Strumenti online per l'analisi e l'annotazione di testi letterari ed epigrafici bilingui

Federico Boschetti

Abstract Methods and tools for the study of bilingual texts in parallel are illustrated. The granularity of the alignments is discussed, distinguishing in particular among document by document, sentence by sentence and word by word alignment. The concept of pericope is defined as a syntactic and semantic unity for the parallelization of consistent and continuous segments. Automated techniques of alignment used by the *Perseus Project* and tools for the dynamic division of texts in parallel pericopes developed at the ILC-CNR are illustrated and criticized. Eventually, the system for the visualization and interrogation of the Latin and Greek version of the *RGDA* edited by Mommsen is described, focusing on the study of the complementarity between the texts.

Keywords Computational Philology, Digital Epigraphy, Parallel texts.

Il presente contributo illustra alcuni metodi e strumenti per l'allineamento di testi bilingui e descrive in particolare il sistema sviluppato presso l'Istituto di Linguistica Computazionale «A. Zampolli» del Consiglio Nazionale delle Ricerche di Pisa, corredato di funzioni specifiche per l'epigrafia digitale.[1] Nato in seno al progetto finanziato dalla Comunità Europea *Greek into Arabic*,[2] il sistema infatti è stato sviluppato in modo tale da poter essere adattato a diverse esigenze di studio legate a differenti discipline umanistiche del mondo classico.

La prima destinazione d'uso dell'applicazione è stata l'allineamento di parte del quarto, del quinto e di parte del sesto libro delle *Enneadi* di Plotino in lingua originale con la *Pseudo-teologia* di Aristotele in lingua araba (D'Ancona 2001). Fra gli scopi principali della ricerca vi è il commento filologico e filosofico dei testi. Lo strumento informatico facilita l'interrogazione e la visualizzazione dei passi in parallelo, oltre a permettere allo studioso di annotare singole parole o porzioni più

1 Durante il convegno veneziano, la presentazione di questo intervento è stata preceduta da un'ampia panoramica sui temi specifici dell'epigrafia digitale, esposta dalla dottoressa Marion Lamé. In attesa della pubblicazione di tale contributo complementare, si possono consultare le note scritte sul taccuino scientifico Épigraphie en Réseau: `http://eer.hypotheses.org` [2013-01-06].

2 Il progetto ERC *Greek into Arabic* vede coinvolti, per la parte Italiana, l'Università di Pisa, la cui unità è diretta dal *principal investigator* Cristina D'Ancona, e l'ILC-CNR di Pisa, la cui unità è guidata da Andrea Bozzi. I materiali inerenti al progetto sono consultabili *on-line* all'indirizzo `http://www.greekintoarabic.eu` [2013-01-06].

estese di testo che si corrispondono, a giudizio delle studioso stesso, in modo più o meno fedele nelle due lingue.

1 Le unità di allineamento e la granularità

Quando si devono allineare testi fortemente correlati (ad esempio un testo greco e la sua traduzione araba, oppure un testo in prosa successivamente versificato etc.), è necessario stabilire la granularità dell'allineamento e di conseguenza le unità dello stesso.

Dalla granularità più fine, che prevede, quando possibile, l'allineamento parola per parola, alla granularità più grossolana, che si limita a stabilire la corrispondenza fra due testi nella stessa lingua o in lingue diverse, si può individuare una gamma di scelte intermedie dettate in parte dalla tipologia dei testi da mettere in parallelo e in parte dalle modalità e dagli scopi della messa in parallelo di tali testi. È la natura dei testi da mettere in parallelo a dare un primo suggerimento per stabilire la granularità degli allineamenti e di questo vi è consapevolezza fin dall'antichità, come sottolinea, fra gli altri, Rochette (1995, pp. 258-259):

