nella loro parzialità, questi primi esperimenti sono stati fondamentali, poiché hanno consentito di chiarire alcuni limiti delle soluzioni attualmente disponibili e di individuare criticità legate in particolare ai dati di addestramento e alle modalità di valutazione, soprattutto per i testi del periodo arcaico. Tali criticità si sono manifestate in tutta la loro evidenza quando nei nostri primi esperimenti abbiamo voluto includere anche, come già accennato, un LLM di tipo cosiddetto generalista: la scelta è caduta su *Mixtral*.

4. Lacune testuali e *Large Language Models*

Oltre agli esperimenti condotti con modelli sviluppati specificamente per il trattamento dei testi antichi, come *Ithaca* e *Ancient GreekBERT*,[36] una parte della sperimentazione è stata dedicata all'impiego di un modello linguistico di uso generale, quale è *Mixtral*.[37] L'obiettivo di questi test non era quello di sostituire gli strumenti specialistici, ma di osservare come modelli non progettati per il dominio epigrafico si comportino nel compito specifico del restauro delle lacune testuali.

I LLMs generalisti sono modelli di intelligenza artificiale addestrati su enormi quantità di testi di varia natura – spesso raccolti automaticamente dal *web* – che possono includere lingue moderne e antiche, testi letterari, conversazioni, articoli scientifici e molto altro. A differenza dei modelli specializzati, costruiti su *corpora* selezionati e controllabili, i LLMs generalisti (come *ChatGPT*, *Gemini* o *Mixtral*) puntano a massimizzare la versatilità, ma non offrono trasparenza sul tipo e sulla qualità dei dati usati durante l'addestramento: non è quindi possibile sapere se, e in che misura, abbiano effettivamente 'letto' testi epigrafici antichi o dati comparabili a quelli del nostro *corpus*. Tuttavia, la loro flessibilità, la capacità di 'ragionamento' contestuale e l'enorme ampiezza

[36] Esperimenti su altri modelli (basati su *BERT* o LLM) sono attualmente in corso e se ne darà conto nel lavoro di cui *supra* nt. 30.

[37] Jiang *et alii* 2024.

del *training set* li rendono strumenti potenzialmente interessanti anche per compiti inediti rispetto al loro uso originario.

In una prima fase, al modello è stato somministrato lo stesso *dataset* ripulito impiegato con *Ithaca*, ovvero testi privi di diacritici, spiriti, accenti, segni editoriali e simboli estranei al testo epigrafico. Nella seconda fase abbiamo sottoposto a *Mixtral* lo stesso corpus, ma in una versione solo parzialmente normalizzata. La rimozione ha riguardato soprattutto gli elementi editoriali, mentre sono stati mantenuti spiriti, accenti e il *digamma*; il testo è stato inoltre oggetto di un intervento mirato di traslitterazione, volto a rendere più esplicite le specificità dialettali: in particolare, le vocali dell'alfabeto 'epigrafico' sono state ricondotte ai corrispondenti valori dell'alfabeto 'letterario' (ad esempio, *epsilon* reso come *eta* nei casi in cui questo ne riflettesse il valore fonetico).[38] Tale scelta rispondeva all'esigenza di rendere più evidenti le differenze dialettali e si fondava sulla considerazione che questi segni potessero risultare familiari a un LLM verosimilmente pre-addestrato su *corpora* letterari greci, in versi o prosa che fossero, e quindi favorire il riconoscimento del contesto e del registro linguistico. Si è dunque ipotizzato che mantenere queste 'irregolarità' potesse migliorare la *performance* del modello, attivando in esso schemi linguistici già appresi. Una simile prospettiva è oggi sostenuta da una crescente letteratura nel campo del NLP, che invita a rivedere il pregiudizio secondo cui la presenza di 'rumore' comprometta necessariamente l'efficacia dei modelli: è stato infatti dimostrato che, in taluni casi, una pulitura eccessiva dei dati può ridurre la capacità del modello di confrontarsi con i testi reali, che raramente sono 'perfetti'.[39]

Per entrambi i *dataset*, sono state create versioni 'mascherate', cioè con lacune simulate in punti strategici della frase (inizio, fine, contesto ambiguo, in contesto chiaro e standard o formulare, in presenza di forme verbali rare ma anche di forme molto ricorrenti). A partire da queste versioni, abbiamo fornito

[38] Si è rimosso l'accento circonflesso e il *macron* sulle vocali brevi, sostituendo queste ultime con vocali lunghe chiuse o aperte annotate secondo l'uso ionico (ει, ου oppure η, ω), il solo 'tollerato' da Unicode (nel caso delle lunghe chiuse ciò altera il numero delle lettere della parola). Nei casi in cui quantità vocalica o accentazione restituiscano un'ambiguità dialettale ritenuta significativa, è stata mantenuta la lezione dell'editore, senza forzare la normalizzazione (è il caso, ad esempio, degli infiniti ἦμεν e ἤμην nella cosiddetta legge costituzionale di Dreros: https://inscriptions.packhum.org/text/201238).

[39] Un contributo particolarmente significativo in questa direzione è stato offerto da Al Sharou, Li, Specia 2021, che hanno mostrato come modelli NLP addestrati anche su dati 'sporchi' — cioè contenenti errori, varianti ortografiche o segni non standardizzati — possano sviluppare una maggiore robustezza e adattabilità. Il loro studio suggerisce che il 'rumore' non va sempre eliminato, ma piuttosto compreso e talvolta valorizzato in funzione degli obiettivi dell'esperimento.

al modello *prompt*[40] guidati, in cui veniva richiesto di completare la frase o proporre possibili ricostruzioni filologicamente plausibili.

La sperimentazione è stata condotta in più fasi e con approcci differenti, cercando di rispecchiare le diverse modalità con cui uno studioso si confronta con una lacuna testuale: dal confronto diretto con il testo lacunoso, al ricorso a iscrizioni parallele, fino all'uso di esempi-guida espliciti.