> Grâce à plusieurs passages de Quintilien, on voit que le verbe *interpretari* s'applique à un travail qui conserve la pensée, mais pas l'ordre des mot de l'original. Le verbe couvre des domaines aussi divers que la philosophie, l'agricuture ou encore l'ethnographie, mais, contrairement à *vertere*, qui désigne principalement – mais pas exclusivement – des adaptations poétiques, *interpretari* qualifie rarement des traductions d'œuvres poétiques. En outre, contrairement aux autres termes dont dispose le latin pour dire «traduire», *interpretari* est, semble-t-il, le seul qui s'applique à la traduction de mots séparés (*uerba interpretari*). [...] Selons les théories des anciens, l'imitation pouvait revêtir trois formes différentes: l'*interpretatio*, qui consiste à traduire une œuvre fidèlement, selon le contenu et la forme; l'*imitatio*, qui renouvelle la forme en conservant les grandes lignes du contenu de l'original; l'*aemulatio*, qui est une refonte complète du contenu et de la forme.

Sembra chiaro, dunque, che la segmentazione a granularità più fine, cioè a livello di singola parola, è possibile soltanto quando un testo sia da intendersi come *interpretatio* dell'altro. Ma anche in questo caso è necessario distinguere fra traduzione letterale, specialmente di documenti ufficiali o di edizioni critiche moderne con testo a fronte (Karas

2007),[3] e traduzione letteraria. La distinzione non può essere fatta in modo netto, in quanto dipende dalla valutazione del minore o maggiore numero di modifiche (dovute a generalizzazione, specificazione, fraintendimento etc.), omissioni, ampliamenti e trasposizioni che accompagnano la traduzione puntuale. Per questi motivi, la granularità a livello di enunciato (e non di parola) è la più diffusa, soprattutto nei sistemi di allineamento automatico, relativamente ai quali la definizione più comune di *corpus* parallelo si può considerare la seguente: «A *parallel corpus* is a sentence-aligned corpus containing bilingual translations of the same document.» (Fung, Cheung 2004, p. 2).

2 La flessibilità del concetto di pericope

Tuttavia, soprattutto per lo studio del mondo antico e tenendo conto dell'assenza di punteggiatura nei testi originali, la parallelizzazione necessita di unità più articolate rispetto alla *sentence.*

Lasciando aperto allo specialista un margine di azione nella scelta dei confini dei segmenti da allineare, il concetto di pericope, mutuato dagli studi biblici (Lee 2007),[4] si presta allo scopo. La pericope si può infatti intendere come un'unità sintattica (non necessariamente delimitata da punto fermo) e semantica (non necessariamente lessicalizzata) costituita da segmenti coerenti e contigui di un testo di origine corrispondenti a segmenti coerenti e contigui di un testo correlato (Bozzi 1981). Questa definizione, come si vede, non fornisce criteri oggettivi di segmentazione dei testi, se non per il divieto della discontinuità in caso di trasposizioni.

Si può parlare di pericopatura statica, quando la segmentazione sia largamente accettata dalla comunità scientifica per ragioni intrinseche alla natura dei documenti (ad esempio suddivisione in *carmina*), ad indicazioni paratestuali presenti nel supporto originario (ad esempio

3 Una riflessione sulle edizioni con testo a fronte si può trovare in Karas 2007, che, fra l'altro, coglie l'occasione per ribadire (§ 20): «Il convient de consolider d'abord le statut interprétatif de nos traductions bilingues au-delà du fait que la traduction est, sans doute, inséparable de l'interprétation, quel que soit le mode de présentation du texte traduit. Les éditions bilingues manifestent, voire revendiquent, ce statut d'interprétation.»

4 Lee 2007 usa la suddivisione in pericopi, definite come «short, coherent passages», per individuare, con metodi quantitativi, la dipendenza del vangelo di Luca dal vangelo di Marco. È da notare inoltre come l'allineamento in molteplici lingue antiche e moderne del testo biblico e delle sue traduzioni operato da BibleWorks (`http://www.bibleworks.com`) [2013-01-06] rispetti la suddivisione in versetti.

spaziature), o a convenzioni entrate nella pratica corrente (ad esempio la suddivisione in versetti del testo biblico). In questi casi i documenti digitali, usualmente marcati in TEI (http://www.tei-c.org [2013-06-01]) o in formati convertibili in TEI, contengono metainformazioni sufficienti ad un adeguato allineamento automatico a livello di pericope.

Si ha invece pericopatura dinamica quando la segmentazione stessa fa parte dell'attività esegetica dello studioso e deve, quindi, essere riveduta e corretta incrementalmente con il procedere del lavoro di interpretazione del testo. È questo il caso di *Greek into Arabic*, dove il commento insiste su segmenti di testo allineati ma dove, a sua volta, l'allineamento richiede aggiustamenti che emergono durante la stesura dei commenti.

3 L'allineamento automatico: impieghi e limiti

Quando si sia in presenza di *corpora* digitali bilingui di testi in lingua originale con traduzione a fronte, come la collezione di testi greci e latini con traduzione inglese messa a disposizione dal *Perseus Project*,[5] è possibile applicare tecniche statistiche per l'allineamento automatico, con risultati promettenti per quanto sperimentali. Tali tecniche sono usate con successo soprattutto per la messa in parallelo di documenti ufficiali scritti in lingue moderne con le relative traduzioni e adattate alle lingue classiche (Koehn 2005).[6] Bamman, Babeu, Crane 2010 illustrano un procedimento in tre fasi per l'allineamento a diversi gradi di granularità dei testi greci e latini con le relative traduzioni inglesi: scelti manualmente i documenti nella prima fase, la seconda fase consiste nell'applicazione del *Moore's Bilingual Sentence Aligner* (Moore 2002) per l'allineamento automatico degli enunciati; infine la terza fase consiste nell'allineamento delle parole tramite MGIZA++, una variante di GIZA++ (Och, Ney 2003), che costituisce lo standard *de facto* in materia di allineamenti bilingui.

L'allineamento automatico nel campo delle *digital humanities* relative

5 La collezione di testi greci e latini e le relative traduzioni in lingua inglese sono consultabili online e scaricabili all'indirizzo http://www.perseus.tufts.edu [2013-01-06]. È in atto un progetto di collaborazione fra il *Perseus Project* e l'Istituto di Linguistica Computazionale «A. Zampolli» di Pisa per la localizzazione in lingua italiana di parte della collezione, in modo da fornire agli studenti e agli studiosi un *corpus* di traduzioni italiane digitalizzate e non più coperte da *copyright* dei maggiori autori del mondo antico.

6 Koehn 2005 illustra il *corpus* parallelo delle trascrizioni degli interventi al parlamento europeo, al fine di addestrare i sistemi di traduzione automatica.

al mondo classico ha almeno due importanti campi di applicazione. In primo luogo, è possibile la proiezione delle informazioni paratestuali (come le divisioni in capitoli e sezioni) o semantiche (come l'identificazione in modo univoco di etnici e toponimi) dal testo annotato in una lingua al testo non annotato nell'altra lingua. In secondo luogo, è possibile predisporre gli enunciati in parallelo per essere corredati, in fasi successive, di annotazione morfosintattica. Le strutture sintattiche sono poi interrogabili e visualizzabili con traduzione interlineare parola per parola, come avviene grazie agli strumenti offerti dall'*Alpheios Project* (`http://www.alpheios.net` [2013-01-06]).

I limiti maggiori dell'allineamento automatico sono legati alla natura delle traduzioni: quanto più la traduzione è letterale, tanto più il sistema è efficace; quanto più la traduzione è letteraria, tanto meno il sistema è in grado di individuare i segmenti correlati. Per questo motivo la parallelizzazione delle pericopi nel progetto *Greek into Arabic* non si può avvalere di sistemi di allineamento automatici: il valore e la novità dal punto di vista informatico consiste invece, come si è accennato sopra, nel fornire allo studioso strumenti sufficientemente dinamici per partire da un'ipotesi di allineamento manuale, valutarla, commentarla e, se necessario, riformularla dinamicamente senza che l'informazione associata ad altri livelli di granularità (come l'analisi morfosintattica a livello di parola o i commenti a livello di sequenze di parole) sia inficiata.