La prima modalità testata è stata quella *zero-shot*, in cui al modello viene semplicemente fornito il testo lacunoso e viene chiesto di proporre una possibile integrazione. Come prevedibile, i risultati si sono rivelati molto deboli: il modello spesso non rispetta la lunghezza della lacuna, produce sequenze troppo lunghe o troppo corte, e in diversi casi tende a ignorare del tutto la richiesta di ricostruzione e/o a ripetere sequenze di lettere o parole che si trovano nell'iscrizione lacunosa. Questo comportamento è legato, almeno in parte, a una delle limitazioni note dei LLMs: la difficoltà di contare o di rispettare vincoli strutturali rigidi.[41]

Abbiamo quindi introdotto due tecniche correttive, mutuate dalle recenti pratiche nel campo del *prompt engineering*: la generazione potenziata dal recupero dati (*Retrieval-Augmented Generation* [RAG]) e l'apprendimento contestuale (*In-Context Learning* [ICL]).[42]

Nel primo caso (RAG), al modello vengono fornite informazioni esterne attraverso un sistema di recupero documentale: insieme al *prompt*, esso riceve tre iscrizioni selezionate dal *corpus* epigrafico arcaico. Tali iscrizioni non sono scelte dall'utente, né direttamente dal modello linguistico, ma vengono individuate dalla struttura di recupero adottata (basata su *Ithaca*), che le identifica come le più simili al testo da integrare.[43] Queste informazioni di contesto vengono

[40] Con *prompt* si intende un'istruzione o una traccia data al modello per orientare la sua risposta; nei modelli di linguaggio, è il testo iniziale che attiva la generazione automatica di contenuti, come se si stesse ponendo una domanda o suggerendo l'inizio di una frase da completare.

[41] Per esempio, la lunghezza precisa di una parola o il numero di lettere attese in una lacuna.

[42] Lewis *et alii* 2020; Dong *et alii* 2024.

[43] La 'somiglianza' tra le iscrizioni è determinata automaticamente dal sistema di recupero sulla base di rappresentazioni numeriche dei testi, che tengono conto delle sequenze di caratteri e delle loro distribuzioni nel *corpus*, senza l'intervento di criteri definiti *a priori* dall'utente.

quindi utilizzate dal modello per formulare proposte di completamento più informate.[44]

Nel secondo caso (ICL) al modello viene di volta in volta mostrato un esempio composto da iscrizione con lacuna, le tre iscrizioni 'simili' di contesto e infine la proposta di integrazione: l'idea è che il modello apprenda per analogia dai casi già risolti.

Riportiamo di seguito un esempio concreto del lavoro condotto con i diversi metodi e sui due *dataset* (ripulito e non del tutto ripulito). L'iscrizione di riferimento (PHI ID 331814; *SEG* 38.948) è un'iscrizione graffita sul fondo di una *kylix*, databile al 500-480 a.C. e caratterizzata da una struttura sintattica piuttosto semplice.

Testo iscrizione:

τοῦτον τὸν σφύφον Πόρρος ἀποδίδōτι ἐς τὸν θίασον τῶν π[αρῶ]ν· αἰ δ' ἐφίλē Φρύναν, οὐκ ἄλλος κ' ἆγε· *h*ο δὲ γράπσας τὸν ἀννέμο(ν)τα πυγίξει

Nel test, è stata mascherata la parola ἀποδίδōτι per simulare una lacuna da colmare. Il completamento offerto dal modello, nei diversi scenari, è il seguente:

- *Dataset* 'pulito' (senza accenti, spiriti, *digamma* ecc.)
  - RAG: τον στρατηγον
  - RAG + ICL: εις
- *Dataset* 'rumoroso' (con accenti, spiriti, *digamma* ecc.)
  - *Zero-shot*: τὸν ἀννέμοντα πυγίξει τούτῳ τῷ σφύφῳ
  - RAG: nessuna risposta
  - RAG + ICL: ποιήσας

Il confronto tra i due gruppi di risposte evidenzia un dato interessante: il *dataset* 'rumoroso' non migliora l'accuratezza, ma genera risposte più coerenti sul piano formale e stilistico, con verbi o espressioni morfologicamente ben costruiti e semanticamente associabili al contesto epigrafico, anche se non perfettamente coerenti. È il caso di ποιήσας o di τὸν ἀννέμοντα πυγίξει τούτῳ τῷ σφύφῳ, con un testo che riprende esattamente la parte finale dell'iscrizione,

[44] Nel linguaggio dell'intelligenza artificiale, si parla di *embedding* per indicare la traduzione di un testo in una serie di numeri che ne rappresentano le caratteristiche linguistiche principali. In questo modo, testi simili tra loro (per lessico, struttura o contesto) risultano vicini anche dal punto di vista numerico.

continuandola con il suo *incipit*. Tuttavia, i risultati non sono sempre cogenti e le ragioni della variabilità dell'efficacia del risultato non sempre comprensibili.[45]

I risultati dei metodi con RAG e ICL, pur mostrando un miglioramento rispetto allo scenario *zero-shot*, sono ancora molto lontani da quelli ottenuti con i modelli specialistici: anche quando il RAG o l'ICL hanno fornito esempi calzanti o testi paralleli, il modello ha faticato a mantenere coerenza formale e soprattutto morfologica – il sistema dei casi e delle diatesi, ad esempio, non è 'comprensibile' all'algoritmo e dunque le integrazioni sono spesso casuali in termini grammaticali – e le sue proposte restano spesso linguisticamente o epigraficamente inaccettabili o assolutamente scorrette o allucinate.[46] L'impressione generale, alla luce degli esperimenti condotti, è che i LLMs generalisti non siano al momento strumenti adatti a risolvere in modo efficace il problema dell'integrazione di lacune nei testi epigrafici greci di età arcaica (cf. la tabella di fig. 2).

L'unico elemento di rilievo emerso da questi test riguarda l'effetto prodotto dall'impiego del *dataset* meno 'normalizzato', quello in cui si mantenevano alcuni tratti grafici caratteristici della scrittura greca (come *digamma*, accenti e spiriti). In questo scenario, si è osservata una modesta ma significativa aderenza alle strutture sintattiche e lessicali proprie del registro epigrafico. Un risultato che conferma quanto recentemente evidenziato da studi sull'importanza del 'rumore' nei dati di *input* e mostra come, in alcuni casi, un certo grado di 'sporcizia' (*useful noise*)[47] può contribuire a rendere i modelli più sensibili al contesto reale d'uso e meno propensi a produrre risposte generiche o astratte.

Nel complesso, le prestazioni del LLM testato restano ben al di sotto della soglia di affidabilità richiesta per un impiego scientificamente fondato. Va tuttavia sottolineato ancora una volta che gli esperimenti qui presentati si inseriscono in uno scenario in rapido sviluppo e costituiscono una fase ancora in corso del progetto: la sperimentazione è stata infatti estesa a modelli non

[45] Esiste un settore specifico dell'informatica noto come *Explainable AI* (XAI), dedicato allo studio dei meccanismi interni dei modelli di intelligenza artificiale, con l'obiettivo di comprenderne il funzionamento e rendere più trasparenti i criteri con cui producono risultati. Si tratta di un ambito ancora in fase di sviluppo, che cerca di superare la cosiddetta *black box* — ovvero l'opacità decisionale — tipica dei modelli più complessi, come quelli basati su reti neurali profonde. Vd. ad esempio Longo *et alii* 2024.

[46] Nel linguaggio tecnico dell'intelligenza artificiale, si definiscono allucinate (dall'inglese *hallucinated*) le risposte fornite da un modello che appaiono plausibili o ben formate dal punto di vista linguistico, ma che in realtà sono prive di fondamento, inventate o errate. È un fenomeno frequente nei modelli di grandi dimensioni (LLMs), soprattutto quando vengono interpellati su argomenti specialistici o poco rappresentati nei dati su cui sono stati addestrati.

[47] Cf. Al Sharou, Li, Specia 2021, p. 53, nt. 1.

generalisti e *domain-specific*,[48] inclusi modelli basati su architetture di tipo *BERT* e modelli della famiglia *Llama* addestrati su *corpora* epigrafici e papirologici, includendo anche l'attiva collaborazione da parte di esperti umani.

Caratteristica	*Ithaca*	*Ancient GreekBERT*	*Mixtral*
Tipo di modello	predizione su rete neurale (carattere per carattere)	modello a trasformatori (*token* per *token*)	modello pre-addestrato su testi di vario genere
Modalità di *input*	testo epigrafico con lacune (mascherato)	frasi o sequenze con parole mascherate	*prompt* in linguaggio naturale
Obiettivo	restauro, datazione, localizzazione	comprensione del contesto, completamento	completamento e ragionamento contestuale
Dataset per addestramento	*corpus* misto, prevalentemente classico/post-classico	testi letterari (TLG, *corpus* standard)	*corpus* moderno e antico (non specifico)
Sensibilità ai dialetti	moderata (più su area geografica)	moderata	variabile, migliorabile con *prompt* mirati
Affidabilità su testi arcaici	discreta per localizzazione; limitata per integrazione	limitata per integrazione	in generale bassa per integrazione; risultati variabili
Possibilità di adattamento	limitata (modello chiuso ma con possibilità di *fine-tuning*)	limitata (modello chiuso ma con possibilità di *fine-tuning*)	alta (via *prompt engineering* o *fine-tuning*)

2 - Confronto sintetico tra i principali modelli testati nel progetto *Lacunae*, che ne evidenzia natura, obiettivi, affidabilità e potenzialità. La tabella intende chiarire le differenze tra modelli specializzati (come *Ithaca* o *Ancient GreekBERT*) e modello generalista (*Mixtral*) in rapporto alla ricerca condotta.

5. Dal testo al documento, dalla lingua all'alfabeto. Una proposta metodologica

Come ben evidenziato nei due paragrafi precedenti, affrontare sul piano esclusivamente testuale e linguistico il problema informatico dell'integrazione delle lacune nelle epigrafi greche di età arcaico-classica non sembra aver sin qui restituito risultati particolarmente incoraggianti: l'eccentricità cronologica del nostro *dataset* rispetto a quello utilizzato per l'addestramento di *Ithaca* sembra influenzare negativamente la *performance* dello strumento, l'utilizzo del quale ha inoltre richiesto un'ulteriore riduzione, dovuta all'eliminazione dei testi troppo

[48] Vd. *supra* ntt. 30 e 36.

lunghi e di quelli troppo corti,[49] del numero delle iscrizioni che componevano un *dataset* già non molto consistente, obbligando infine a classificare i dialetti soltanto secondo i più ampi macrogruppi. Comparabile a quella di *Ithaca* – una volta che il modello è stato messo a punto per lavorare con il nostro *dataset* – è a sua volta la *performance* di *Ancient GreekBERT*, un risultato che, *en passant*, asseconda forse l'idea che i testi delle iscrizioni del periodo arcaico-classico non andrebbero 'trattati' separatamente da quelli della letteratura del medesimo periodo, che sono spesso di ispirazione proprio nell'integrazione delle loro lacune; quanto alle risposte, spesso allucinatorie, del LLM da noi utilizzato, sottolineano senz'altro, senza destare eccessiva sorpresa, la necessità di un adeguato addestramento, a fronte di compiti così specifici sotto diversi rispetti, di modelli certamente nati per altro genere di scopi.

L'esperienza maturata sin qui nel campo del NLP, nell'ambito del nostro progetto, è senz'altro servita a calibrare e affinare i parametri dei successivi esperimenti, attualmente in corso;[50] al tempo stesso, è proprio grazie a questa esperienza che si è fatta strada l'idea, perfettamente in linea con le aspirazioni iniziali del progetto che, come già detto, in nessun modo intende trascurare gli aspetti materiali di un documento epigrafico, che fosse anche opportuno adottare un approccio nuovo, appunto più attento a quegli stessi aspetti. Come è ben noto, la materialità di un testo epigrafico, specialmente se di età arcaico-classica, si esprime innanzitutto nel particolare alfabeto epicorico in cui esso è redatto e nel fatto di essere stato redatto in *scriptio continua* perlopiù non 'impaginata', prendendo così vita quella combinazione che, come è altrettanto noto e risulterà tra poco ancora più evidente, semplicemente e letteralmente determina il calcolo del numero di lettere che la lacuna può contenere, a seconda della loro combinazione. Come già evidenziato, a qualunque strumento informatico ci si rivolga, fra quelli attualmente disponibili per il riempimento delle lacune testuali, è in ogni caso all'utente che spetta l'onere di compiere quel calcolo, con una approssimazione che inevitabilmente diminuisce, anche di molto, con l'aumentare dello spazio occupato dalla lacuna.[51]

[49] Tra le iscrizioni cretesi, sono ad esempio troppo lunghe sia il codice di Gortina (vd. *infra* e nt. 61) sia il contratto stretto tra i *Dataleis* e lo scriba Spensithios (su cui, ad esempio, Tribulato 2017).

[50] Vd. *supra* e ntt. 30 e 36.