4 Preparazione del testo latino e del testo greco delle *Res Gestae Divi Augusti* (*RGDA*)

Si procede ora ad illustrare nel dettaglio l'implementazione del sistema di visualizzazione e ricerca delle pericopi in parallelo della versione greca e latina delle *RGDA* secondo il testo dell'edizione critica mommseniana (Mommsen 1883).

Come pianificato in Lamé, Valchera, Boschetti 2012, l'analisi morfologica del testo è stata realizzata in *stand-off*, cioè separando in file indipendenti le annotazioni e le unità testuali su cui tali annotazioni insistono. Quindi, a ciascuna forma flessa del testo è fatta corrispondere in *stand-off* la forma normalizzata, il lemma e la relativa categoria gram-

maticale.[7] La normalizzazione è necessaria per neutralizzare varianti ortografiche conservate nell'edizione critica di Mommsen in conformità alla sua edizione diplomatica corrispondente alla realtà epigrafica, come il nesso nasale + gutturale. La lemmatizzazione è realizzata, tanto per il testo latino quanto per il testo greco, con *Morpheus*, il motore morfologico creato da Gregory Crane presso il *Perseus Project.* In questa fase del lavoro si è proceduto ad una lemmatizzazione completamente automatica, senza correzioni manuali. Nel caso di forme lemmatizzabili in più modi, il sistema sceglie la prima proposta, raggruppando quindi insieme forme che dovrebbero appartenere a paradigmi diversi; quando si effettua una ricerca sui lemmi viene così privilegiata la *recall*, cioè la probabilità di trovare molte occorrenze corrette insieme ad occorrenze non pertinenti, rispetto alla *precision*, cioè la possibilità di trovare solo occorrenze pertinenti. Viene lasciato quindi al vaglio dello studioso lo scarto delle occorrenze non pertinenti.

Come peculiarità del testo epigrafico, a ciascuna forma flessa è associato lo status di conservazione sulla pietra all'epoca di Mommsen, secondo le indicazioni dei segni di integrazione nella sua edizione critica (coerente con l'edizione diplomatica e verificabile sui calchi). I tre livelli codificati sono: forma attestata (completamente leggibile); forma parzialmente attestata (non completamente leggibile); forma totalmente congetturata (illeggibile).

La sequenza delle pericopi, sia nella versione latina che nella versione greca, segue la moderna suddivisione in sezioni costituite da unità sintattiche e semantiche su cui c'è universale consenso da parte degli studiosi. L'allineamento viene effettuato quindi in modo automatico grazie alla semplice corrispondenza dei numeri di sezione.