[51] Significativo il commento di Cullhed 2025, p. 6: «Moreover, these models offer a notable advantage: they are particularly useful when scholars have an approximate sense of the number of missing characters but lack precise information about word boundaries — the most common situation in the study of

È sembrato di conseguenza necessario o persino naturale compiere, almeno apparentemente, un passo indietro, che ci ha riportato proprio davanti a quella porzione danneggiata di testo da riempire con un numero di lettere (e, se del caso, di segni di interpunzione) la cui variabilità è una questione di proporzioni tra lo spazio disponibile e i singoli elementi di differente grandezza che lo riempiono: determinare questa variabilità può appunto essere questione particolarmente delicata e difficile, quando la disposizione dei caratteri non sia di tipo stoichedico (la norma nei testi che compongono il nostro *dataset*, la cui scrittura può piuttosto definirsi a spaziatura proporzionale) e la lacuna sia di una certa ampiezza. È appunto di compiere innanzitutto questo calcolo – che deve necessariamente precedere qualunque proposta scientificamente fondata di integrazione di una lacuna – ciò che abbiamo deciso di demandare all'intelligenza artificiale, potendo ottenere, per dir così, un duplice risultato: data una lacuna, il numero minimo e massimo, con tutte le declinazioni intermedie, di caratteri (ivi compresi gli eventuali segni di interpunzione) da essa ospitabili, calcolato in ragione della loro grandezza commisurata a quella della lacuna; essendo a sua volta questo stesso numero (variabile) una conseguenza della capacità della macchina di mostrare tutte le possibili combinazioni di caratteri in grado di riempire quella stessa lacuna, utilizzando le lettere dell'alfabeto epicorico dell'iscrizione presa in esame.

Trattandosi infatti di un aspetto materiale, per il quale non è affatto necessario, almeno in questa prima fase, che la macchina conosca il significato dei segni grafici utilizzati dalle iscrizioni (è bene ripetere: caratteri alfabetici ed eventuali segni di interpunzione), ci si è rivolti alle tecniche della visione artificiale (*Computer Vision* [CV]) – appunto utilizzate quando, nell'ambito della collaborazione fra antichistica e informatica, l'attenzione si focalizzi sui supporti e le scritture che li caratterizzano, piuttosto che sui testi e le lingue in cui sono redatti –, che hanno a loro volta richiesto il contributo di altre competenze informatiche.[52]

ancient texts». Tale situazione sembra tuttavia molto meno comune proprio nel caso dei testi epigrafici greci di età arcaico-classica: vd. *infra*.

[52] Messe appunto a disposizione da chi sta curando, lato informatico, la parte 'visuale' del progetto: vd. *supra* nt. 1.

Potenziamento della qualità delle immagini, che ne aumenta la leggibilità, binarizzazione e analisi del *layout*,[53] che a loro volta rendono possibile il riconoscimento ottico dei caratteri (*Optical Character Recognition* [OCR]), anche quando realizzati a mano (*Handwritten Text Recognition* [HTR]),[54] sono le tecniche di cui ci si può avvalere nella trascrizione e analisi delle scritture antiche, ma anche nell'identificazione dei redattori di manoscritti o papiri, per tacere dei casi in cui esse sono piuttosto sfruttate per la ricostituzione di supporti frammentari.[55]

Dovendo sottoporre alla macchina immagini e non stringhe di testo, è infine parsa quasi obbligata l'opzione per uno strumento tradizionale del lavoro dell'epigrafista che, già a rischio di inaffidabilità,[56] ha ormai riacquistato piena dignità scientifica proprio grazie alla possibilità di realizzarlo (a partire da una buona fotografia) o aumentarne decisamente la precisione anche con l'ausilio dei più diffusi e utilizzati programmi di grafica;[57] l'apografo o facsimile, un'immagine decisamente più leggera (in termini di spazio di archiviazione) di qualunque buona fotografia e spesso preferita a quest'ultima, in molte edizioni epigrafiche anche recenti, per la sua maggiore leggibilità[58] (oltre che per il minor costo editoriale), ha nel nostro caso il vantaggio aggiuntivo di essere proprio quel tipo di immagine bidimensionale richiesta dall'occhio digitale.[59]

Per i primi esperimenti compiuti nell'ambito del progetto, di cui tratta il prossimo paragrafo, sono state dunque scelte due iscrizioni per le quali disponiamo sia di ottime fotografie, sia di attendibili apografi, e che danno

[53] A titolo di esempio, fra altri lavori dedicati al tema: Zottin *et alii* 2024.

[54] Sánchez-DelaCruz, Loeza-Mejía 2024; Krithiga *et alii* 2025.

[55] Possiamo senz'altro limitarci a citare i due importanti progetti *D-Scribes* (9.2018-5.2023) e, in continuità e ulteriore sviluppo, *EGRAPSA* (6.2023-5.2028), diretti da Isabelle Marthot Santaniello e dedicati ai documenti papiracei dell'Egitto: https://d-scribes.philhist.unibas.ch/en.

[56] È ben noto il caso dell'identificazione di uno degli alleati di Atene menzionati nella colonna di sinistra del decreto di Aristotele (Rhodes-Osborne, *GHI*, pp. 92-105, nr. 22, ll. 97-98): [Θη]ραίων | [ὁ δ]ῆμος secondo la lettura di Coleman, Bradeen 1967, che hanno corretto i precedenti e impossibili [Ἐρυθ]ραίων e [Κερκυ]ραίων dopo aver confrontato l'apografo, pubblicato nell'*editio princeps* (Eustratiades 1851, nr. 61) e accolto in *IG* II.1.17, con il calco da loro realizzato.

[57] Shaus *et alii* 2016; Davis Parker, Rollston 2019; Martín Hernández, Shaus 2022.

[58] Ad esempio Dana, *Corresp. gr. privée*, in cui le fotografie dei documenti sono regolarmente affiancate dagli apografi realizzati dall'autrice.

[59] L'apografo, inoltre, è nella maggioranza dei casi un'immagine in scala di grigi, perciò facilmente binarizzabile: vd. *infra*.

testimonianza di tutte le lettere dell'alfabeto epicorico in cui sono state redatte.[60] Questi stessi esperimenti hanno anche fornito l'occasione per cominciare a immaginare lo strumento informatico (e intanto, con esso, il logo) che di questo progetto costituisce l'obiettivo finale. La strada è di sicuro ancora lunga, ma l'inizio del viaggio sembra promettente.

6. Dal dialetto all'alfabeto epicorico: gli esperimenti

Per testare sul campo questo nuovo approccio, abbiamo selezionato due iscrizioni ben note e in buone condizioni di conservazione: il Codice di Gortina (in particolare, la quinta colonna) e il decreto di fondazione della colonia di Naupatto. I due testi, incisi rispettivamente su pietra e su bronzo, sono rappresentativi di due varietà dell'alfabeto greco arcaico – rispettivamente del gruppo verde e del gruppo rosso, secondo la classificazione di Kirchhoff – e offrono dunque un terreno ideale per testare la capacità dell'algoritmo di riconoscere e gestire le specificità formali delle diverse tradizioni epigrafiche e alfabetiche.[61]

Le immagini degli apografi, prima binarizzate (fig. 3),[62] sono state analizzate tramite tecniche di CV, senza prendere in considerazione in alcun modo la componente semantica o linguistica dei documenti.

La macchina, dunque, non ha 'letto' il testo, ma ha elaborato visivamente gli spazi, le forme delle lettere e la loro distribuzione. Questo ha permesso di stimare quante lettere possono essere ospitate da una determinata lacuna, in base

[60] È chiaro che la maggior parte delle iscrizioni, soprattutto di età arcaica, non contiene tutte le lettere dell'alfabeto in cui è redatta e che questo, per lo strumento che abbiamo in mente, costituisce un problema. La soluzione, alla quale stiamo pensando, sembrerebbe al momento poter essere offerta vuoi dal raggruppamento delle lettere complessivamente note dalle iscrizioni di ciascun singolo alfabeto epicorico in (teoriche) classi di grandezza, in modo che lo stesso spazio occupato da una determinata lettera possa considerarsi occupabile dalle altre di analoghe dimensioni; vuoi dalla creazione di dati (alfabetici) sintetici, che suppliscano le lettere non presenti in una determinata iscrizione, generate ancora una volta sulla base di quelle complessivamente note per l'alfabeto epicorico in oggetto e commisurate caso per caso, quanto a dimensioni, a quelle della singola iscrizione in esame.

[61] Il facsimile del codice è quello realizzato per Margherita Guarducci da Enrico Stefani, per l'edizione *I.Cret.* IV, 72, e inserito anche nell'edizione di Willetts 1967, a sua volta corredata dalle fotografie di Peter Gautel, specializzato in fotografia archeologica. Il facsimile del decreto di fondazione di Naupatto (*IG* IX 1^2, 3, 718) è tratto da Roehl *IGA*, p. 92, nr. XXIX.1; foto a colori e in bianco e nero sono disponibili nella collezione *online* del British Museum, in cui il bronzo è conservato (https://www.britishmuseum.org/collection/object/G_1896-1218-1), e in *British Mus.* IV, pp. 119-122, nr. 954.

[62] La binarizzazione è un processo che trasforma un'immagine a toni di grigio o a colori in un'immagine composta solo da due colori (tipicamente nero e bianco). Questa semplificazione rende l'immagine più adatta all'elaborazione da parte dei sistemi informatici, facilitando il riconoscimento delle lettere incise rispetto allo sfondo e a eventuali 'rumori'.

3 - Immagine binarizzata della V colonna del Codice di Gortina (*I.Cret.* IV, 72), tratta dall'apografo realizzato da E. Stefani per M. Guarducci (in Willetts 1967).

alla spaziatura tra i segni, alla dimensione delle singole occorrenze delle lettere nel documento oggetto di indagine; si è inoltre tenuto conto di eventuali segni di interpunzione o eccezionali distanziamenti tra le lettere o intorno a quegli stessi segni. Il sistema ha restituito una serie di possibili stime per ogni lacuna, indicando il minimo e massimo numero di lettere plausibilmente contenibili.

L'esito più rilevante dell'esperimento è stato la capacità del modello di adattarsi alla varietà formale delle due iscrizioni, riconoscendo con precisione i margini dei segni e suggerendo soluzioni coerenti con la forma, la dimensione e il *ductus* delle lettere nel documento di partenza.

È importante sottolineare che il risultato così ottenuto si colloca in un momento preliminare rispetto alla proposta di integrazione avanzata dall'epigrafista: non fornisce risposte sul piano del significato o della coerenza grammaticale, ma si propone come strumento di supporto per lo studioso, nella delicata fase di valutazione materiale della lacuna. In un contesto come quello delle iscrizioni arcaiche – redatte in *scriptio continua*, con alfabeti locali e forme grafiche non standardizzate –, poter disporre di una stima non arbitraria sulla capacità effettiva della lacuna rappresenta un primo passo verso proposte di integrazione che tengano sempre conto anche della materialità del supporto, ossia dell'ampiezza delle singole lettere, dello spazio effettivo della lacuna, delle variabili paleografiche incontrate nell'iscrizione che si sottopone all'indagine.

Gli esiti, pur ancora in fase esplorativa, hanno mostrato che un'analisi visuale automatizzata è in grado di affiancare con profitto le valutazioni tradizionali dell'epigrafista. L'approccio adottato consente infatti di rendere espliciti e misurabili alcuni vincoli materiali – spesso valutati in modo intuitivo – relativi all'ampiezza dello spazio disponibile, alla dimensione delle lettere e alle variabili paleografiche proprie del singolo documento.

Su queste basi è in corso di sviluppo un prototipo di strumento digitale, concepito per supportare lo studioso nelle fasi preliminari dell'integrazione epigrafica. Attraverso un'interfaccia visuale (fig. 4), l'utente può caricare l'immagine dell'iscrizione – fotografia o facsimile – e selezionare una o più aree danneggiate. Il sistema analizza quindi l'immagine di partenza, calcola la dimensione dei caratteri dell'intera iscrizione e stima l'ampiezza della lacuna nella parte selezionata dall'utente. A partire da tali vincoli materiali, lo strumento genera automaticamente tutte le combinazioni di lettere compatibili con lo spazio disponibile, evidenziando in modo trasparente i limiti materiali entro cui l'epigrafista

può formulare le proprie ipotesi di integrazione. In questo senso, il prototipo non fornisce soluzioni interpretative, né interviene sul piano semantico o grammaticale, ma si propone come uno strumento di supporto che formalizza e rende riproducibili alcune operazioni preliminari del lavoro epigrafico, mantenendo centrale il ruolo del giudizio critico dell'epigrafista.[63]

4 - Interfaccia dimostrativa del prototipo *online* del progetto *Lacunae* per l'analisi e l'integrazione delle lacune epigrafiche mediante metodi di *Computer Vision* (il logo del progetto è visibile in alto a sinistra).

[63] I dettagli relativi ai metodi informatici impiegati e ai risultati ottenuti saranno accessibili in un articolo al quale stiamo lavorando insieme a chi si occupa della parte informatica del progetto relativa alla CV (vd. *supra* nt. 1).

7. Riflessioni conclusive. Risultati attuali e prospettive future del progetto

Come già detto, gli esiti di questi primi esperimenti con le tecniche di CV, per quanto anch'essi provvisori, sembrano promettenti: il computer è in grado di commisurare l'ampiezza di una lacuna di un testo epigrafico alla variabile grandezza dei segni grafici che lo compongono (lettere ed eventuali segni di interpunzione) e di riempirla con quegli stessi segni, il cui numero può di conseguenza cambiare a seconda delle loro combinazioni.[64] I risultati, sinora restituiti nella sola forma di file grafico (.PNG), sono appunto presentati nell'alfabeto epicorico dell'iscrizione nonché in *scriptio continua* (peraltro, in direzione tanto destrorsa quanto sinistrorsa) e, grazie alla loro compiutezza, rappresentano tutte le possibilità di integrazione di quella stessa lacuna: sta all'epigrafista scegliere la più plausibile.