5 Visualizzazione e ricerca in parallelo

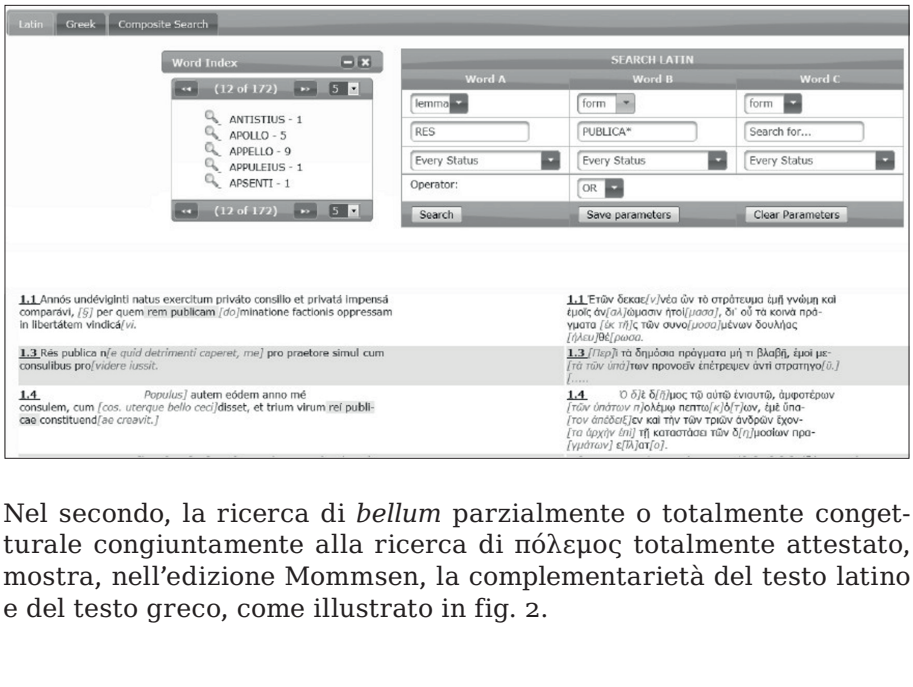
La finalità principale della visualizzazione e della ricerca in parallelo è lo studio della complementarietà fra il testo latino e il testo greco essendo, come è noto, corrotti in luoghi diversi.

7 Il sistema usa il medesimo *framework* adottato per il progetto *Greek into Arabic*, tuttavia in quel caso le informazioni associate al testo arabo sono differenti dalle informazioni associate al testo greco, in quanto l'analisi morfologica dell'arabo prevede forma vocalizzata, lemma, radice e categoria grammaticale. Tuttavia il sistema, progettato dall'autore e dall'ing. Angelo Mario del Grosso con la collaborazione dell'arabista Ouafae Nahli, è sufficientemente flessibile da poter essere esteso o modificato in base alle peculiarità delle lingue prese in esame.

Il motore di ricerca permette l'inserimento di chiavi d'interrogazione per una singola lingua o per entrambe le lingue, separatamente o congiuntamente prese. Coerentemente con quanto detto sopra, la chiave di ricerca, oltre ad essere costituita da forme flesse, può essere costituita da forme normalizzate o da forme lemmatizzate. Lo status della forma (attestata, parzialmente attestata o congetturale) permette di filtrare ulteriormente la ricerca, con opportune combinazioni relative alle due lingue.

Due esempi.[8] Nel primo, illustrato in fig. 1, la ricerca lemmatizzata di *res publica* trova le occorrenze delle forme flesse 1, 1 *rem publicam*; 1, 3 *res publica*; 1, 4 *rei publicae* etc., cui corrispondono in greco, all'interno delle sezioni correlate, 1, 1 τὰ κοινὰ πράγματα; 1, 3 τὰ δημόσια πράγματα; 1, 4 τῶν δ[η]μοσίων πρα[γμάτων] etc.

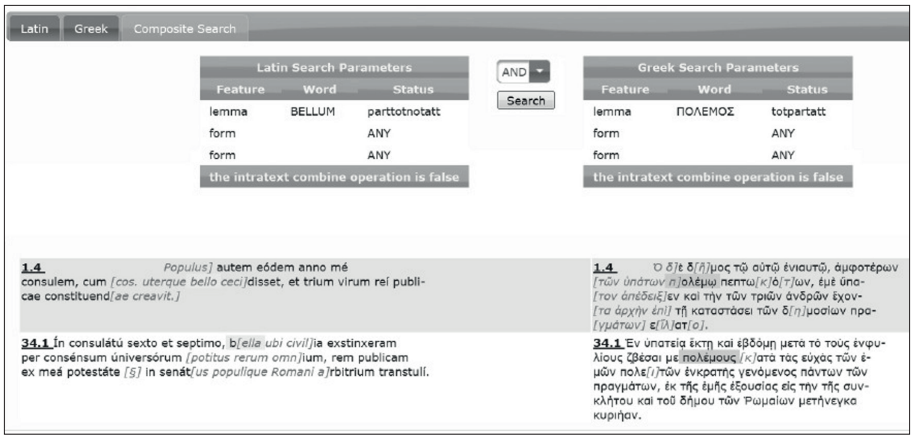
Figura 1. Ricerca lemmatizzata su una sola lingua



Nel secondo, la ricerca di *bellum* parzialmente o totalmente congetturale congiuntamente alla ricerca di πόλεμος totalmente attestato, mostra, nell'edizione Mommsen, la complementarietà del testo latino e del testo greco, come illustrato in fig. 2.