Quest'ultima caratteristica senz'altro positiva dell'algoritmo è causa tuttavia, al tempo stesso, del suo principale difetto: proprio l'esaustività dei risultati può mettere l'epigrafista di fronte a un numero decisamente consistente di opzioni, tanto maggiore quanto più grande è la lacuna, mentre la loro elaborazione richiede un tempo computazionale davvero elevato. La correzione di questo inconveniente è il nostro prossimo obiettivo, al conseguimento del quale stiamo già lavorando: oltre all'idea di permettere all'utente di stabilire preliminarmente in quale direzione di scrittura vuole di volta in volta il suggerimento e di escludere *a priori* tutti i risultati che siano oggettivamente impossibili sul piano linguistico (ad esempio la ricorrenza in successione del medesimo segno per un numero eccessivo di volte),[65] sono attualmente allo studio, in costante contatto con chi cura la parte informatica del progetto che coinvolge la CV,[66] ulteriori possibilità di filtraggio o, per meglio dire, di gerarchizzazione dei risultati, perché vengano presentati a partire da quelli statisticamente più probabili, tenuto innanzitutto conto del dialetto dell'iscrizione, anche nella sua declinazione epicorica, nonché dell'alfabeto che essa utilizza – senza tuttavia venir meno, nell'un caso come nell'altro, al rispetto della *scriptio continua*. Al medesimo, rigoroso, rispetto sarà infine improntata anche l'auspicata collaborazione, nello stesso *tool*,

[64] Non sembra del tutto inutile precisare esplicitamente che, a fronte di un apografo attendibile, contano appunto le proporzioni (tra la grandezza della lacuna e quella delle singole lettere) e non le misure assolute, che la macchina in ogni caso ignora.

[65] Nei casi in cui, ovviamente, si tratti di iscrizioni che hanno linguisticamente senso.

[66] Vd. *supra* nt. 1.

fra le tecniche della CV e quelle del NLP, che costituisce uno degli obiettivi finali del progetto *Lacunae*.

8. Riflessioni conclusive. Il dialogo non sempre facile tra epigrafia e informatica

A margine degli esperimenti finora condotti, sembra opportuno soffermarsi su alcune riflessioni nate dal confronto diretto – spesso produttivo, talvolta problematico – tra le competenze umanistiche e quelle informatiche. Più che offrire risposte conclusive, queste osservazioni mirano a ripercorrere criticamente il processo di collaborazione tra saperi diversi, evidenziandone potenzialità e limiti, in un ambito – quello dell'epigrafia greca – che pone sfide peculiari tanto alla filologia quanto alla modellizzazione computazionale.

Negli ultimi anni, si è spesso fatto riferimento al dialogo tra scienze umane e informatica. Tuttavia, l'esperienza concreta mostra che si tratta, molto spesso, di un confronto parziale: le due discipline si accostano, infatti, con aspettative, metodi e priorità profondamente diversi. Il punto di attrito principale non riguarda tanto il piano della comunicazione tra studiosi di diversa formazione e *background*, quanto la diversa concezione della verità e del dato. Da un lato, la propensione alla cautela, alla sfumatura interpretativa, alla gestione dell'ambiguità; dall'altro, l'esigenza di efficienza, di formalizzazione e riduzione dell'incertezza. Il caso delle iscrizioni frammentarie lo mostra chiaramente: ciò che per il filologo è uno spazio ermeneutico aperto, per l'algoritmo è un vuoto da colmare con esattezza.

Eppure, è proprio in questa distanza che può maturare una prospettiva diversa. L'intelligenza artificiale, anche quando fallisce nel suo obiettivo più immediato, può agire come strumento euristico: forzare lo studioso a esplicitare i propri criteri, a interrogare abitudini interpretative spesso implicite, a ripensare i propri strumenti in funzione di un nuovo contesto operativo. In questo senso, l'errore della macchina può diventare occasione di consapevolezza, e il fallimento un punto di partenza per una riflessione più profonda sul metodo.

La collaborazione tra umanisti e informatici – al cuore di questo progetto – non è mai stata un processo lineare. All'entusiasmo iniziale si sono alternati momenti di disorientamento, incomprensioni reciproche, frustrazioni legate alle inevitabili differenze di linguaggio e di priorità. I tempi della ricerca, le aspettative sui risultati, la gestione dell'errore o dell'imprevisto: tutto, in una collaborazione del genere, richiede negoziazione continua.

Eppure, proprio in questa fatica si è creata una forma di fiducia operativa. Non una fusione dei saperi, ma una loro collaborazione consapevole. La filologia non ha ceduto alla tentazione dell'automazione, e l'informatica ha imparato a convivere con l'incertezza e la frammentarietà dei dati antichi. Questo non ha prodotto – almeno per ora – uno strumento risolutivo, ma ha generato qualcosa di forse più interessante: un terreno comune in cui sperimentare e discutere metodi e approcci.

Bile, *Dialecte crétois*	M. Bile, *Le dialecte crétois ancien. Étude de la langue des inscriptions. Recueil des inscriptions postérieures aux IC* (École Francaise d'Athènes. Études crétoises, 27), Paris 1988.
Dana, *Corresp.gr.privée*	M. Dana, *La correspondance grecque privée sur plomb et sur tesson. Corpus épigraphique et commentaire historique* (Vestigia. Beiträge zur alten Geschichte, 73), München 2021.
I.British Mus. IV	G. Hirschfeld, S.H. Marshall, *The Collection of Ancient Greek Inscriptions in the British Museum*, IV. *Knidos, Halikarnassos and Branchidae. Supplementary and Miscellaneous Inscriptions*, London 1893-1916.
I.Cret.	M. Guarducci, *Inscriptiones Creticae*, Romae 1935-1950.
IG	*Inscriptiones Graecae*, Berolini 1877–.
Rhodes-Osborne, *GHI*	P.J. Rhodes, R. Osborne, *Greek Historical Inscriptions, 404-323 BC*, Oxford 2003.
Roehl, *IGA*	H. Roehl, *Imagines Inscriptionum Graecarum Antiquissimarum in Usum Scholarum. Editio tertia*, Berolini 1907.
SEG	*Supplementum Epigraphicum Graecum*, Leiden 1923–.
Wachter, *Non-Attic Vase Inscr.*	R. Wachter, *Non-Attic Greek Vase Inscriptions*, Oxford 2001.