8 Si ringrazia la dottoressa Lamé per la scelta degli esempi.

Figura 2. Ricerca congiunta per lo studio della complementarietà



6 Conclusioni

In conclusione, si è cercato di illustrare alcuni metodi e strumenti per la visualizzazione e la ricerca di testi in parallelo nell'ambito delle *digital humanities* relative agli studi classici, passando poi a descrivere un'applicazione sviluppata presso l'ILC-CNR di Pisa sufficientemente versatile per essere utilizzata tanto per lo studio di testi filosofici greco-arabi quanto per lo studio di un aspetto specifico delle *RGDA* come la complementarietà del testo greco-latino.

Bibliografia

Bamman, David; Babeu, Alison; Crane Gregory (2010). «Transferring Structural Markup Across Translations Using Multilingual Alignment and Projection». In: *Proceedings of the Tenth ACM/IEEE-CS Joint Conference on Digital Libraries, Gold Coast, Australia, June 21-25*. New York: Association for Computing Machinery, pp. 11-20.

Bozzi, Andrea (1981). *Il trattato ippocratico Sulle arie, le acque e i luoghi e la sua traduzione latina tardo-antica. Concordanze contrastive con il calcolatore elettronico e commento linguistico-filologico al lessico tecnico latino*. Pisa: Giardini Editori e Stampatori.

D'Ancona, Cristina (2001). «Pseudo-Theology of Aristotle, Chapter 1: Structure and Composition». *Oriens, Zeitschrift der internationalen Gesellschaft für Orientforschung*, 36, pp. 78-112.

Fung, Pascale; Cheung, Percy (2004). «Multi-level Bootstrapping for Extracting Parallel Sentences from a Quasi-Comparable Corpus». In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04), Stroudsburg, PA, Association for Computational Linguistics*. Genève, pp. 1051-1057.

Karas, Hilla (2007). «Le statut de la traduction dans les éditions bilingues: de l'interprétation au commentaire». *Palimpsestes*, 20, pp. 137-160.

Koehn, Philipp (2005). «Europarl: A Parallel Corpus for Statistical Machine Translation». In: *Proceedings of the MT Summit X*. Phuket: AAMT, pp. 79-86.

Lamé, Marion; Valchera, Valeria; Boschetti, Federico (2012). «Epigrafia digitale. Paradigmi di rappresentazione per il trattamento digitale delle epigrafi». *Epigraphica*, 74 (1-2), pp. 331-338.

Lee, John (2007). «A Computational Model of Text Reuse in Ancient Literary Texts». In: *Proceedings of the 45th Annual Meeting of the association of Computational Linguistics*. Prague: Association for Computational Linguistics, pp. 472-479.

Mommsen, Theodor (1883). *Res Gestae Divi Augusti. Ex Monumentis Ancyrano et Apolloniensi*. 2a ed. Berlin: apud Weidmannos.

Moore, Robert C. (2002). «Fast and Accurate Sentence Alignment of Bilingual Corpora». In: *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: from Research to Real Users*. London: Springer-Verlag, pp. 135-144.

Och, Franz Josef; Ney, Hermann (2003). «A Systematic Comparison of Various Statistical Alignment Models». *Computational Linguistic*, 29 (1), pp. 19-51.

Rochette, Bruno (1995). «Du grec au latin et du latin au grec. Les problèmes de la traduction dans l'antiquité gréco-latine». *Latomus*, 54 (2), pp. 245-261.