Al Sharou K., Li Z., Specia L. 2021. *Towards a Better Understanding of Noise in Natural Language Processing*, in R. Mitkov, G. Angelova (eds.), *Deep Learning for Natural Language Processing Methods and Applications. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), 1–3 September 2021 (Online)*, Shoumen (Bulgaria), pp. 53-62 (https://aclanthology.org/2021.ranlp-1.7).

Angliker E., Bultrighini I. (eds.) 2023. *New Approaches to the Materiality of Text in the Ancient Mediterranean. From Monuments and Buildings to Small Portable Objects* (Archaeology of the Mediterranean World, 4), Turnhout.

Assael Y., Sommerschield Th., Cooley A., Shillingford B., Pavlopoulos J., Suresh P., Herms B., Grayston J., Maynard B., Dietrich N., Wulgaert R., Prag J., Mullen A., Mohamed Sh. 2025. *Contextualizing Ancient Texts with Generative Neural Networks*, in *Nature* 645, pp. 141–147 (https://doi.org/10.1038/s41586-025-09292-5).

Assael Y., Sommerschield Th., Prag J. 2019. *Restoring Ancient Text Using Deep Learning: A Case Study on Greek Epigraphy*, in K. Inui, J. Jiang, V. Ng, X. Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong (China) 3-7 November 2019*, Stroudsburg (PA), pp. 6368-6375 (https://doi.org/10.18653/v1/D19-1668).

Assael Y., Sommerschield Th., 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*, in *Nature* 603, pp. 280–283 (https://doi.org/10.1038/s41586-022-04448-z).

Bamman D., Burns P.J. 2020. *Latin BERT: A Contextual Language Model for Classical Philology*, in *arXiv:2009.10053v1* (prePrint) (https://doi.org/10.48550/arXiv.2009.10053).

Buck C.D. 1955. *The Greek Dialects. Grammar, Selected Inscriptions, Glossary*, Chicago.

Butskhrikidze M. 2021, *What do Modern Languages with* Scriptio Continua *Have in Common?*, in *Journal of Linguistics / Jazykovedný časopis* 72, pp. 821-838 (https://doi.org/10.2478/jazcas-2022-0006).

Coleman J.E., Bradeen D.V. 1967. *Thera on I.G., II^2, 43*, in *Hesperia* 36, pp. 102-104.

Cullhed E. 2025. *Instruction-Tuning Pretrained Causal Language Models to Restore Ancient Greek Papyri and Inscriptions*, in *DSH*, pp. 1-8 (https://doi.org/10.1093/llc/fqaf131).

Davis Parker H.D., Rollston Ch.A. 2019. *Teaching Epigraphy in the Digital Age*, in D. Hamidović, C. Clivaz, S. Bowen Savant, A. Marguerat (eds.), *Ancient Manuscripts in Digital Culture. Visualisation, Data Mining, Communication* (Digital Biblical Studies), Leiden-Boston, pp. 189-216 (https://brill.com/downloadpdf/edcollbook-oa/title/34930.pdf).

Devlin J., Chang M.-W., Lee K., Toutanova K. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, in J. Burstein, Ch. Doran, Th. Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis (Minnesota) 2-7 June 2019*, I. *Long and Short Papers*, Stroudsburg (PA), pp. 4171-4186 (https://doi.org/10.18653/V1/N19-1423).

Domínguez Casado R. 2014. *El dialecto de Tera. Gramática y estudio dialectal* (Tesis Doctoral), Madrid.

Dong Q., Li L., Dai D., Zheng C., Wu J., Chang B., Sun X., Xu L., Sui Z., 2024. *A Survey on In-context Learning*, in Y. Al-Onaizan, M. Bansal, Y.-N. Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami (Florida) 12-16 November 2024*, Kerrville (Texas), pp. 1107-1128 (https://aclanthology.org/2024.emnlp-main.64.pdf).

Eustratiades P. 1851. *Ἐπιγραφαὶ ἀνέκδοτοι ἀνακαλυφθείσαι καὶ ἐκδοθείσαι ὑπό τοῦ ἀρχαιολογικοῦ συλλόγου*, II, Ἀθήνησιν.

Fujii T., Shibata K., Yamaguchi A., Morishita T., Sogawa Y. 2023. *How do Different Tokenizers Perform on Downstream Tasks in Scriptio Continua Languages?: A Case Study in Japanese*, in V. Padmakumar, G. Vallejo, Y. Fu (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto 10-12 July 2023*, IV. *Student Research Workshop*, Stroudsburg (PA), pp. 39-49 (https://aclanthology.org/2023.acl-srw.5.pdf).

Giannakis G.K. 2014. *Encyclopedia of Ancient Greek Language and Linguistics*, Leiden-Boston.

Hoogendijk F.A.J., Gompel S.M.V. (eds.) 2018. *The Materiality of Texts from Ancient Egypt: New Approaches to the Study of Textual Material from the Early Pharaonic to the Late Antique Period* (Papyrologica Lugduno-Batava, 35), Leiden-Boston.

Inglese A. 2008. *Thera arcaica. Le iscrizioni rupestri dell'*agora *degli dei* (Themata, 1), Tivoli.

Jiang A.Q., Sablayrolles A., Roux A., Mensch A., Savary B., Bamford Ch., Sing Chaplot D., de las Casas D., Bou Hanna E., Bressand F., Lengyel G., Bour G., Lample G., Renard Lavaud L., Saulnier L., Lachaux M.-A., Stock P., Subramanian S., Yang S., Antoniak S., Le Scao T., Gervet Th., Lavril Th., Wang Th., Lacroix T., El Sayed W. 2024. *Mixtral of Experts. arXiv preprint* (https://arxiv.org/pdf/2401.04088).

Kang K., Jin K., Yang S., Jang S., Choo J., Kim Y. 2021. *Restoring and Mining the Records of the Joseon Dynasty via Neural Language Modeling and Machine Translation*, in K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 6-11 June 2021 (Online)*, Stroudsburg (PA), pp. 4031–4042 (https://doi.org/10.18653/v1/2021.naacl-main.317).

Kirchhoff A. 1887. *Studien zur Geschichte des griechischen Alphabets. Vierte umgearbeitete Auflage*, Gütersloh.

Krithiga R., Varsini S., Gabriel Joshua R., Om Kumar C.U. 2025. *Ancient Character Recognition: A Comprehensive Review*, in *IEEE Access* 13, pp. 88847-88857 (https://doi.org/10.1109/ACCESS.2023.3341352).

Lazar K., Saret B., Yehudai A., Horowitz W., Wasserman N., Stanovsky G. 2021. *Filling the Gaps in Ancient Akkadian Texts: A Masked Language Modelling Approach*, in K. Lazar, B. Saret, A. Yehudai, W. Horowitz, N. Wasserman, G. Stanovsky (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana (Dominican Republic) 7-11 November 2021*, Stroudsburg (PA), pp. 4682-4691 (https://doi.org/10.18653/v1/2021.emnlp-main.384).

Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., Kielaet D. 2020. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, in H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (eds.), *NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver BC Canada, 6-12 December 2020*, XII, New York, pp. 9459-9474 (https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf).

Longo L., Brcic M., Cabitza F., Choi J., Confalonieri R., Del Ser J., Guidotti R., Hayashi Y., Herrera F., Holzinger A., Jiang R., Khosravi H., Lecue F., Malgieri G., Páez A., Samek W., Schneider J., Speith T., Stumpf S. 2024. *Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions*, in *Information Fusion* 106, art. 102301 (https://doi.org/10.1016/j.inffus.2024.102301).

Martín Hernández R., Shaus A. 2022. *New Technologies for Tracing Magical Texts and Drawings: Experience with Automatic Binarization Algorithms*, in S. Torallas Tovar, R. Martín Hernández (eds.), *The Materiality of Greek and Roman Curse Tablets. Technological Advances*, Chicago, pp. 33-43 (https://isac.uchicago.edu/research/publications/misc/materiality-greek-and-roman-curse-tablets-technological-advances).

Nieto Izquierdo E. 2025. *The making of Paradeigmata VI.4: Unveiling the Doric Islands of Ancient Greece*, in *CHS Research Bulletin* 13 (https://nrs.harvard.edu/URN-3:HLNC.ESSAY:106296652).

Papavassileiou K., Kosmopoulos D.I., Owens G. 2023. *A Generative Model for the Mycenaean Linear B Script and Its Application in Infilling Text from Ancient Tablets*, in *Journal on Computing and Cultural Heritage* 16, art. 52 (https://doi.org/10.1145/3593431).

Petrovic A., Petrovic I., Thomas E.V. (eds.) 2019. *The Materiality of Text – Placement, Perception, and Presence of Inscribed Texts in Classical Antiquity* (Brill Studies in Greek and Roman Epigraphy, 11), Leiden-Boston.

Riemenschneider F., Frank A. 2023a. Graecia capta ferum victorem cepit. *Detecting Latin Allusions to Ancient Greek Literature*, in A. Anderson, S. Gordin, S. Klein, B. Li, Y. Liu, M.C. Passarotti (eds.), *Proceedings of the Ancient Language Processing Workshop associated with the 14th International Conference on Recent Advances in Natural Language Processing (RANLP 2023), Varna (Bulgaria), 8 September 2023*, Shoumen (Bulgaria), pp. 30-38 (https://aclanthology.org/2023.alp-1.4.pdf).

Riemenschneider F., Frank A. 2023b. *Exploring Large Language Models for Classical Philology*, in A. Rogers, J. Boyd-Graber, N. Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Volume I. Long Papers, Toronto, July 9-14, 2023*, Stroudsburg (PA), pp. 15181-15199 (https://aclanthology.org/2023.acl-long.846).

Sánchez-DelaCruz E., Loeza-Mejía C.-I. 2024. *Importance and Challenges of Handwriting Recognition with the Implementation of Machine Learning Techniques: A Survey*, in *Applied Intelligence* 54, pp. 6444–6465 (https://doi.org/10.1007/s10489-024-05487-x).

Shaus A., Sober B., Faigenbaum-Golovin Sh., Mendel-Geberovich A., Piasetzky E., Turkel E. 2016. *Facsimile Creation: Review of Algorithmic Approaches*, in I. Finkelstein, Ch. Robin, Th. Römer (eds.), *Alphabets, Texts and Artifacts in the Ancient near East. Studies presented to Benjamin Sass*, Paris, pp. 474-488.

Singh P., Rutten G., Lefever E. 2021. *A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek*, in S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, S. Szpakowicz (eds.), *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Punta Cana (Dominican Republic) 11 November 2021 (Online)*, Stroudsburg (PA), pp. 128–137 (https://doi.org/10.18653/v1/2021.latechclfl-1.15).

Sommerschield Th., Assael Y., Pavlopoulos J., Stefanak V., Senior A., Dyer Ch., Bodel J., Prag J., Androutsopoulos I., de Freitas N. 2023. *Machine Learning for Ancient Languages: A Survey*, in *Computational Linguistics* 49, pp. 704-747 (https://doi.org/10.1162/coli_a_00481).

Steele Ph.M. 2020. *The Development of Greek Alphabets: Fluctuations and Standardisations*, in Ph.J. Boyes, Ph.M. Steele (eds.), *Understanding Relations between Scripts II. Early Alphabets. Proceedings of a Conference held at the Faculty of Classics in Cambridge 21-22 March 2017*, Oxford-Philadelphia, pp. 125-149 (https://library.oapen.org/handle/20.500.12657/57138).

Tribulato O. 2017. *Decisione della* polis *per lo scriba Spensithios*, in *Axon* 1, pp. 75-87 (http://doi.org/10.14277/2532-6848/Axon-1-1-17-7).

Widiarti A.R., Pulungan R. 2020. *A Method for Solving Scriptio Continua in Javanese Manuscript Transliteration*, in *Heliyon* 6 (https://doi.org/10.1016/j.heliyon.2020.e03827).

Willets L.R. 1967. *The Law Code of Gortyn. With Introduction, Translation, and a Commentary*, Berlin.

Zottin S., De Nardin A., Colombi E., Piciarelli C., Pavan F., Foresti G.L. 2024. *U-Diads-Bib: A Full and Few-Shot Pixel-Precise Dataset for Document Layout Analysis of Ancient*

Manuscripts, in *Neural Computing and Applications* 36, pp. 11777-11789 (https://doi.org/10.1007/s00521-023-09356-5).

Abstract: This paper presents the first results of the LACUNAE project, which aims to develop digital support for the study and restoration of lacunae in ancient Greek inscriptions. While recent research has primarily addressed the problem from a linguistic and textual perspective using Natural Language Processing and Large Language Models, this study proposes a shift in focus by approaching lacunae as material phenomena before treating them as textual ones.

For the first time, Computer Vision (CV) is applied not to character recognition, but to estimating the spatial capacity of a lacuna. The method calculates the minimum and maximum number of characters that a gap may contain based on the proportional relationship between the available space and the size of epichoric letter forms. Through image enhancement, binarization, and layout analysis, the system examines inscribed surfaces and models graphic variability in letter shapes.

This approach makes it possible to objectify a crucial yet traditionally intuitive step in epigraphic practice — estimating the length of a restoration — and to provide quantitative constraints that assist scholars, without automating or determining the philological restoration itself.